

# Evaluation of OpenAI o1: Opportunities and Challenges of AGI

Tianyang Zhong <sup>\*1</sup>, Zhengliang Liu <sup>\*2</sup>, Yi Pan <sup>\*2</sup>, Yutong Zhang <sup>\*3</sup>, Yifan Zhou <sup>\*4,40</sup>, Shizhe Liang <sup>\*5</sup>, Zihao Wu <sup>\*2</sup>, Yanjun Lyu <sup>\*6</sup>, Peng Shu <sup>\*2</sup>, Xiaowei Yu <sup>\*6</sup>, Chao Cao <sup>\*6</sup>, Hanqi Jiang <sup>\*2</sup>, Hanxu Chen <sup>\*7</sup>, Yiwei Li <sup>†2</sup>, Junhao Chen <sup>†2</sup>, Huawei Hu <sup>†8</sup>, Yihen Liu <sup>†9</sup>, Huaqin Zhao <sup>†2</sup>, Shaochen Xu <sup>†2</sup>, Haixing Dai <sup>†2</sup>, Lin Zhao <sup>†2</sup>, Ruidong Zhang <sup>†10</sup>, Wei Zhao <sup>†11,12,13</sup>, Zhenyuan Yang <sup>†14</sup>, Jingyuan Chen <sup>†15</sup>, Peilong Wang <sup>†15</sup>, Wei Ruan <sup>†2</sup>, Hui Wang <sup>†16</sup>, Huan Zhao <sup>†17</sup>, Jing Zhang <sup>‡6</sup>, Yiming Ren <sup>‡18</sup>, Shihuan Qin <sup>‡18</sup>, Tong Chen <sup>‡6</sup>, Jiaxi Li <sup>‡2</sup>, Arif Hassan Zidan <sup>‡19</sup>, Afrar Jahin <sup>‡19</sup>, Minheng Chen <sup>‡6</sup>, Sichen Xia <sup>‡9</sup>, Jason Holmes <sup>‡15</sup>, Yan Zhuang <sup>‡6</sup>, Jiaqi Wang <sup>‡8</sup>, Bochen Xu <sup>‡20</sup>, Weiran Xia <sup>‡21,22</sup>, Jichao Yu <sup>‡2</sup>, Kaibo Tang <sup>‡22</sup>, Yaxuan Yang <sup>‡23</sup>, Bolun Sun <sup>‡24</sup>, Tao Yang<sup>25</sup>, Guoyu Lu<sup>26</sup>, Xianqiao Wang<sup>27</sup>, Lilong Chai<sup>28</sup>, He Li<sup>29</sup>, Jin Lu<sup>2</sup>, Lichao Sun<sup>30</sup>, Xin Zhang<sup>3</sup>, Bao Ge<sup>20</sup>, Xintao Hu<sup>9</sup>, Lian Zhang<sup>18</sup>, Hua Zhou<sup>31</sup>, Lu Zhang<sup>32</sup>, Shu Zhang<sup>8</sup>, Ninghao Liu<sup>2,40</sup>, Bei Jiang<sup>1</sup>, Linglong Kong<sup>1</sup>, Zhen Xiang<sup>2</sup>, Yudan Ren<sup>33</sup>, Jun Liu<sup>11,12</sup>, Xi Jiang<sup>17</sup>, Yu Bao<sup>34</sup>, Wei Zhang<sup>19</sup>, Xiang Li<sup>35</sup>, Gang Li<sup>22</sup>, Wei Liu<sup>15</sup>, Dinggang Shen<sup>36,37</sup>, Andrea Sikora<sup>38</sup>, Xiaoming Zhai<sup>39,40</sup>, Dajiang Zhu<sup>6</sup>, Tuo Zhang<sup>9</sup>, and Tianming Liu<sup>2,40</sup>

<sup>1</sup>Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada

<sup>2</sup>School of Computing, University of Georgia, GA, USA

<sup>3</sup>Institute of Medical Research, Northwestern Polytechnical University, Xi'an, China

<sup>4</sup>College of Arts and Sciences, University of Georgia, Athens, USA

<sup>5</sup>Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA, USA

<sup>6</sup>Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA

<sup>7</sup>The Lamar Dodd School of Art, University of Georgia, GA, USA

<sup>8</sup>School of Computer Science, Northwestern Polytechnical University, Xi'an, China

<sup>9</sup>School of Automation, Northwestern Polytechnical University, Xi'an, China

<sup>10</sup>University of California, Los Angeles, CA, USA

<sup>11</sup>Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha, China

<sup>12</sup>Clinical Research Center for Medical Imaging in Hunan Province, Changsha, China

<sup>13</sup>Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

<sup>14</sup>Guanghua School of Management, Peking University, Beijing, China

<sup>15</sup>Department of Radiation Oncology, Mayo Clinic, Phoenix, Arizona, USA

<sup>16</sup>Second Language Acquisition and Teaching, University of Arizona, Tucson, AZ, USA

<sup>17</sup>School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

<sup>18</sup>The First Hospital of Hebei Medical University, Shijiazhuang, Hebei, China

<sup>19</sup>School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA

<sup>20</sup>School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China

<sup>21</sup>School of Future Technology, South China University of Technology, Guangzhou, China

- <sup>22</sup>Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC, USA  
<sup>23</sup>Department of Educational Psychology, University of Georgia, Athens, GA, USA  
<sup>24</sup>Johns Hopkins University, Baltimore, MD, USA  
<sup>25</sup>School of Architecture, Tsinghua University, Beijing, China  
<sup>26</sup>College of Engineering, University of Georgia, Athens, USA  
<sup>27</sup>School of Environmental, Civil, Agricultural and Mechanical Engineering, College of Engineering, University of Georgia, Athens, GA, USA  
<sup>28</sup>Department of Poultry Science, College of Agricultural and Environmental Sciences, University of Georgia, Athens, GA, USA  
<sup>29</sup>School of Chemical, Materials and Biomedical Engineering, University of Georgia, Athens, USA  
<sup>30</sup>Department of Computer Science and Engineering, Lehigh University, PA, USA  
<sup>31</sup>Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA  
<sup>32</sup>Department of Computer Science, Indiana University Indianapolis, IN, USA  
<sup>33</sup>School of Information Science & Technology, Northwest University, Xi'an, China  
<sup>34</sup>Department of Graduate Psychology, James Madison University, Harrisonburg, VA, USA  
<sup>35</sup>Department of Radiology, Massachusetts General Hospital and Harvard Medical School, MA, USA  
<sup>36</sup>School of Biomedical Engineering, ShanghaiTech University, and Shanghai Clinical Research and Trial Center, Shanghai, China  
<sup>37</sup>Shanghai United Imaging Intelligence Co., Ltd.  
<sup>38</sup>College of Pharmacy, University of Georgia, GA, USA  
<sup>39</sup>Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA  
<sup>40</sup>AI4STEM Education Center, University of Georgia, Athens, GA, USA

---

\*Co-first authors.

<sup>†</sup>Co-second authors.

<sup>‡</sup>Co-third authors.

---

## Abstract

This comprehensive study evaluates the performance of OpenAI's o1-preview large language model across a diverse array of complex reasoning tasks, spanning multiple domains, including computer science, mathematics, natural sciences, medicine, linguistics, and social sciences. Through rigorous testing, o1-preview demonstrated remarkable capabilities, often achieving human-level or superior performance in areas ranging from coding challenges to scientific reasoning and from language processing to creative problem-solving. Key findings include:

- 83.3% success rate in solving complex competitive programming problems, surpassing many human experts.
- Superior ability in generating coherent and accurate radiology reports, outperforming other evaluated models.
- 100% accuracy in high school-level mathematical reasoning tasks, providing detailed step-by-step solutions.
- Advanced natural language inference capabilities across general and specialized domains like medicine.
- Impressive performance in chip design tasks, outperforming specialized models in areas such as EDA script generation and bug analysis.
- Remarkable proficiency in anthropology and geology, demonstrating deep understanding and reasoning in these specialized fields.
- Strong capabilities in quantitative investing. O1 has comprehensive financial knowledge and statistical modeling skills.
- Effective performance in social media analysis, including sentiment analysis and emotion recognition.

The model excelled particularly in tasks requiring intricate reasoning and knowledge integration across various fields. While some limitations were observed, including occasional errors on simpler problems and challenges with certain highly specialized concepts, the overall results indicate significant progress towards artificial general intelligence. This evaluation not only highlights o1-preview's current strengths and limitations but also identifies crucial areas for future development, including multi-modal integration, domain-specific validation, and ethical considerations for real-world applications. The findings provide valuable insights into the potential of large language models in numerous fields and pave the way for further advancements in AI research and application.

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                                       | <b>11</b> |
| 1.1      | Background: What is New with o1 . . . . .                 | 11        |
| 1.1.1    | OpenAI o1 & Chain-of-Thought Reasoning . . . . .          | 11        |
| 1.1.2    | OpenAI o1 & Reinforcement Learning . . . . .              | 11        |
| 1.2      | Motivation . . . . .                                      | 12        |
| 1.3      | Key Findings . . . . .                                    | 12        |
| 1.4      | AGI-Benchmark 1.0 . . . . .                               | 13        |
| <b>2</b> | <b>Scope of the Study and Used Public Datasets</b>        | <b>15</b> |
| 2.1      | Code Generation . . . . .                                 | 15        |
| 2.2      | Radiology Report Generation . . . . .                     | 16        |
| 2.3      | Robot Command Planning . . . . .                          | 17        |
| 2.4      | Nature Language Inference . . . . .                       | 17        |
| 2.5      | Quantitative Investing . . . . .                          | 17        |
| 2.6      | Low-Resource Language Translation . . . . .               | 19        |
| 2.7      | Educational Q&A . . . . .                                 | 19        |
| 2.8      | Student Writing Improvement in Higher Education . . . . . | 20        |
| 2.9      | 3D Layout Generation . . . . .                            | 20        |
| 2.10     | Chip Design . . . . .                                     | 21        |
| 2.10.1   | Transformative Potential of LLMs in Chip Design . . . . . | 21        |
| 2.10.2   | Experimenting with o1-preview in Chip Design . . . . .    | 22        |
| 2.10.3   | A Leap Toward AGI and the Future of Chip Design . . . . . | 22        |
| 2.11     | Logical Reasoning . . . . .                               | 23        |
| 2.12     | Table-to-Text Generation . . . . .                        | 24        |
| 2.13     | High-School-Level Math Competition . . . . .              | 24        |
| 2.14     | College-level Math Problems . . . . .                     | 26        |
| 2.15     | Electronic Health Record Diagnosis . . . . .              | 26        |
| 2.16     | Stochastic Processes in Statistics . . . . .              | 27        |
| 2.17     | Medical Text Anonymization . . . . .                      | 28        |
| 2.18     | Social Media Analysis . . . . .                           | 28        |
| 2.19     | Analogical Reasoning . . . . .                            | 29        |
| 2.20     | Sentiment Analysis . . . . .                              | 29        |
| 2.21     | Anthropology and Geology . . . . .                        | 30        |
| 2.22     | Educational Measurement and Psychometrics . . . . .       | 31        |
| 2.23     | Public Health Policy Analysis . . . . .                   | 31        |
| 2.24     | Medical Genetics and Genomics Reasoning . . . . .         | 32        |
| 2.25     | Medical Knowledge Question Answer . . . . .               | 33        |
| 2.26     | Art Education . . . . .                                   | 33        |
| 2.27     | Content Summarization . . . . .                           | 34        |
| <b>3</b> | <b>Related Work</b>                                       | <b>34</b> |
| 3.1      | Foundation Models . . . . .                               | 34        |
| 3.2      | Prompt Engineering . . . . .                              | 35        |
| 3.3      | Chain of Thought . . . . .                                | 36        |
| 3.4      | Multi-modal Large Language Models . . . . .               | 37        |
| 3.5      | Fine-tuning Large Language Models . . . . .               | 38        |

---

|          |   |            |
|----------|---|------------|
| 3.6      | Large Language Model Agent and Retrieval-Augmented Generation . . . . .           | 39         |
| 3.7      | Large Language Models & Reasoning . . . . .                                       | 40         |
| 3.8      | Reinforcement Learning with Human Feedback . . . . .                              | 41         |
| 3.9      | Evaluation Complex Reasoning Tasks . . . . .                                      | 42         |
| <b>4</b> | <b>Experiments and Observation</b>  | <b>45</b>  |
| 4.1      | Test Procedure . . . . .  | 45         |
| 4.2      | Code Generation . . . . .   | 46         |
| 4.3      | Radiology Report Generation . . . . .   | 49         |
| 4.4      | Robot Command Planning . . . . .  | 52         |
| 4.5      | Nature Language Inference . . . . .   | 54         |
| 4.6      | Quantitative Investing . . . . .  | 56         |
| 4.7      | Low-Resource Language Translation . . . . .                                       | 58         |
| 4.8      | Educational Q&A . . . . .   | 60         |
| 4.9      | Student Writing Improvement in Higher Education . . . . .                         | 62         |
| 4.10     | 3D Layout Generation . . . . .  | 66         |
| 4.11     | Chip Design . . . . .   | 72         |
| 4.11.1   | Engineering Assistant Chatbot . . . . .   | 72         |
| 4.11.2   | EDA Script Generation . . . . .   | 73         |
| 4.11.3   | Bug Summary & Analysis . . . . .  | 74         |
| 4.12     | Logical Reasoning . . . . .   | 88         |
| 4.13     | Table-to-Text Generation . . . . .  | 91         |
| 4.14     | High School Level Math Competition . . . . .                                      | 91         |
| 4.15     | College-level Math Problems . . . . .   | 99         |
| 4.16     | Electronic Health Record Diagnosis . . . . .                                      | 105        |
| 4.17     | Stochastic Processes in Statistics . . . . .                                      | 110        |
| 4.18     | Medical Text Anonymization . . . . .  | 113        |
| 4.19     | Social Media Analysis . . . . .   | 117        |
| 4.20     | Analogical Reasoning . . . . .  | 120        |
| 4.21     | Sentiment Analysis . . . . .  | 123        |
| 4.22     | Anthropology and Geology . . . . .  | 126        |
| 4.23     | Educational Measurement and Psychometrics . . . . .                               | 133        |
| 4.24     | Public Health Policy Analysis . . . . .   | 137        |
| 4.25     | Medical Genetics and Genomics Reasoning . . . . .                                 | 140        |
| 4.26     | Medical Knowledge Question Answer . . . . .                                       | 143        |
| 4.27     | Art Education . . . . .   | 147        |
| 4.28     | Content Summarization . . . . .   | 151        |
| <b>5</b> | <b>Discussion</b>   | <b>154</b> |
| 5.1      | LLM Agents and o1: Advancing Problem-Solving in Science and Engineering . . . . . | 155        |
| <b>6</b> | <b>Conclusion</b>   | <b>156</b> |
| <b>A</b> | <b>Appendix</b>   | <b>157</b> |
| A.1      | Code Generation . . . . .   | 157        |
| A.2      | Radiology Report Generation . . . . .   | 167        |
| A.3      | Robot Command Planning . . . . .  | 175        |
| A.4      | Nature Language Inference . . . . .   | 179        |

---

|      |   |     |
|------|---|-----|
| A.5  | Quantitative Investing . . . . .                  | 181 |
| A.6  | Low-Resource Language Translation . . . . .       | 184 |
| A.7  | Educational Q&A . . . . .                         | 187 |
| A.8  | Logical Reasoning . . . . .                       | 190 |
| A.9  | High School Level Math Competition . . . . .      | 198 |
| A.10 | College-level Math Problems . . . . .             | 204 |
| A.11 | Electronic Health Record Diagnosis . . . . .      | 213 |
| A.12 | Stochastic Processes in Statistics . . . . .      | 219 |
| A.13 | Medical Text Anonymization . . . . .              | 223 |
| A.14 | Social Media Analysis . . . . .                   | 229 |
| A.15 | Analogical Reasoning . . . . .                    | 231 |
| A.16 | Sentiment Analysis . . . . .                      | 233 |
| A.17 | Public Health Policy Analysis . . . . .           | 235 |
| A.18 | Medical Genetics and Genomics Reasoning . . . . . | 242 |
| A.19 | Medical Knowledge Question Answer . . . . .       | 243 |
| A.20 | Art Education . . . . .                           | 258 |
| A.21 | Content Summarization . . . . .                   | 264 |

## List of Figures

|    |   |    |
|----|---|----|
| 1  | Schematic Overview of the Evaluation Methodology . . . . .                        | 15 |
| 2  | An example of a patient table and its corresponding clinical description. . . . . | 25 |
| 3  | Code Generation: Case 1 . . . . .   | 47 |
| 4  | Code Generation: Case 2 . . . . .   | 48 |
| 5  | Radiology Report Generation: Case 1 . . . . .                                     | 50 |
| 6  | Radiology Report Generation: Case 2 . . . . .                                     | 51 |
| 7  | Robot Command Planning: Case 1 . . . . .  | 52 |
| 8  | Quantitative Investing: Case 1 . . . . .  | 56 |
| 9  | Quantitative Investing: Case 2 . . . . .  | 57 |
| 10 | Low-Resource Language Translation: Case 1 . . . . .                               | 58 |
| 11 | Low-Resource Language Translation: Case 2 . . . . .                               | 59 |
| 12 | Educational Q&A: Case 1 . . . . .   | 60 |
| 13 | Educational Q&A: Case 2 . . . . .   | 61 |
| 14 | Academic Writing Improvement: Case 1 . . . . .                                    | 63 |
| 15 | Academic Writing Improvement: Case 2 . . . . .                                    | 64 |
| 16 | Academic Writing Improvement: Case 3 . . . . .                                    | 65 |
| 17 | 3D Layout Generation: Case 1 . . . . .  | 68 |
| 18 | 3D Layout Generation: Case 2 . . . . .  | 69 |
| 19 | 3D Layout Generation: Case 3 . . . . .  | 70 |
| 20 | 3D Layout Generation: Case 4 . . . . .  | 71 |
| 21 | Chip Design-Engineering Assistant Chatbot: Case 1 . . . . .                       | 76 |
| 22 | Chip Design-Engineering Assistant Chatbot: Case 2 . . . . .                       | 77 |
| 23 | Chip Design-Engineering Assistant Chatbot: Case 3 . . . . .                       | 79 |
| 24 | Chip Design-EDA Script Generation: Case 1 . . . . .                               | 81 |
| 25 | Chip Design-EDA Script Generation: Case 2 . . . . .                               | 84 |
| 26 | Chip Design-Bug Summary & Analysis: Case 1 . . . . .                              | 87 |
| 27 | Logical Reasoning: Case 1 . . . . .   | 89 |
| 28 | Logical Reasoning: Case 2 . . . . .   | 90 |

---

|    |   |     |
|----|---|-----|
| 29 | Table-to-Text Generation: Case 1 . . . . .                  | 92  |
| 30 | Table-to-Text Generation: Case 2 . . . . .                  | 93  |
| 31 | High School Level Math Competition: Case 1 . . . . .        | 95  |
| 32 | High School Level Math Competition: Case 2 . . . . .        | 96  |
| 33 | High School Level Math Competition: Case 3 . . . . .        | 97  |
| 34 | High School Level Math Competition: Case 4 . . . . .        | 98  |
| 35 | College-Level Math: Case 1 . . . . .                        | 102 |
| 36 | College-Level Math: Case 5 . . . . .                        | 103 |
| 37 | College-Level Math: Case 7 . . . . .                        | 104 |
| 38 | College-Level Math: Case 7 (GPT-4o) . . . . .               | 105 |
| 39 | Electronic Health Record Diagnosis: Case 1 . . . . .        | 106 |
| 40 | Electronic Health Record Diagnosis: Case 2 . . . . .        | 107 |
| 41 | Electronic Health Record Diagnosis: Case 3 . . . . .        | 108 |
| 42 | Electronic Health Record Diagnosis: Case 4 . . . . .        | 109 |
| 43 | Stochastic Processes in Statistics: Case 1 . . . . .        | 111 |
| 44 | Stochastic Processes in Statistics: Case 2 . . . . .        | 112 |
| 45 | Medical Text Anonymization: Case 1 . . . . .                | 114 |
| 46 | Medical Text Anonymization: Case 2 . . . . .                | 115 |
| 47 | Medical Text Anonymization: Case 3 . . . . .                | 116 |
| 48 | Social Media Analysis: Case 1-3 . . . . .                   | 118 |
| 49 | Social Media Analysis: Case 4-5 . . . . .                   | 119 |
| 50 | Analogical Reasoning: Case 1-3 . . . . .                    | 121 |
| 51 | Analogical Reasoning: Case 4-5 . . . . .                    | 122 |
| 52 | Sentiment Analysis: Case 1. . . . .                         | 124 |
| 53 | Sentiment Analysis: Case 2. . . . .                         | 124 |
| 54 | Sentiment Analysis: Case 3. . . . .                         | 125 |
| 55 | Sentiment Analysis: Case 4. . . . .                         | 125 |
| 56 | Anthropology Reasoning: Case 1. . . . .                     | 126 |
| 57 | Anthropology Reasoning: Case 2. . . . .                     | 127 |
| 58 | Anthropology Reasoning: Case 3. . . . .                     | 129 |
| 59 | Anthropology Reasoning: Case 4. . . . .                     | 131 |
| 60 | Geology Reasoning: Case 1. . . . .                          | 132 |
| 61 | Educational Measurement and Psychometrics: Case 1 . . . . . | 134 |
| 62 | Educational Measurement and Psychometrics: Case 2 . . . . . | 135 |
| 63 | Educational Measurement and Psychometrics: Case 3 . . . . . | 136 |
| 64 | Public Health Policy Analysis: Case 1-2. . . . .            | 138 |
| 65 | Public Health Policy Analysis: Case 3. . . . .              | 139 |
| 66 | Medical Genetics and Genomics Reasoning: Case 1. . . . .    | 141 |
| 67 | Medical Genetics and Genomics Reasoning: Case 2. . . . .    | 142 |
| 68 | Medical Knowledge Question Answer: Case 1 . . . . .         | 144 |
| 69 | Medical Knowledge Question Answer: Case 2 . . . . .         | 146 |
| 70 | Art Education: Case 1 . . . . .                             | 148 |
| 71 | Art Education: Case 2 . . . . .                             | 150 |
| 72 | Content Summarization: Case 1. . . . .                      | 152 |
| 73 | Content Summarization: Case 2. . . . .                      | 153 |
| 74 | Code Generation: Case 3 . . . . .                           | 157 |
| 75 | Code Generation: Case 4 . . . . .                           | 158 |
| 76 | Code Generation: Case 5 . . . . .                           | 159 |

---

---

|     |   |     |
|-----|---|-----|
| 77  | Code Generation: Case 6 . . . . .                     | 160 |
| 78  | Code Generation: Case 7 . . . . .                     | 161 |
| 79  | Code Generation: Case 8 . . . . .                     | 162 |
| 80  | Code Generation: Case 9 . . . . .                     | 163 |
| 81  | Code Generation: Case 10 . . . . .                    | 164 |
| 82  | Code Generation: Case 11 . . . . .                    | 165 |
| 83  | Code Generation: Case 12 . . . . .                    | 166 |
| 84  | Radiology Report Generation: Case 3 . . . . .         | 167 |
| 85  | Radiology Report Generation: Case 4 . . . . .         | 168 |
| 86  | Radiology Report Generation: Case 5 . . . . .         | 169 |
| 87  | Radiology Report Generation: Case 6 . . . . .         | 170 |
| 88  | Radiology Report Generation: Case 7 . . . . .         | 171 |
| 89  | Radiology Report Generation: Case 8 . . . . .         | 172 |
| 90  | Radiology Report Generation: Case 9 . . . . .         | 173 |
| 91  | Radiology Report Generation: Case 10 . . . . .        | 174 |
| 92  | Robot Command Planning: Case 2 . . . . .              | 175 |
| 93  | Robot Command Planning: Case 3 . . . . .              | 176 |
| 94  | Robot Command Planning: Case 4 . . . . .              | 177 |
| 95  | Robot Command Planning: Case 5 . . . . .              | 178 |
| 96  | Quantitative Investing: Case 3 . . . . .              | 181 |
| 97  | Quantitative Investing: Case 4 . . . . .              | 182 |
| 98  | Quantitative Investing: Case 5 . . . . .              | 183 |
| 99  | Low-Resource Language Translation: Case 3 . . . . .   | 184 |
| 100 | Low-Resource Language Translation: Case 4 . . . . .   | 185 |
| 101 | Low-Resource Language Translation: Case 5 . . . . .   | 186 |
| 102 | Educational Q&A: Case 3 . . . . .                     | 187 |
| 103 | Educational Q&A: Case 4 . . . . .                     | 188 |
| 104 | Educational Q&A: Case 5 . . . . .                     | 189 |
| 105 | Logical Reasoning: Case 3 . . . . .                   | 190 |
| 106 | Logical Reasoning: Case 4 . . . . .                   | 191 |
| 107 | Logical Reasoning: Case 5 . . . . .                   | 192 |
| 108 | Logical Reasoning: Case 6 . . . . .                   | 193 |
| 109 | Logical Reasoning: Case 7 . . . . .                   | 194 |
| 110 | Logical Reasoning: Case 8 . . . . .                   | 195 |
| 111 | Logical Reasoning: Case 9 . . . . .                   | 196 |
| 112 | Logical Reasoning: Case 10 . . . . .                  | 197 |
| 113 | High School Level Math Competition: Case 5 . . . . .  | 198 |
| 114 | High School Level Math Competition: Case 6 . . . . .  | 199 |
| 115 | High School Level Math Competition: Case 7 . . . . .  | 200 |
| 116 | High School Level Math Competition: Case 8 . . . . .  | 201 |
| 117 | High School Level Math Competition: Case 9 . . . . .  | 202 |
| 118 | High School Level Math Competition: Case 10 . . . . . | 203 |
| 119 | College-Level Math: Case 2 . . . . .                  | 204 |
| 120 | College-Level Math: Case 3 . . . . .                  | 205 |
| 121 | College-Level Math: Case 4 . . . . .                  | 206 |
| 122 | College-Level Math: Case 6 . . . . .                  | 207 |
| 123 | College-Level Math: Case 8 . . . . .                  | 208 |
| 124 | College-Level Math: Case 8.5 . . . . .                | 209 |

---

---

|     |   |     |
|-----|---|-----|
| 125 | College-Level Math: Case 9 . . . . .                      | 210 |
| 126 | College-Level Math: Case 10 . . . . .                     | 211 |
| 127 | College-Level Math: Case 10.5 . . . . .                   | 212 |
| 128 | Electronic Health Record Diagnosis: Case 5 . . . . .      | 213 |
| 129 | Electronic Health Record Diagnosis: Case 6 . . . . .      | 214 |
| 130 | Electronic Health Record Diagnosis: Case 7 . . . . .      | 215 |
| 131 | Electronic Health Record Diagnosis: Case 8 . . . . .      | 216 |
| 132 | Electronic Health Record Diagnosis: Case 9 . . . . .      | 217 |
| 133 | Electronic Health Record Diagnosis: Case 10 . . . . .     | 218 |
| 134 | Stochastic Processes in Statistics: Case 3 . . . . .      | 219 |
| 135 | Stochastic Processes in Statistics: Case 4 . . . . .      | 220 |
| 136 | Stochastic Processes in Statistics: Case 5 . . . . .      | 221 |
| 137 | Stochastic Processes in Statistics: Case 6 . . . . .      | 222 |
| 138 | Medical Text Anonymization: Case 4 . . . . .              | 223 |
| 139 | Medical Text Anonymization: Case 5 . . . . .              | 224 |
| 140 | Medical Text Anonymization: Case 6 . . . . .              | 225 |
| 141 | Medical Text Anonymization: Case 7 . . . . .              | 226 |
| 142 | Medical Text Anonymization: Case 8 . . . . .              | 227 |
| 143 | Medical Text Anonymization: Case 9 . . . . .              | 228 |
| 144 | Social Media Analysis: Case 6-9 . . . . .                 | 229 |
| 145 | Social Media Analysis: Case 10-11 . . . . .               | 230 |
| 146 | Analogical Reasoning: Case 6-8 . . . . .                  | 231 |
| 147 | Analogical Reasoning: Case 9-10 . . . . .                 | 232 |
| 148 | Sentiment analysis: Case 5 . . . . .                      | 233 |
| 149 | Sentiment analysis: Case 6 . . . . .                      | 234 |
| 150 | Public Health Policy Analysis: Case 4 . . . . .           | 235 |
| 151 | Public Health Policy Analysis: Case 5 . . . . .           | 236 |
| 152 | Public Health Policy Analysis: Case 6 . . . . .           | 237 |
| 153 | Public Health Policy Analysis: Case 7 . . . . .           | 238 |
| 154 | Public Health Policy Analysis: Case 8 . . . . .           | 239 |
| 155 | Public Health Policy Analysis: Case 9 . . . . .           | 240 |
| 156 | Public Health Policy Analysis: Case 10 . . . . .          | 241 |
| 156 | Medical Genetics and Genomics Reasoning: Case 3 . . . . . | 242 |
| 157 | Medical Knowledge Question Answer: Case 3 . . . . .       | 243 |
| 158 | Medical Knowledge Question Answer: Case 4 . . . . .       | 245 |
| 159 | Medical Knowledge Question Answer: Case 5 . . . . .       | 246 |
| 160 | Medical Knowledge Question Answer: Case 6 . . . . .       | 248 |
| 161 | Medical Knowledge Question Answer: Case 7 . . . . .       | 251 |
| 162 | Medical Knowledge Question Answer: Case 8 . . . . .       | 253 |
| 163 | Medical Knowledge Question Answer: Case 9 . . . . .       | 255 |
| 164 | Medical Knowledge Question Answer: Case 10 . . . . .      | 257 |
| 165 | Art Education: Case 3 . . . . .                           | 258 |
| 166 | Art Education: Case 4 . . . . .                           | 260 |
| 167 | Art Education: Case 5 . . . . .                           | 262 |
| 168 | Art Education: Case 6 . . . . .                           | 263 |
| 169 | Content Summarization: Case 3 . . . . .                   | 264 |
| 170 | Content Summarization: Case 4 . . . . .                   | 265 |
| 171 | Content Summarization: Case 5 . . . . .                   | 266 |

---

---

|     |   |     |
|-----|---|-----|
| 172 | Content Summarization: Case 6 . . . . . | 267 |
| 173 | Content Summarization: Case 7 . . . . . | 268 |
| 174 | Content Summarization: Case 8 . . . . . | 269 |

---

# 1 Introduction

large language models (LLMs) have rapidly advanced in recent years, demonstrating impressive capabilities across a wide range of tasks [91, 217, 215, 101]. This study aims to comprehensively evaluate the performance of OpenAI’s **o1** model, widely perceived as the model with the highest reasoning capabilities in history so far, on complex reasoning tasks spanning multiple disciplines. O1 represents a significant leap forward in AI reasoning capabilities. According to OpenAI’s blog post [129], o1 demonstrates remarkable proficiency across various domains, including competitive programming, advanced mathematics, and PhD-level scientific problem-solving. It ranks in the 89th percentile on competitive programming questions, places among the top 500 students in the US in a qualifier for the USA Math Olympiad, and surpasses human PhD-level accuracy on a benchmark of physics, biology, and chemistry problems. These achievements underscore o1’s potential to revolutionize AI applications in science, coding, mathematics, and related fields.

While standard benchmarks have shown promising results [129], they can be manipulated and are often not comprehensive enough to fully assess the model’s capabilities [116, 121]. By rigorously testing the model’s abilities across a wide range of domains with real world tasks, we seek to provide a more holistic evaluation of its progress towards artificial general intelligence and identify areas for further development.

## 1.1 Background: What is New with o1

LLMs, built on the Transformer [176] architecture, have evolved from early work such as BERT [26] and GPT [143] to more advanced models such as GPT-3 and GPT-4 [14, 2]. These models, trained on vast corpora of text data, have shown significant proficiency in understanding context, generating human-like text, and performing complex reasoning tasks [91, 217, 215, 101, 111, 93, 51, 213, 70, 107, 182, 167, 106, 169].

### 1.1.1 OpenAI o1 & Chain-of-Thought Reasoning

Chain-of-thought reasoning [191] is a recent advancement in LLMs that enables models to break down complex problems into intermediate steps, mirroring human-like problem-solving processes. This approach has shown particular promise in enhancing performance on tasks requiring multi-step reasoning or mathematical problem-solving. By explicitly generating a series of coherent thoughts leading to a conclusion, models can tackle more sophisticated problems and provide more interpretable outputs.

Compared to its predecessor GPT-4, o1 explicitly incorporates chain-of-thought into its inference process [129]. While GPT-4 could utilize chain-of-thought reasoning when prompted, o1 integrates this approach as a fundamental part of its architecture. This allows o1 to "think before it answers," producing a long internal chain of thought before responding to the user. This integration enables o1 to handle more complex reasoning tasks and provide more transparent explanations of its problem-solving process, potentially surpassing GPT-4’s capabilities in areas requiring deep, multi-step reasoning.

### 1.1.2 OpenAI o1 & Reinforcement Learning

Reinforcement Learning from Human Feedback (RLHF) is a powerful technique that has significantly advanced the capabilities of large language models. It combines reinforcement learning principles with human preferences to fine-tune models, aligning their outputs more closely with human expectations.

---

RLHF typically involves supervised fine-tuning, reward modeling based on human preferences, and policy optimization through reinforcement learning. For a more detailed discussion of RLHF, see Section 3.8.

O1 employs advanced reinforcement learning techniques that significantly evolve beyond traditional RLHF methods. According to OpenAI, o1’s performance consistently improves with more reinforcement learning (train-time compute) and with more time spent thinking (test-time compute) [129]. This novel approach likely incorporates Chain-of-Thought reasoning into its reinforcement learning framework, which allows it to generate and evaluate multiple reasoning paths before producing a final output.

Unlike traditional models that primarily spend compute during training, o1 scales its performance with increased compute during inference. This suggests a form of online learning or search that occurs at test time, which could involve real-time exploration and refinement of reasoning strategies. The model potentially rewards not just the final answer, but also the quality and effectiveness of its reasoning steps. O1 appears to have mechanisms for self-reflection and improvement, which implement a form of self-supervised learning where the model’s thoughts become training data for further enhancement [129, 118, 66].

It might generate multiple candidate reasoning paths in parallel, which use reinforcement learning to score and select the most promising paths, similar to the Quiet-STaR method [205, 118]. The model’s ability to improve with more thinking time” suggests a continuous learning loop, which possibly implements a form of meta-learning that adapts its reasoning strategies based on the specific task.

These advanced techniques represent a paradigm shift which focuses on enhancing the model’s reasoning capabilities during inference, rather than solely aligning its outputs with human preferences during training.

## 1.2 Motivation

As LLMs grow more sophisticated, there is a critical need to understand their true capabilities and limitations beyond standard benchmarks. This study is driven by the goal of assessing o1-preview’s ability to handle complex, multi-disciplinary tasks that require deep reasoning and knowledge integration. Such a comprehensive evaluation provides valuable insights into the current state of LLM technology and its potential for real-world applications, which is of significant interest to the scientific community.

To comprehensively assess the capabilities of o1-preview, we structured our evaluation around five major domains: Creation and Design, Planning, Reasoning, Diagnosis, and Reflection (see Figure 1). Each domain encompasses a set of relevant tasks that test specific aspects of the model’s performance. In total, 27 tasks were designed to evaluate the model’s adaptability and effectiveness across a diverse array of cognitive and real-world challenges.

## 1.3 Key Findings

Our comprehensive evaluation of o1-preview across various domains revealed several main insights:

- **Advanced Reasoning Capabilities:** o1-preview demonstrated exceptional logical reasoning abilities in multiple fields, including high school mathematics, quantitative investing, and chip

---

design. It showed a strong capacity for step-by-step problem-solving and the ability to handle complex, multi-layered tasks.

- **Domain-Specific Knowledge:** The model exhibited impressive knowledge breadth across diverse fields such as medical genetics, radiology, anthropology, and geology. It often performed at a level comparable to or exceeding that of graduate students or early-career professionals in these domains.
- **Creative and Practical Applications:** In areas such as 3D layout generation and art education, o1-preview showed creativity and practical application skills, generating functional designs and structured lesson plans. However, it still lacks the flexibility and adaptability of human experts in these fields.
- **Natural Language Understanding:** The model excelled in tasks requiring nuanced language understanding, such as sentiment analysis, social media analysis, and content summarization. It demonstrated the ability to capture complex expressions like irony and sarcasm, though it still struggles with very subtle emotional nuances.
- **Scientific and Medical Reasoning:** o1-preview showed strong capabilities in medical diagnosis, radiology report generation, and answering complex medical exam questions. While it performed well in these areas, its reasoning process sometimes differed from that of trained medical professionals.
- **Limitations and Areas for Improvement:** Despite its impressive performance, o1-preview showed limitations in handling extremely abstract logical puzzles, adapting to real-time dynamic situations, and consistently performing well on the most complex tasks in fields like advanced mathematics and stochastic processes.
- **Potential for Real-World Applications:** The model's performance suggests significant potential for applications in various fields, from educational support and medical assistance to financial analysis and scientific research. However, further refinement and validation are necessary before deployment in critical real-world scenarios.

## 1.4 AGI-Benchmark 1.0

To contribute to the field of AI research and evaluation, we are introducing **AGI-Benchmark 1.0**, a comprehensive collection of the complex reasoning tasks used in this study to evaluate o1-preview. Unlike existing language model benchmarks such as MMLU [45], which primarily focus on question-answering and multiple-choice formats, AGI-Benchmark 1.0 is designed to assess a model's ability to tackle intricate, multi-step reasoning problems across a diverse set of domains.

This benchmark encompasses tasks from **27 distinct categories**, grouped into five major cognitive faculties:

- **Reasoning:**
  - Natural Language Inference
  - Logical Reasoning
  - High School Level Math Competition
  - College-level Math Problems

- 
- Analogical Reasoning
  - Anthropology and Geology

- **Planning:**

- Robot Command Planning
- Quantitative Investing
- Public Health Policy Analysis
- Low-resource Language Translation
- Medical Knowledge Question Answering

- **Creation & Design:**

- Code Generation
- 3D Layout Generation
- Chip Design
- Table-to-Text Generation
- Art Education
- Educational Measurement and Psychometrics

- **Diagnosis:**

- Radiology Report Generation
- Electronic Health Record Diagnosis
- Sentiment Analysis
- Stochastic Processes in Statistics
- Medical Genetics and Genomics Reasoning

- **Reflection:**

- Educational Q&A
- Student Writing Improvement in Higher Education
- Medical Text Anonymization
- Social Media Analysis
- Content Summarization

These tasks reflect the complexity of real-world problems and challenge models to demonstrate not just factual knowledge, but the ability to reason through novel contexts, engage in multi-step problem-solving, and exhibit creativity. **AGI-Benchmark 1.0** resists manipulation and memorization, providing a more authentic evaluation of a model’s reasoning capabilities.

By releasing AGI-Benchmark 1.0 to the public, we aim to foster transparency, reproducibility, and collaborative progress in the pursuit of artificial general intelligence. This benchmark will be an invaluable resource for researchers and developers, guiding advancements in AI systems capable of

solving real-world, complex problems across multiple domains. This benchmark will be available at <https://github.com/UGA-CAID/AGI-Bench>.

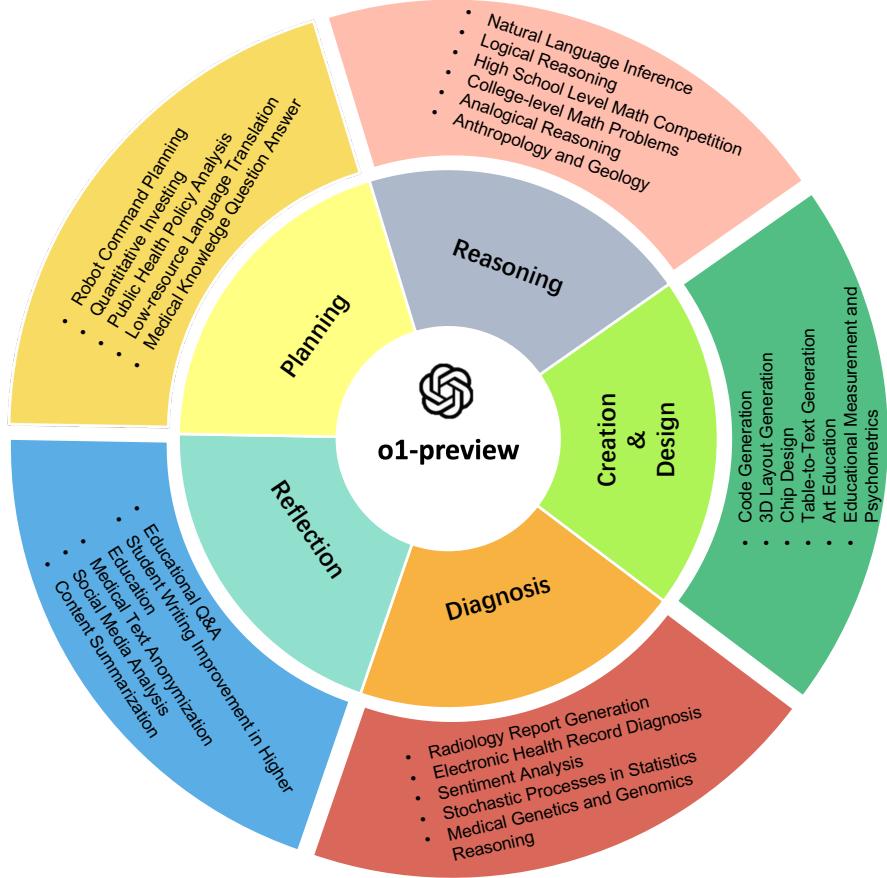


Figure 1: **Schematic Overview of the Evaluation Methodology.** This diagram illustrates the five major evaluation domains for o1-preview: Creation and Design, Planning, Reasoning, Diagnosis, and Reflection. Each domain is tested through relevant tasks. The 27 distinct tasks evaluate the model’s adaptability and effectiveness across a diverse set of cognitive and real-world challenges.

## 2 Scope of the Study and Used Public Datasets

In our study, we aim to explore and evaluate o1-preview’s capabilities and limitations across various domains and complex tasks. Below is a comprehensive list of the domains and tasks we have included in our research.

### 2.1 Code Generation

Following the initial evaluation of o1-preview’s coding capabilities, we extended our assessment by testing its performance in a competitive programming environment, specifically within Leetcode contests.

Leetcode is a widely recognized platform for coding challenges, designed to test and improve problem-solving skills in a variety of programming languages. It offers a range of problems covering topics

---

such as algorithms, data structures, dynamic programming, and system design. Leetcode’s contests are time-bound competitive events that challenge participants to solve multiple algorithmic problems of varying difficulty within a limited timeframe. These contests provide an ideal benchmark for evaluating the real-world coding abilities of language models, as they require not only syntactic correctness but also efficient problem-solving and optimization skills. Leetcode has hosted 415 weekly contests and 139 biweekly contests, each consisting of four problems designed to comprehensively cover the domain of data structures and algorithms. These contests serve as a thorough test of coding knowledge, spanning topics such as sorting, dynamic programming, graph theory, and more.

For the purpose of evaluating o1-preview, we specifically test its performance on Leetcode Weekly Contests 414 and 413, as well as Biweekly Contest 138. In each contest, o1-preview is given three submission attempts per problem. A problem was considered successfully solved if any of the submissions passed Leetcode’s automated system judgement, which evaluates the correctness and efficiency of the code against predefined test cases. This evaluation framework allows for a detailed assessment of o1-preview’s ability to reason through complex algorithmic challenges and produce accurate solutions under typical coding contest conditions.

## 2.2 Radiology Report Generation

OpenAI’s next-generation large language model, o1-preview, has exhibited considerable potential in medical report generation. To evaluate its capabilities in this domain, we performed an assessment using the Chinese radiology report dataset SXY [216, 211] to test o1-preview’s effectiveness in generating medical reports.

The SXY dataset, sourced from The Second Xiangya Hospital, Central South University, is a private Chinese radiology report dataset designed for training and evaluating models in radiology report generation. It includes 317,339 radiological reports from five categories (94,097 chest reports, 64,550 abdominal reports, 46,092 musculoskeletal reports, 69,902 head reports, and 42,698 maxillofacial & neck reports). These reports offer comprehensive documentation of patients’ radiological imaging analyses, encompassing a wide spectrum of pathological findings and providing diverse opportunities for model training and evaluation. Leveraging the SXY dataset, researchers can assess a model’s performance in producing accurate, clinically relevant radiology reports, particularly its proficiency in understanding and generating complex medical terminology. This dataset serves as a valuable benchmark for radiology report generation tasks, spanning various medical imaging domains, and provides an ideal resource for validating model performance across different clinical scenarios.

To assess the performance of o1-preview, we randomly selected 10 radiology reports from the SXY dataset for evaluation. During the testing process, we conducted several experiments to determine the optimal prompt phrasing, ensuring that the prompts remained consistent across all trials. For the model-generated outputs, we employed the ROUGE metric to measure the degree of correspondence between the radiology reports produced by the large language models and the reference reports authored by medical professionals. Specifically, this study utilized three evaluation metrics: ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L), as defined in Eq. (1).

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

---

### 2.3 Robot Command Planning

o1-preview can analyze real-time sensor data and adapts to dynamic environments, providing flexible, intelligent control solutions. Its ability to generate robot control commands and control code tailored to various robotic platforms reduces manual intervention, allowing developers to optimize control algorithms on the fly. o1-preview can potentially refine its strategies, enhancing autonomy and resilience across industrial, household, and autonomous vehicle applications.

In this section, we evaluate the performance of o1-preview on the ROS official code repository dataset, ROS [32], which contains the official ROS code’s usage. The task involves analyzing code snippets and determining their functionality and correctness, structured as a classification task where the given code either performs as expected, contains logical errors, or has undefined behavior. Code understanding tasks require advanced technical comprehension and reasoning to identify functional correctness and are widely used to evaluate AI models in software engineering contexts. Some domain-specific code datasets not only demand sophisticated reasoning but also assess the model’s understanding of domain-specific programming principles, providing a more robust evaluation of its potential for real-world development scenarios. For detailed analysis and results, please see Section 4.4.

### 2.4 Nature Language Inference

In this section, we evaluate o1-preview on the natural language inference (NLI) task. The NLI task involves determining the logical relationship between two sentences, structured as a classification task where the second sentence either logically follows from the first, contradicts the first, or is neutral (possibly true). NLI tasks require advanced language understanding and reasoning to analyze logical coherence and are widely used to evaluate LLMs. Some domain-specific NLI datasets not only demand advanced reasoning but also assess the domain-specific knowledge of LLMs, providing a more comprehensive evaluation of their potential for real-world applications.

Here, we evaluate o1-preview using data samples from five NLI datasets: MNLI, ANLI, QNLI, MedNLI, and RadQNLI [194, 125, 179, 152, 195]. Table 1 presents a summary of these datasets with detailed descriptions, covering different formats and domains, ensuring a thorough assessment of o1-preview’s reasoning capabilities.

### 2.5 Quantitative Investing

The stock-trading-QA [203] dataset offers several distinct advantages that set it apart from other financial question-and-answer datasets. First and foremost, it provides deep insights into trading strategies, financial models, and market analysis techniques, which are essential components of quantitative trading. This dataset covers a wide range of topics, from statistical models used in market forecasting to the role of automation in real-time trading, as well as how fundamental analysis can be integrated with technical signals to create more robust trading strategies.

One key advantage of this dataset is its focus on quantitative investment, a crucial aspect of modern finance. While many other financial QA datasets may address broader economic questions or general investment advice, this dataset narrows its scope to topics specifically relevant to algorithmic and quantitative trading. This allows it to dive deeper into concepts such as ARIMA models, machine learning techniques for market prediction, event-driven trading strategies, and other technical aspects that are fundamental to quantitative finance. By emphasizing these specialized topics, the dataset is highly applicable to traders, quants, and researchers focused on developing automated trading

---

Table 1: Summary of NLI datasets used for o1-preview evaluation.

| Dataset       | Size | Text Source        | Description  |
|---------------|------|--------------------|--|
| MNLI [194]    | 433k | 10 distinct genres | Determines the relationship in sentence pairs as entailment, neutral, or contradiction. Includes a broad range of diverse sources, making it ideal for evaluating model generalizability across unseen text. |
| ANLI [125]    | 169k | Wikipedia          | Employs an iterative and adversarial Human-And-Model-in-the-Loop Entailment Training process, in which annotators craft increasingly complex examples to challenge models across three rounds.               |
| QNLI [179]    | 115k | Wikipedia          | Given question-context sentence pairs with high lexical overlap, determines whether the context sentence contains the answer to the question.  |
| MedNLI [152]  | 14k  | MIMIC-III [59]     | Follows the MNLI schema, but sourced from the medical domain, specifically the MIMIC-III dataset.  |
| RadQNLI [195] | 10k  | MIMIC-CXR [58]     | Adapts the QNLI schema to the radiology domain, using data from MIMIC-CXR.   |

systems.

Another significant strength of the stock-trading-QA dataset is its ability to highlight the reasoning and numerical computation skills of models, especially those designed for quantitative reasoning. Unlike standard instruction-following datasets, which primarily test a model’s ability to comprehend and follow simple commands or guidelines, this dataset is structured to challenge models with complex, domain-specific queries that require logical reasoning, numerical understanding, and advanced financial knowledge. For example, questions about optimizing trading algorithms, statistical model selection, or implementing automation in trading systems require much more than rote instruction following—they demand quantitative problem-solving and a deep understanding of financial markets.

To rigorously test and measure the performance of models using this dataset, a comprehensive evaluation framework is employed. Models are assessed based on their ability to accurately solve quantitative problems, generate coherent and logically sound responses, and correctly apply advanced financial concepts. Key performance metrics include accuracy for classification tasks, mean squared error (MSE) and root mean squared error (RMSE) for regression analyses, and precision and recall for information retrieval tasks. Additionally, the F1-score is utilized to provide a balanced measure of a model’s precision and recall capabilities.

Beyond quantitative metrics, qualitative assessments are conducted to evaluate the models’ reasoning processes. This involves analyzing the logical flow of their solutions, the correctness of numerical computations, and the appropriateness of the financial methodologies applied. Models are also benchmarked against established financial theories and real-world market data to ensure their outputs are not only theoretically sound but also practically relevant. Human expert evaluations further validate the models’ performance by comparing their answers to those provided by experienced professionals in the field of quantitative finance.

This multifaceted evaluation approach ensures that the models are not merely performing superficial computations but are genuinely understanding and engaging with complex financial concepts. It allows researchers and practitioners to identify specific areas where models excel or need improvement, thereby facilitating the development of more advanced and reliable quantitative trading systems.

---

In summary, the stock-trading-QA dataset excels in offering insights into critical areas of quantitative finance and trading models. Its highly specialized focus on topics like statistical modeling, automation, and signal integration makes it a valuable resource for testing the abilities of models in a quantitative context. By emphasizing model reasoning and numerical computation over simple instruction-following tasks, this dataset provides a robust platform for assessing the performance of AI models in handling complex financial queries.

o1-preview’s mathematical reasoning capabilities enable it to perform complex, multi-factor, real-time quantitative model analysis, going beyond mere memorization. o1-preview can dynamically analyze numerous market variables, adjusting predictions and strategies in response to evolving conditions. This adaptability is crucial in quantitative finance, where models must factor in diverse data points such as price trends, trading volumes, and macroeconomic indicators. o1-preview’s ability to integrate these factors in real time allows for the continuous refinement of models and decision-making processes. Additionally, its capacity for multi-dimensional analysis ensures that it can identify nuanced correlations and patterns that simpler, rule-based systems might overlook, enhancing the accuracy and robustness of its quantitative trading strategies.

## 2.6 Low-Resource Language Translation

Low-resource translation is particularly challenging for Transformer-based models, like the previous GPT-4o, due to their reliance on large amounts of high-quality training data to learn language patterns effectively. Our study will evaluate o1-preview model’s ability to handle low-resource language translation by using data from the Cherokee Corpus Main Section of the Cherokee-English Dictionary (CED) project [126]. The Cherokee Corpus is a collection of 1776 Cherokee sentences paired with corresponding English translations as ground truth. This resource will serve as a valuable benchmark to evaluate how effectively the model can handle translation tasks between Cherokee and English, particularly in the context of limited parallel data availability for the Cherokee language.

Throughout the experimental phase, o1-preview generated translations and grammatical breakdowns for Cherokee sentences, focusing on both word meanings and sentence structure. The model was instructed to translate five Cherokee sentences and provide detailed analyses of key components, such as noun phrases and verb conjugations. Each analysis covered the subject, action, and contextual meaning of the sentence, demonstrating the model’s ability to handle translations for a low-resource language.

The experimental results show that o1-preview can successfully translate common phrases and identify grammatical structures such as plural nouns and verb tenses in Cherokee. However, due to Cherokee being a low-resource language, the model sometimes fails to recognize certain words, leading to incomplete or inaccurate translations. Despite this, the model is able to provide reasonable guesses for unknown words, ensuring a degree of consistency in the overall translation. This ability to infer meaning allows the model to generate plausible translations, even when faced with unfamiliar vocabulary. However, expert intervention is still often required to refine these guesses and ensure full accuracy. This indicates the need for more detailed linguistic data and expert guidance to improve the model’s performance in low-resource language translation tasks.

## 2.7 Educational Q&A

In the field of educational science, our study utilized the SciQ dataset [193], consisting of 13,679 scientifically oriented questions across various disciplines, including physics, biology, chemistry, and earth sciences. We selected the SciQ dataset to specifically evaluate o1-preview’s capabilities in

---

understanding and reasoning about scientific education knowledge. The dataset, with its wide range of scientifically oriented questions, allows us to assess the model’s ability to comprehend complex concepts, make logical inferences, and generate accurate, well-explained answers across various scientific disciplines such as physics, biology, chemistry, and earth sciences.

The results demonstrated that o1-preview performed exceptionally well on this dataset, showcasing its robust ability to select the correct answer despite the presence of distractor options. The model exhibited a strong understanding of key scientific concepts and was able to navigate through misleading distractors to consistently choose the correct answer. This performance highlights o1-preview model’s exceptional capacity for discerning relevant information and ignoring irrelevant or incorrect options, further underscoring its potential in educational applications.

In the field of education, o1-preview holds significant potential to transform the way students learn and teachers deliver instruction. By effectively understanding and reasoning through complex scientific concepts, o1-preview can assist educators in creating more personalized learning experiences, allowing students to receive tailored feedback and guidance. As technology continues to evolve, o1-preview could play a pivotal role in alleviating the workload on teachers, enabling them to focus on higher-level pedagogical tasks. Furthermore, it has the potential to contribute to a more balanced distribution of educational resources. With continued advancements, o1-preview is poised to become an indispensable tool in modern education, facilitating more efficient, scalable, and equitable learning environments.

## 2.8 Student Writing Improvement in Higher Education

This section aims to evaluate o1-preview’s potential to enhance student writing in higher education. Effective writing in higher education requires appropriate language conventions, coherent structure, rhetorical awareness, etc, which can be supported by o1-preview’s advanced language capabilities. While current LLMs predominantly focus on grammar checking, we offer a more comprehensive assessment, covering linguistic accuracy, coherence, outline generation, citation management, and creativity/personalization. The student writing samples for this evaluation are sourced from Corpus & Repository of Writing (CROW) [164], a large-scale collection of student drafts from various writing courses at three universities in the US, providing diverse scenarios and levels for a thorough analysis of o1-preview’s capabilities.

## 2.9 3D Layout Generation

In the domain of computer vision and 3D scene understanding, our study utilized the 3D-FRONT dataset [34], which comprises a large collection of high-quality 3D indoor scenes with detailed room layouts and furniture arrangements. We selected the 3D-FRONT dataset to specifically evaluate o1-preview’s capabilities in generating realistic and functional 3D room layouts. The dataset, with its extensive variety of room types, objects, and spatial configurations, allows us to assess the model’s ability to comprehend complex spatial relationships, adhere to design principles, and produce layouts that are both aesthetically pleasing and functionally sound.

The results demonstrated that o1-preview performed exceptionally well on this dataset, showcasing its robust ability to generate coherent and realistic 3D layouts despite the complexity of the scenes. The model exhibited a strong understanding of spatial constraints and design guidelines, effectively placing objects within rooms while avoiding overlaps and ensuring accessibility. This performance highlights o1-preview’s exceptional capacity for spatial reasoning and adherence to design constraints,

---

further underscoring its potential in applications such as interior design and virtual environment creation.

In the field of 3D layout generation, the o1-preview holds significant potential to transform the way virtual spaces are designed and visualized. By effectively understanding spatial relationships and design principles, o1-preview can assist designers in creating more efficient and appealing layouts, allowing for rapid prototyping and customization. As technology continues to evolve, o1-preview could play a pivotal role in automating aspects of interior design, enabling professionals to focus on more creative and complex tasks. Furthermore, it has the potential to contribute to more immersive virtual environments in gaming and virtual reality applications. With continued advancements, o1-preview is poised to become an indispensable tool in the field of 3D layout generation, facilitating more efficient, scalable, and high-quality spatial design solutions.

## 2.10 Chip Design

The intersection of LLMs and MLLMs with chip design is poised to transform the semiconductor industry, offering capabilities that surpass traditional methods in efficiency, precision, and scalability. In a field where the margins between success and failure are razor-thin, the introduction of LLMs into chip design holds profound significance. The complex workflows, intricate trade-offs, and multi-dimensional challenges inherent to chip design make it an ideal candidate for AI-driven innovation. LLMs and MLLMs, with their ability to process vast amounts of data, perform high-level reasoning, and optimize processes, are uniquely positioned to revolutionize this domain.

### 2.10.1 Transformative Potential of LLMs in Chip Design

The chip design process is a delicate balance of performance optimization, power efficiency, and manufacturability, requiring advanced techniques to manage the trade-offs between these factors. Today's semiconductor industry is tasked with creating ever-smaller, more efficient chips, all while maintaining lower costs and faster time-to-market. Traditional engineering tools and processes, while highly sophisticated, often fall short in providing the speed and depth of analysis required to stay ahead in this competitive environment. Herein lies the transformative potential of LLMs and MLLMs.

LLMs can rapidly process and analyze vast datasets, including prior chip designs, performance reports, and error logs, generating insights that might elude even the most experienced human engineers. This not only accelerates the design process but also improves its outcomes by ensuring more optimal circuit layouts, better power management, and early error detection. Moreover, MLLMs, with their ability to process multiple data forms—text, images, and simulations—offer a broader range of applications, enabling AI to assist across the full spectrum of chip design and manufacturing.

One particularly powerful application of LLMs lies in error prediction and mitigation. By analyzing historical manufacturing data, these models can identify patterns and foresee potential design flaws long before they reach the fabrication stage, reducing costs associated with defective chips. Furthermore, LLMs can also optimize the logistical supply chains in semiconductor production, minimizing material delays and ensuring timely production cycles, which are critical in an industry where time-to-market is a key differentiator.

---

### 2.10.2 Experimenting with o1-preview in Chip Design

As part of our exploration into the capabilities of LLMs, we are testing o1-preview in the domain of chip design. Given the complexity of this task and the absence of standardized chip design datasets, we draw inspiration from approaches like those found in ChipNeMo [90], which pioneered the adaptation of LLMs for chip design. However, our evaluation assess the o1-preview’s versatility and efficacy in this domain.

The test will focus on three critical areas: Engineering Assistant Chatbot, EDA Script Generation, and Bug Summary and Analysis. These tasks represent key stages in the chip design and manufacturing process and serve as proxies to gauge the model’s ability to address complex, real-world engineering challenges.

The first task, Engineering Assistant Chatbot, will evaluate the o1-preview’s ability to serve as a technical consultant for engineers, providing expert guidance on complex design questions. This test will explore whether the model can synthesize information from technical documentation, past designs, and performance metrics to deliver high-quality responses in real time. In chip design, where decisions need to be made quickly and with high precision, such a tool could significantly enhance the productivity of engineers by reducing time spent on research and troubleshooting.

In the second task, EDA Script Generation, the o1-preview will generate code for Electronic Design Automation (EDA) tools based on specific design requirements. EDA tools (i.e., specifically Verilog in this experiment) are essential for automating various steps in the design and verification of electronic systems, and generating efficient code is critical for ensuring that chips meet performance and power requirements. This phase of the experiment will evaluate the quality and efficiency of the generated scripts, testing them in real-world scenarios to compare the performance of AI-generated scripts against human-designed ones.

Finally, the Bug Summary and Analysis task will focus on the model’s ability to analyze error logs and simulation data, summarizing the root causes of design errors and suggesting potential solutions. Chip design is highly susceptible to errors, especially at the simulation stage, and early detection and resolution of bugs can save significant time and resources. By testing o1-preview’s ability to automate this process, we aim to evaluate whether the model can reduce the burden on human engineers and provide clearer, more actionable bug reports.

### 2.10.3 A Leap Toward AGI and the Future of Chip Design

The success of o1-preview model in chip design could signify a major breakthrough, not only for semiconductor engineering but also for the pursuit of AGI. o1-preview’s advanced reasoning capabilities and ability to handle complex, multi-step workflows mark it as a potential game-changer in the development of AI systems that can operate with human-level understanding in specialized domains. Chip design, with its demanding balance of technical precision and high-stakes decision-making, serves as an ideal proving ground for the model’s capabilities.

What distinguishes o1-preview from earlier models is its enhanced ability to engage with intricate, domain-specific tasks, such as circuit optimization, bug analysis, and EDA script generation. These are not just routine engineering problems; they require deep understanding, real-time analysis, and adaptive problem-solving—skills that align closely with the core requirements for AGI. By successfully performing these tasks, o1-preview demonstrates its potential to transcend narrow AI applications and begin tackling more generalized, high-level intellectual challenges.

---

Furthermore, if o1-preview can meaningfully enhance chip design processes, reducing both time and error rates while optimizing overall design performance, it will not only revolutionize the semiconductor industry but also signal a profound step forward in AI development. This success would show that o1-preview is capable of the kind of holistic problem-solving and adaptive reasoning that is critical for the realization of AGI. Unlike traditional LLMs, which often struggle with highly specialized or technical tasks, o1-preview is proving itself capable of functioning in a variety of high-complexity domains, each with its own unique set of challenges.

By bridging the gap between task-specific expertise and general-purpose reasoning, o1-preview is laying the groundwork for a new generation of AI systems capable of solving a wide range of problems across diverse fields. Its potential success in chip design would be an important milestone on the path to achieving AGI, demonstrating that AI can not only learn from vast datasets but also apply that knowledge in novel, practical, and high-stakes scenarios.

In conclusion, The introduction of o1-preview into the field of chip design marks a significant step forward in AI's potential to revolutionize this highly technical domain. By successfully handling critical tasks such as circuit layout optimization, EDA script generation, and bug summary analysis, o1-preview showcases its ability to manage the intricate balance of performance, power, and manufacturability that defines chip design.

More importantly, o1-preview's capabilities represent a major advancement toward the realization of AGI. Unlike earlier LLMs that were confined to narrow tasks, o1-preview's ability to reason through complex, technical workflows in chip design points to its potential as a general-purpose problem-solver. As it continues to demonstrate success in high-stakes, specialized environments, o1-preview is paving the way for future AI systems capable of tackling diverse and complex intellectual challenges across industries. This success will not only transform the semiconductor industry but also bring us closer to the era of AGI, where AI systems can perform at a human-like level across a wide range of tasks and domains.

## 2.11 Logical Reasoning

We attempted to explore the performance of o1-preview in the field of logical reasoning. We defined five types of logical reasoning for manual instance classification, namely categorical reasoning, sufficient condition reasoning, necessary condition reasoning, disjunctive reasoning, and conjunction reasoning. o1-preview holds numerous application advantages in the field of logical reasoning. Firstly, it can efficiently handle a large amount of data. Whether it is complex text, images, or other forms of data, it can conduct rapid analysis in an extremely short time. Moreover, it has parallel processing capabilities and can simultaneously handle multiple logical reasoning tasks, greatly enhancing work efficiency. Secondly, it has high accuracy. Unaffected by human emotions, fatigue, and subjective biases, it always maintains objective and accurate reasoning. Based on sophisticated algorithms and extensive data analysis techniques, it can establish an accurate logical model to ensure reliable reasoning results. Furthermore, it has repeatability and consistency. For the same logical problem and input data, it can generate consistent reasoning results every time. The quality is stable. No matter when and where it is used, it can maintain a high-level reasoning ability. Additionally, it has strong learning and adaptability. Through continuous improvement and optimization based on new data and feedback, it can refine the logical reasoning model, adapt to different problem types and changing situations, and can also be applied in multiple fields by learning knowledge and logical rules in different fields. Finally, it can assist human decision-making, provide objective logical analysis and suggestions for human decision-makers, help make wiser decisions, and at the same

---

time undertake some repetitive and cumbersome logical reasoning work, allowing humans to focus on more creative and strategic tasks.

Multiple choice questions from the LogiQA dataset [76] were used for testing because logical reasoning problems do not rely heavily on complex contexts. This dataset consists of 8,678 QA instances.(Train:7376; Eval:651; Test:651)The files is divided into English version: Train, Eval, Test., and Chinese version: zh\_train, zh\_eval, zh\_test.Each 8 lines constitute an example of a problem. ( $8,678 * 8 = 69,424$ )

Moreover, in the form of multiple choice questions, the average paragraph length of each sample is relatively small, and correct answers are provided. It is reported that the LogiQA dataset collects questions about logical understanding from public questions in the Chinese civil service examination, aiming to test candidates' critical thinking and problem-solving abilities. These raw data collected from the official website form 8678 question - answer data sample pairs.

The results show that o1-preview performs extremely well on this dataset, maintaining high accuracy even after multiple recalculations. When faced with interference factors in the options, it demonstrates strong anti-interference ability. Through repeated consideration of problem descriptions and attempts at different options, it always selects the correct answer in multiple tests, highlighting the model's overall excellent performance in logical reasoning instead of being limited to a certain independent type of logical reasoning problem.

## 2.12 Table-to-Text Generation

In this section, we aim to evaluate the effectiveness of o1-preview in performing table-to-text generation tasks for medical datasets. Specifically, we utilize data samples from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [132]. The objective is to determine how well the model can convert structured tabular data into comprehensive and coherent natural language descriptions, focusing on the medical domain where accuracy and clarity are paramount.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a large-scale, longitudinal study that was launched in 2004 with the primary aim of identifying biomarkers for the early detection and progression tracking of Alzheimer's disease (AD). ADNI collects a wide range of data from participants, including clinical, imaging, genetic, and biomarker information from individuals with normal cognitive function, mild cognitive impairment (MCI), and Alzheimer's disease. By leveraging neuroimaging techniques such as MRI and PET scans, ADNI seeks to deepen the understanding of the disease and support the development of new therapies. The collaborative nature of ADNI has enabled significant data sharing, which has greatly accelerated research in the field of Alzheimer's disease. An example of a patient table and its corresponding clinical description is shown in Figure 2. In this work, we harness o1-preview to transform the ADNI dataset's tabular data into fluent and accurate diagnostic reports, offering a powerful tool for medical professionals.

## 2.13 High-School-Level Math Competition

Recent efforts have extensively explored the math problem-solving abilities of the ChatGPT and its derivative models, with promising results achieved in multiple areas of mathematical reasoning and problem solving [157, 156, 136, 79]. Some studies have even examined the potential for ChatGPT to serve as a tool for teaching and learning mathematics [188]. With the key improvements in the o1-preview over previous versions of ChatGPT is its enhanced logical reasoning capability, we aim to assess o1-preview's ability to solve mathematical problems with a specific focus on its reasoning

| Participant_ID                            |             |               | clinical Description Reference  |
|---|-------------|---------------|---|
|   | Attribute   | Value         |   |
| Basic Personal information                | Age         | 72.0          | <b>Basic Personal Information:</b> Subject 098_S_0896 is a 72.0-year-old Female who has completed 15 years of education. The ethnicity is Not Hisp/Latino and race is White. Marital status is Married. Initially diagnosed as AD, as of the date 2007-10-24, the final diagnosis was Dementia.   |
|   | Sex         | Female        |   |
|   | Education   | 15            |   |
|   | Race        | White         |   |
|   | DX_bl       | AD            |   |
|   | DX          | Dementia      | <b>Biomarker Measurements:</b> The subject's genetic profile includes an ApoE4 status of 0.0...   |
|   | ...         | ...           |   |
| Biomarker measurements                    | APOE4       | 1.0           |   |
|   | TAU         | 212.5         |   |
|   | ...         | ...           | <b>Cognitive and Neurofunctional Assessments:</b> The Mini-Mental State Examination score stands at 29.0. The Clinical Dementia Rating, sum of boxes, is 1.0. ADAS 11 and 13 scores are 4.67 and 4.67 respectively, with a score of 1.0 in delayed word recall...   |
| Cognitive and neurofunctional Assessments | MMSE        | 29.0          |   |
|   | CDRSB       | 0.0           |   |
|   | ...         | ...           | <b>Volumetric Data:</b> Under MRI conditions at a field strength of 1.5 Tesla MRI Tesla, using Cross-Sectional FreeSurfer (FreeSurfer Version 4.3), the imaging data recorded includes ventricles volume at 54422.0, hippocampus volume at 6677.0, whole brain volume at 1147980.0, entorhinal cortex volume at 2782.0, fusiform gyrus volume at 19432.0, and middle temporal area volume at 24951.0. The intracranial volume measured is 1799580.0.... |
| Volumetric data                           | FLDSTRENG   | 1.5 Tesla MRI |   |
|   | Ventricles  | 84599         |   |
|   | Hippocampus | 5319          |   |
|   | ...         | ...           |   |

Figure 2: An example of a patient table and its corresponding clinical description.

skills. Mathematics, as a highly structured and logic-driven discipline, provides an ideal testing ground for evaluating this reasoning ability. To investigate o1-preview’s performance, we designed a series of tests covering various difficulty levels. We begin with high school-level math competition problems in this section, followed by college-level mathematics problems in the next section, allowing us to observe the model’s logical reasoning across varying levels of complexity.

In this section, we selected two primary areas of mathematics: algebra and counting and probability in this section. We chose these two topics because of their heavy reliance on problem-solving skills and their frequent use in assessing logical and abstract thinking [46]. The dataset used in testing is from the MATH dataset [46]. The problems in the dataset cover a wide range of subjects, including Prealgebra, Intermediate Algebra, Algebra, Geometry, Counting and Probability, Number Theory, and Precalculus. Each problem is categorized based on difficulty, ranked from level 1 to 5, according to the Art of Problem Solving (AoPS). The dataset mainly comprises problems from various high school math competitions, including the American Mathematics Competitions (AMC) 10 and 12, as well as the American Invitational Mathematics Examination (AIME), and other similar contests. Each problem comes with detailed reference solutions, allowing for a comprehensive comparison of o1-preview’s solutions.

In addition to evaluating the final answers produced by o1-preview, our analysis delves into the step-by-step reasoning process of the o1-preview’s solutions. By comparing o1-preview’s solutions with the dataset’s solutions, we assess its ability to engage in logical reasoning, handle abstract problem-solving tasks, and apply structured approaches to reach correct answers. This deeper analysis offers insights into o1-preview’s overall reasoning capabilities, using mathematics as a reliable indicator for logical and structured thought processes.

---

## 2.14 College-level Math Problems

We further investigate o1-preview’s ability to solve college-level mathematics problems. Recently, numerous studies have focused on evaluating and enhancing the ability of LLMs to solve math problems [136, 46, 83, 33, 79, 55, 119, 138, 137, 202, 85, 141]. These models have shown potential in symbolic computation, mathematical reasoning and automated theorem proving. However, there remain significant challenges in evaluating the depth of their mathematical understanding, particularly when it comes to handling problems requiring long reasoning chains, abstract concepts, and intricate symbolic manipulation. Testing o1-preview’s ability to solve such problems offers an opportunity to assess its capacity for mathematical reasoning and identify areas where it excels or struggles, providing valuable insights for future advancements in AI-driven mathematics education and problem-solving.

The problems presented in this section are manually created by the authors, drawing inspiration from various sources [133, 22, 67, 1]. They are particularly challenging due to their high level of abstraction and the complexity of reasoning required. These problems can be broadly categorized as follows:

- Basic discrete mathematics problems.
- Advanced discrete mathematics problems.
- Calculus problems.
- Proofs of advanced theorems.

We use discrete mathematics problems to evaluate o1-preview’s ability to recognize discrete relations and patterns. Calculus problems are designed to test the model’s comprehension of continuous concepts. Finally, by asking o1-preview to reproduce proofs of advanced mathematical theorems, we assess its capacity to manage abstract concepts and extended chains of reasoning.

Compared to the high school-level math problems tested in the previous subsection, college-level math problems typically involve higher levels of abstraction, more complex notations and longer reasoning chains. By challenging o1-preview with these advanced mathematical problems and analyzing its outputs in details, we can gain a deeper understanding of its strengths and weaknesses in mathematical reasoning tasks. Although the problem set is relatively small, it provides a comprehensive representation of college-level mathematics, and the results reveal several intriguing aspects of o1-preview’s intrinsic problem-solving workflow. Specifically, we empirically observed that o1-preview can easily solve basic discrete math problems. However, when confronted with more challenging problems, o1-preview tends to exhibit various types of mistakes. See the detailed discussion in Section 4.15.

## 2.15 Electronic Health Record Diagnosis

Electronic health records (EHRs) have become an integral component of modern healthcare, revolutionizing the way patient data is stored, accessed and used [63, 42, 78]. EHRs serve as digital repositories that contain comprehensive patient medical histories, ranging from clinical notes to diagnostic images, making them invaluable resources to improve patient care and medical decision making. In this article, we explore the diagnostic capabilities of o1-preview by leveraging EHR data for patient diagnosis, aiming to assess how effectively o1-preview can interpret and analyze complex medical information.

---

We use the OHSUMED dataset (Download available online<sup>\*</sup>) [48, 47] , a specialized subset of the MEDLINE database focused on biomedical literature, to evaluate the diagnostic capabilities of o1-preview. The OHSUMED dataset, curated by William Hersh and his team at Oregon Health Science University [170], was designed for interactive retrieval evaluation and research on large text collections. OHSUMED consists of medical abstracts drawn from 270 medical journals over five years (1987-1991), comprising 13,929 documents. Each document is categorized into one of 23 cardiovascular disease categories, which presents a challenge for classification due to the large number of categories. To test o1-preview, we randomly selected 10 abstracts of biomedical texts related to different types of human cardiovascular disease from the MEDLINE database as our test cases.

By presenting each EHR text to o1-preview as a prompt and asking it to provide a diagnosis, the model has demonstrated the potential to support medical decision making. Notably, it tends to perform better with shorter texts, where diagnostic accuracy is higher. In longer and more complex records, while o1-preview still identifies key patterns, significant challenges remain in evaluating its performance. In addition, we observed that o1-preview not only provides diagnostic results but also offers reasoning based on the input text. This is evident in the model's ability to generate explanations and assess the relevance of the information to specific medical categories, demonstrating its capacity for inference. Testing o1-preview's ability to analyze EHR texts provides a valuable opportunity to assess its diagnostic reasoning capabilities. By evaluating its ability to interpret complex medical data, identify relevant conditions, and deliver accurate diagnoses, we can gain crucial insights to advance AI-driven diagnostic tools and enhance their application in real-world healthcare settings.

## 2.16 Stochastic Processes in Statistics

Initial tests of o1-preview have already demonstrated promising results in mathematical domains. However, to further explore its reasoning capabilities, it is essential to evaluate its performance in statistics, a field that requires more nuanced and complex decision-making. By expanding testing into the domain of statistics, we aim to understand how well the model can handle probabilistic reasoning, which is crucial for applications in both theoretical and applied sciences.

The domain of stochastic processes was chosen for this evaluation because it is one of the most challenging areas within statistics. Stochastic processes involve systems that evolve over time under the influence of randomness, making them ideal for testing a model's ability to handle uncertainty, temporal dynamics, and predictive reasoning. Problems in this area often require deep, step-by-step thought processes, demanding an advanced capacity for reasoning and understanding of random systems, which directly aligns with o1-preview's enhanced chain-of-thought capabilities.

For this study, we selected problems from the well-known textbook "Stochastic Processes" by Sheldon Ross (Download available online<sup>†</sup>) [154], a staple resource in senior undergraduate and graduate-level courses. The problems were chosen specifically for their complexity, covering key topics such as Markov chains, Poisson processes, and renewal theory. These problems require not only mathematical precision but also a deep conceptual understanding, providing a robust test of o1-preview's statistical reasoning abilities.

This assessment is critical because it evaluates o1-preview's potential to solve complex, real-world problems where randomness and uncertainty play a central role. By comparing its performance

---

\*<https://disi.unitn.it/moschitti/corpora/ohtsumed-first-20000-docs.tar.gz>

†[https://www.academia.edu/download/56739884/stochastic-processes-ross\\_2.pdf](https://www.academia.edu/download/56739884/stochastic-processes-ross_2.pdf)

---

against human-level reasoning in these advanced topics, we can better understand the practical applicability of the model in academic and professional contexts within statistics.

## 2.17 Medical Text Anonymization

Most medical texts, such as clinical notes, are highly private because they contain a large amount of sensitive information [210]. These texts are often highly technical, context-specific, and embedded with domain-specific terminology, requiring expertise for meaningful extraction. Due to the inherent challenges of completely anonymizing these texts, researchers encounter significant obstacles in accessing valuable medical text resources, all while adhering to strict ethical rules. Leveraging LLMs as powerful automatic tools for anonymizing medical texts to expand medical data resources has become increasingly urgent in current medical LLM development.

In this section, we utilize the latest o1-preview to anonymize medical texts. We use the 2014 i2b2/UTHealth de-identification challenge dataset [166], for which a data use agreement must be signed to gain access to the dataset. This dataset has been annotated to highlight privacy features, including names, professions, locations, ages, dates, contracts, and IDs. We have implemented scripts to extract information from XML files and store them as the model’s inputs in text files.

The objective is to employ o1-preview to detect and remove all privacy-related information from the given content. Three prompt styles, ranging from coarse to fine, have been designed to evaluate o1-preview’s anonymization capabilities at different levels [97]. The diversity of the annotated dataset helps avoid bias, and varying prompt levels ensure a comprehensive assessment of the model’s performance. Based on our testing, o1-preview demonstrates an efficient ability to identify privacy features. Although performance may vary depending on the prompt style, the results remain robust, with most privacy information accurately detected and removed.

## 2.18 Social Media Analysis

Social media analysis plays a critical role in modern business and societal contexts by providing valuable insights into public opinion, market trends, and consumer behavior. With the rise of platforms like Twitter, Instagram, and Facebook, individuals and organizations generate vast amounts of data that can be analyzed to understand patterns, sentiment, and influence. Testing the capability of large language models for social media analysis is essential for improving their ability to interpret complex human language, detect trends, and understand public sentiment at scale [148]. It also plays a key role in detecting and addressing biases in social media content, ensuring more balanced and fair analyses while driving the automation of key tasks like sentiment analysis and trend forecasting.

In this section, we will test the performance of o1-preview on four tasks in the field of social media analysis based on existing benchmarks [9]: *sentiment analysis*, *irony detection*, *emotion recognition*, and *offensive language identification*. By testing o1-preview’s performance on social media analysis data, we were able to further explore its potential for text analysis in fields such as psychology and the social sciences.

For sentiment analysis, we use a dataset from SemEval-2017 Task 4 [153], which contains tweets from Twitter in English and Arabic. The task requires predicting the sentiment of tweets in three categories: *positive*, *neutral*, and *negative*. Irony detection is performed on the Semeval-2018 dataset [175]. note that this dataset includes two subtasks: subtask A is a binary irony detection task, while subtask B is a multi-class irony classification task that identifies whether the text contains a specific type of

---

irony or no irony. And we focus solely on subtask A in this paper. In short, we define this task as a binary classification task, that is, we need to predict whether a tweet is ironic or not. We use the most participated SemEval-2018, "Affect in Tweets" [122], to analyze the emotion recognition ability of o1-preview. Given the limited number of tweets with a single label, we selected the four most common emotions—*anger*, *joy*, *sadness*, and *optimism*—as our labels. As for offensive language identification, the task involves identifying whether a tweet contains some form of offensive language. Our test relies on the SemEval-2019 OffensEval dataset [204]. The above datasets are all reorganized by [9] (Publicly available<sup>†</sup>).

## 2.19 Analogical Reasoning

Recent studies have shown that LLMs can match human performance across a range of analogical reasoning tasks, particularly those involving semantic structure mapping [189, 54, 120, 74]. This capability is not limited to abstract symbol manipulation, but also includes semantically meaningful symbols, thus providing a more nuanced understanding of analogical reasoning. To evaluate the analogical reasoning ability of o1-preview, we tested its performance on a dataset specifically designed for evaluating semantic structure mapping ability. This dataset comprises 13 distinct task types and a total of 136 samples, and provides a comprehensive evaluation of the model’s analogical reasoning capability [123].

Each of the 13 task types in this dataset is uniquely challenging, ranging from simple relational analogies to more complex tasks involving multiple layers of semantic content and structure. The 136 samples were carefully selected to encompass a wide range of semantic relations, ensuring a comprehensive evaluation of the model’s analogical reasoning ability.

To test o1-preview’s analogical reasoning capability, we employed tasks that require the model to transfer semantic structure and content from one domain to another. These tasks involve identifying and mapping semantic relations between words and phrases in a manner similar to human cognition and language acquisition. The goal of this evaluation is to determine whether o1-preview can perform analogical reasoning in a human-like manner by accurately aligning the relational structure between the source and target domains.

## 2.20 Sentiment Analysis

Recent studies have demonstrated that LLMs can achieve human-level performance in various sentiment analysis tasks. This capability extends beyond simple sentiment classification to include nuanced understanding of contextual and aspect-based sentiments, providing a more comprehensive perspective on natural language processing. To evaluate the sentiment analysis ability of o1-preview, we tested its performance on three benchmark datasets: IMDB [114], SemEval-2014 Task 4 [139], and SemEval-2015 Task 12 [140]. These datasets collectively offer a diverse and thorough assessment of the model’s capabilities in sentiment analysis across different domains and text types.

The IMDB dataset comprises 50,000 movie reviews divided evenly between positive and negative sentiments. It is widely used for binary sentiment classification tasks, challenging models to accurately interpret sentiment in lengthy and varied user-generated content. This dataset tests the model’s ability to handle informal language, sarcasm, and nuanced opinions expressed in movie reviews.

SemEval-2014 Task 4 focuses on aspect-based sentiment analysis in the domains of restaurants and laptops. It requires models not only to determine the overall sentiment but also to identify sentiments

---

<sup>†</sup><https://github.com/cardiffnlp/tweeteval>

---

toward specific aspects within the text, such as service quality or battery life. This provides a more granular evaluation of the model’s ability to understand and analyze sentiments at the aspect level.

SemEval-2015 Task 12 extends the challenge by including sentiment analysis in tweets. This dataset presents unique difficulties due to the informal language, use of slang, abbreviations, and character limitations inherent in social media content. It tests the model’s robustness in handling noisy and brief text snippets, which is essential for real-world applications where data is unstructured and varied.

To assess o1-preview’s sentiment analysis capabilities, we employed tasks that require the model to classify sentiment polarity, identify aspect-specific sentiments, and interpret informal and condensed language. These tasks involve understanding and analyzing sentiments expressed in different contexts and formats, paralleling the complexity of human language understanding. The goal of this evaluation is to determine whether o1-preview can perform sentiment analysis in a human-like manner by accurately interpreting and classifying sentiments across diverse datasets and textual styles.

## 2.21 Anthropology and Geology

LLMs have demonstrated significant reasoning capabilities [39] and a wealth of accumulated knowledge [155], enabling it to address questions across various disciplines. However, certain limitations persist, such as providing overly generalized answers, lacking specificity, and ambiguity in key areas [62]. To evaluate the performance of o1-preview, we posed questions related to paleoanthropology and geology. This was done to assess its understanding of these relatively niche disciplines.

Given the scarcity of test sets in the fields of geology and paleoanthropology, and to prevent the inadvertent reuse of test sets that may have been utilized during the training phase, we collaborated with experts in these areas to create new, specialized questions. Although the final test consists of only five questions, each is crafted with considerable difficulty, aiming to evaluate o1-preview’s capabilities from multiple perspectives. Following this, the experts carefully assessed the quality of o1-preview’s responses. This approach offers a more comprehensive reflection of o1-preview’s reasoning abilities.

In the field of paleoanthropology (Figure 56), we tasked o1-preview with answering questions related to ancient human genetics, including issues such as genetic contamination, genetic changes over time, and exploratory questions. For example, we posed a question about how discovering ancient human hair could be used to study various aspects of their civilization, living environment, and migration patterns. o1-preview’s responses closely resembled those of an industry expert, progressing from broad conceptual explanations to more specific and in-depth answers. The model also demonstrated the ability to refine its responses based on interactive input, continuously deepening its understanding of the questions and providing increasingly precise and professional answers. This capability is remarkably similar to engaging with a domain expert, as it handles questions of varying complexity with ease, showcasing an impressive level of expertise.

To prevent the o1-preview model from having already seen the geology dataset during training, we had experts create new questions for o1-preview to answer. The experts then evaluated these responses to test the model’s true capabilities. In the field of geology (Figure 60), we examined o1-preview’s ability to respond to true/false questions by describing a particular rock phenomenon and asking about its origin. Not only did o1-preview correctly identify the cause of the phenomenon, but it also provided supporting evidence, analyzing the observed features in a manner akin to that of a geologist. The model demonstrated the capacity to approach such questions with expert-level

---

insight, engaging in in-depth discussions on topics like geology or paleoanthropology, much like a specialist in the field. This ability to reason and converse on complex scientific matters showcases its potential for facilitating expert-level discourse.

## 2.22 Educational Measurement and Psychometrics

Educational measurement and psychometrics refer to the science and practice of developing the most appropriate test to accurately measure human characteristics, such as student knowledge, skills, abilities, and interests [7, 64]. Since these characteristics cannot be directly measured, researchers often rely on samples of behaviors, such as test responses or observations, to make inferences about these characteristics. Specifically, measurement involves assigning numerical values to observed events based on predefined rules. For example, in a test, each correct response may be assigned one point, and the total score represents the sum of these points that serves as a measure of the knowledge of the student [64]. Measurement in social sciences involves developing valid and reliable instruments, designing effective data collection procedures, analyzing and evaluating the quality of the data, and communicating the results to multiple stakeholders, such as educators, policymakers, students, and parents. The key goals of educational measurement are (1) to draw valid conclusions about the abilities, achievements, or attitudes of the students, (2) to assess progress towards educational objectives, and (3) to improve teaching and learning [64].

Therefore, educational measurement and psychometrics is an interdisciplinary field closely intertwined with education, psychology, and statistics [7]. However, it faces notable challenges, particularly due to limited research and development compared to fields like computer science. This scarcity is due in part to sluggish enrollment and slow growth in graduate education programs. From 2007 to 2016, the number of conferred doctoral degrees in areas like "educational evaluation and research," "educational statistics and research methods," and "educational assessment, testing, and measurement" ranged from a low of 128 in 2012 to a high of 157 in 2015 [73]. This restricted academic growth highlights potential limitations in training datasets for large language models. As such, exploring the performance of LLMs like o1-preview in the context of educational measurement and psychometrics becomes increasingly valuable for the development of the field. Notably, the dataset we used for this task was selected from multiple representative quiz questions in an introduction-level course named *Measurement Theory* based on Bandalo's textbook [7] for first-year graduate students. This dataset was built by content experts who had more than 10 years of research and teaching experience for a graduate-level course at James Madison University.

## 2.23 Public Health Policy Analysis

The intersection of public health and LLMs represents a unparalleled opportunity to enhance public health surveillance, healthcare administration, and health policy-making. In this section, we assess o1-preview's ability to analyze and evaluate the Affordable Care Act (ACA), one of the most significant healthcare reforms in recent U.S. history.

Our focus is on the significant aspects of the ACA, such as the expansion of insurance coverage, improvements in access to care, and the broader public health impacts of these changes. The evaluation follows the Q&A format, where the prompts derived from the article named The Affordable Care Act at 10 Years: Evaluating the Evidence and Navigating an Uncertain Future[168]. It is important to note that the dataset used for this task is limited in size and scope, consisting of key questions focused on ACA findings. Thus, our evaluation emphasizes depth of reasoning, factual accuracy, and consistency in how the model addresses complex health policy. For further evaluations,

---

we suggest using larger datasets, such as the ACA state dataset[44], to provide more comprehensive training for the model.

The evaluation involved 10 distinct prompts, each addressing critical areas of the ACA’s public health impact, including insurance coverage expansion, Medicaid expansion, surgical outcomes, preventive services, and healthcare disparities. These prompts challenged o1-preview to demonstrate nuanced reasoning and provide answers based on real-world health outcomes. The generated responses compared against expert insights from the article to determine their accuracy, depth, and relevance.

Overall, this study determine how effectively o1-preview can process and reason about health policy, particularly its performance in analyzing the ACA. Although the current dataset size is limited, this evaluation offers insight into o1-preview’s great potential for policy analysis and its capability to support public health decision-making processes.

## 2.24 Medical Genetics and Genomics Reasoning

Genetics and diseases are closely related, and understanding the cause-and-effect relationship between them is crucial for the biomedical field. From a view of the knowledge base, the Inferential information can be used to establish deep-learning frameworks to reason the mutual relationship between genes and proteins. Several studies utilize the Gene Ontology (GO) knowledgebase to design neural network architectures that simulate gene or protein interactions within cells [4, 201, 113]. At the same time, using the bio-text data along with a large language model approach to infer the relationship between genomics and biomedical entities has become more prevalent. For example, A precise GPT model can significantly aid genomics research by minimizing the time interdisciplinary researchers, who may have limited genomics expertise, spend searching for and retrieving information from reliable sources. GeneGPT [57] use a language model as an agent to connect NCBI Web APIs, which use in-context learning and an augmented decoding algorithm to conduct automatic API calls. Recent work, GP-GPT [111] exhibits proficiency in accurately retrieving medical genetics information and executing common genomics analysis tasks, including information retrieval and relationship determination.

In this section, we seek to evaluate the model’s ability to reason through problems related to medical genetics and genomics. To assess the model’s genomics and genetics medical reasoning ability, we designed a set of experiments that focused on genomics questions and answers, where the model was required to generate reasonable answers towards the pre-defined genomics questions. These questions are about the relationship predictions and corresponding explanations. The tasks were created to test various aspects of genomics reasoning using a dataset comprising 20 question-and-answer (QA) tasks from the GenTuring benchmarks dataset [52]. GeneTuring is a comprehensive QA database used to assess the performance of GPT models in genomics (Methods), which is based on bio-texts. It consists of twelve modules, encompassing a total of 600 question-answer pairs, which can be categorized into four main groups. These modules reflect tasks commonly encountered in genomics research. The experimental tasks were designed based on the GeneTuring benchmark. The delicate selected gene-disease-related QA terms from the GeneTuring database have been reviewed and significantly extended by manually adding the corresponding gene/disease information from OMIM [5]. One extended term is utilized to build one full testing instance, which contains the instruction prompt, the gene’s description and the disease’s description. The model needed to follow the prompt instructions to respond to the gene-disease questions, in the meantime, the model was asked to give inference details and explanations step by step provided with the extended information.

---

## 2.25 Medical Knowledge Question Answer

In this section, we evaluate o1-preview’s performance on the Medical Knowledge question answer (QA). Medical knowledge QA involves tackling real-world examination datasets from complex medical disciplines, including pharmacology, medicine, and surgery. These tasks demand a deep and comprehensive understanding of the field, requiring knowledge that is comparable to that of human experts. Such QA systems must not only grasp foundational medical concepts but also integrate interdisciplinary knowledge, clinical reasoning, and problem-solving skills, making them particularly challenging in medical AI development.

In this section, we will utilize the MedMCQA dataset [131] to assess o1-preview’s capability in answering medical knowledge questions. MedMCQA is a newly developed, large-scale Multiple-Choice Question Answering (MCQA) dataset, specifically designed to tackle real-world medical entrance exam questions. The dataset comprises 194,000 high-quality multiple-choice questions (MCQs) from the medical domain, covering 2,400 healthcare topics and 21 distinct medical subjects, making it a robust and diverse benchmark for evaluation. Each question is accompanied by the correct answer(s), alternative options, and additional auxiliary data, most notably detailed explanations of the solutions. This rich dataset allows for an in-depth assessment of model performance in the medical field.

Specifically, we extracted 10 questions from the MedMCQA dataset, covering various medical knowledge areas such as Anatomy, Pathology, Pharmacology, and more. These questions are categorized into two levels of difficulty: easy and difficult. Easy questions typically consist of straightforward knowledge-based queries, while difficult questions require scenario-based inference (see Section 4.26 for more details). All questions are multiple-choice and include relevant explanations, allowing us to directly assess the accuracy of o1-preview’s responses and the validity of its explanations. We will analyze the model’s selected options and explanations to provide a systematic and comprehensive evaluation.

## 2.26 Art Education

This study evaluates the performance of the o1-preview model in art education, focusing on its ability to assist art educators, contribute to curriculum theory, and support educational standards. By engaging with diverse prompts—including art lesson development, creative writing exercises, reflections on artistic processes, and explanations of key educational concepts—the research examines the model’s capacity to assist in both practical teaching strategies and deeper curriculum design.

We develop our testing data from a set of prompts and scenarios covering tasks such as curriculum development, lesson planning, artistic writing, and engagement with educational theories. The questions were structured around real-world classroom challenges and reflective exercises, such as designing a cardboard assemblage art activity [109], creating a dramatic monologue from the perspective of a non-human object [15], and exploring high school students’ identities under the influence of artist Nick Cave’s *Soundsuits* [24]. Key aspects of educational theory, including William Pinar’s concept of *currere* [134], inclusive education, such as cross-disciplinary connections with disability studies [25], were incorporated to test the model’s ability to engage with both practical strategies and theoretical concepts.

Grounded in educational theory, particularly Pinar’s *currere*[134], the study explores how o1-preview can contribute to reflective learning by interpreting and explaining complex ideas within curriculum theory and educational practice. This involves defining concepts like *currere* and inclusive education

---

and adapting them to specific educational contexts, requiring the model to comprehend theoretical ideas and apply them in practical scenarios such as art activity planning and student engagement.

The research also examines the model’s ability to support reflective practices, encouraging students and educators to consider how personal experiences shape their understanding of art and culture. It analyzes o1-preview’s capacity to assist art educators in setting objectives, aligning lessons with educational standards, and engaging students creatively. Furthermore, the study assesses whether the model can provide coherent, contextually relevant responses that support comprehensive curriculum development in art education, including suggestions on representation, inclusivity, and interdisciplinary connections to improve instructional quality.

Overall, the study assesses o1-preview’s potential as a tool for art educators in guiding artistic writing, planning lessons, and fostering inclusive and reflective learning environments. It explores whether the model can enhance curricula by contributing to personalized, culturally responsive, and critically reflective educational practices.

## 2.27 Content Summarization

Automatic summarization, one of the central problems in Natural Language Processing, is the process of creating a short, concise summary of a longer text document. The goal is to distill the most important information from the original text into a shorter version that retains the essence of the content. It obtains high requirements for language models in context relevance, text content understanding and text generation.

In this section, we evaluate o1-preview’s performance on the Journalism Content Summarizing task, which focuses on summarizing news articles in one sentence. We use a dataset called *XSum*, from the article named *Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization* [124]. This dataset consists of BBC articles and accompanying one single sentence summary. It includes 226,711 articles ranging from 2010 to 2018 and covers a wide range of categories (including sports, technology, business, family, art, health, education, etc.) [124].

In this section, we design several considerable judgment points for testing o1-preview’s performance on textual content summarization. One is whether the summary’s content is consistent with the original text. Another is whether the summary generated by the model is too detailed or too general. Moreover, the length of the summary is also considered in our research.

## 3 Related Work

### 3.1 Foundation Models

The Transformer architecture [176] has quickly become a foundational framework for both language and vision-based models. First introduced in the field of Natural Language Processing (NLP), it revolutionized the handling of long sequences through the use of a self-attention mechanism, setting new benchmarks in machine translation. This architectural breakthrough laid the groundwork for models like BERT and GPT [26, 143], which significantly improved performance across a wide range of NLP tasks. As datasets expanded and models grew more complex, Transformer-based models such as RoBERTa, GPT-2, and GPT-3 [94, 144, 14] emerged, further pushing the limits of what was possible in natural language understanding.

---

OpenAI’s ChatGPT and GPT-4 [128, 127] have been particularly noteworthy, demonstrating impressive language comprehension and reasoning capabilities. These models, thanks to vast training data and techniques like Reinforcement Learning from Human Feedback (RLHF) [130], perform effectively on a variety of tasks without the need for specific fine-tuning. Meanwhile, open-source models such as LLaMA and Mistral [173, 56] have also shown strong performance, contributing to ongoing research into fine-tuning large language models for specialized domains, including medicine, education, robotics, and more [104, 182].

The success of Transformers in NLP has naturally extended to the realm of computer vision. The Vision Transformer (ViT) [27] adapted the architecture for image processing, and subsequent models like DeiT, Swin Transformer, and Masked Autoencoders (MAE) [174, 95, 43] have driven further advancements in visual tasks. As with NLP, larger datasets and increased parameters have enabled the development of even more powerful models like ViT-22B [23], leading to significant improvements in image recognition and analysis.

### 3.2 Prompt Engineering

With the rapid advancement of artificial intelligence technology, particularly in the field of natural language processing, the applications of large language models, such as OpenAI’s GPT series models [14], are becoming increasingly widespread. These models have demonstrated significant capabilities in tasks such as text generation, translation, and dialogue systems. However, effectively guiding these models to generate outputs that meet expectations and avoiding erroneous or inappropriate content has become a pressing challenge. prompt engineering has emerged in response to this need. Its goal is to maximize the performance and practical value of LLMs by designing precise and effective input prompts that guide the models toward generating desired results.

The importance of prompt engineering is reflected on multiple levels. First, it serves as a crucial link between user needs and model capabilities. While LLMs possess powerful generative abilities, they may also produce content that deviates from the topic or contains bias [159]. Through carefully designed prompts, users can better control the direction and quality of the model’s outputs. Additionally, in the context of few-shot learning and zero-shot learning, prompt engineering can significantly enhance model performance [36]. When large amounts of labeled data are lacking, well-crafted prompts can tap into the model’s latent knowledge, enabling it to perform satisfactorily on new tasks. Furthermore, prompt engineering helps reduce model biases and errors, improving the accuracy and reliability of generated content, thereby enhancing user trust in the model [10].

There are two main approaches to prompt engineering: manual design and automated optimization. Manual prompt design is the most intuitive and commonly used method. Researchers or practitioners craft appropriate prompts based on their understanding of task requirements and model characteristics to guide the model in generating the desired output [144]. For instance, in a question-answering system, including clear questions and necessary contextual information in the prompt can improve the accuracy of the model’s responses. The advantage of this method lies in its flexibility, allowing designers to make adjustments according to specific situations. However, its disadvantage is the significant time and effort required, and the quality of the design heavily depends on the expertise and experience of the designer.

To improve efficiency and effectiveness, automated prompt optimization methods have gradually gained traction. Among these, Prompt Tuning [72] and Prefix Tuning [81] are representative techniques. Prompt Tuning adds trainable prompt vectors before the model input, enabling fine control over model behavior with minimal parameter adjustment. Prefix Tuning, on the other hand,

---

inserts trainable prefixes into the model’s hidden layers to influence its internal representations and generation process. The common feature of these methods is that they achieve excellent performance on specific tasks by fine-tuning only a small portion of the model’s parameters, avoiding the need for large-scale model retraining and saving computational resources.

Despite the significant results achieved in practice, prompt engineering still faces several challenges. First, there is the issue of interpretability and transparency. Automated prompts often involve high-dimensional continuous vectors, making it difficult to intuitively understand their specific impact on model outputs [142]. Secondly, the generalizability of prompts across different tasks and domains is limited. A prompt effective for a specific task may not be directly applicable to other tasks, requiring redesign or adjustment [178]. Moreover, controlling model bias and undesirable outputs, while avoiding the introduction of subjective or discriminatory content in prompts, is also an important area of research [86].

Looking ahead, prompt engineering is expected to integrate with other techniques, such as human-computer interaction and reinforcement learning, to further improve the efficiency and effectiveness of prompt design. At the same time, developing more intelligent prompt generation tools that leverage the model’s own capabilities to assist in prompt design is a promising direction. Additionally, strengthening the theoretical foundation of prompt engineering and gaining a deeper understanding of the relationship between prompts and model behavior will help elevate the level of research in this field. As a critical component in the application of large language models, prompt engineering holds significant theoretical and practical value. Through the careful design and optimization of prompts, models can be effectively guided to generate high-quality, expectation-aligned content that meets diverse task requirements. Although challenges remain, continued research and technological development will undoubtedly enable prompt engineering to play an increasingly important role, leading to new breakthroughs and innovations in the field of AI.

### 3.3 Chain of Thought

Chain-of-Thought (CoT) prompting is a powerful technique that enhances the reasoning capabilities of large language models. Specifically, CoT prompting enables models to solve complex problems by generating intermediate reasoning steps instead of directly providing the final answer. This approach is particularly effective for multi-step tasks, such as mathematical problem-solving and logical reasoning [191]. CoT aims to overcome the limitations of traditional prompting methods. In traditional prompting, models are typically given input-output examples and are expected to produce a final answer directly. However, this method often falls short for tasks that require multi-step reasoning [146]. CoT addresses this issue by instructing the model to explicitly articulate each step in the reasoning process.

Following the initial work on CoT, subsequent research on annotation methods can be categorized into manual, automatic, and semi-automatic approaches [19]. Manual CoT relies on human-crafted prompts to guide the reasoning process. For instance, the Prompting-based Arithmetic Learner (PAL) framework uses programming-language rationales to generate programs, manually constructing prompts for arithmetic tasks [35]. On the other hand, automatic prompting techniques like zero-shot CoT and Auto-CoT generate reasoning steps automatically [61, 212]. Some research has focused on semi-automatic methods that combine automatic generation with limited human supervision. For example, AutoMate CoT [163] and BoostedPrompt [135] use a small set of human-annotated rationales to bootstrap the process, enabling the generation of higher-quality reasoning chains even in the absence of extensive labeled data.

---

Beyond basic CoT prompting, various extensions have been developed to tackle specific tasks or challenges. For instance, techniques like Self-Refine [115] and CoSelfRevisions [69] allow models to revise or verify their reasoning processes as they generate solutions. Other advanced frameworks, such as Tree of Thought (ToT)[200] and Graph of Thought (GoT)[12], further extend the CoT structure by exploring multiple reasoning paths or incorporating network-based structures.

### 3.4 Multi-modal Large Language Models

Recent years have witnessed remarkable advancements in multi-modal large language models (MLLMs), which have demonstrated unprecedented capabilities in understanding and generating content across various modalities. These models have significantly impacted numerous domains, including computer vision, natural language processing, and human-computer interaction.

The evolution of MLLMs can be traced back to early work on combining visual and textual information. [110] proposed ViLBERT, one of the first models to jointly process visual and linguistic inputs using a two-stream architecture. Building upon this foundation, [84] introduced Oscar, which utilized object tags as anchor points to align image regions with text tokens, significantly improving vision-language pre-training. These early efforts laid the groundwork for more sophisticated multi-modal models. A significant breakthrough came with the introduction of CLIP by [145], which demonstrated the effectiveness of contrastive learning in aligning visual and textual representations at scale. This work not only achieved impressive zero-shot performance across various tasks but also paved the way for subsequent research into scaling multi-modal models to billions of parameters. Following this trend, [3] presented Flamingo, a visual language model capable of few-shot learning across a wide range of vision and language tasks. The model's architecture, combining a vision encoder with a language model through cross-attention layers, has become a blueprint for many subsequent MLLMs. As the field progressed, researchers began to focus on enhancing the cognitive abilities of MLLMs. In 2023, [80] introduced CogVLM, a large vision-language model that achieves strong performance across various vision-language tasks while maintaining high efficiency. The model's cognitive training approach enables it to handle complex reasoning tasks involving both visual and textual inputs, representing a significant step forward in multi-modal understanding. Concurrently, efforts were made to improve the versatility and applicability of MLLMs. [208] proposed mPLUG-Owl, an MLLM that demonstrates impressive capabilities in multi-modal dialogue and image editing. The model's unique architecture allows for efficient fine-tuning and adaptation to downstream tasks, making it particularly valuable for real-world applications. This work highlighted the potential of MLLMs to bridge the gap between research and practical implementation. More recently, [17] introduced MM1, a multi-modal model that achieves state-of-the-art performance on a wide range of vision and language tasks. MM1's novel pre-training strategy, which incorporates large-scale web-crawled image-text pairs, enables the model to develop a deep understanding of the relationships between visual and textual content. This research underscores the importance of diverse and extensive training data in developing robust MLLMs. As the capabilities of MLLMs continue to expand, researchers have increasingly focused on addressing key challenges in the field. One such challenge is the need for more efficient architectures and compression techniques, given the growing size of these models. [187] made significant strides in this direction by proposing a novel pruning method that substantially reduces the computational requirements of MLLMs without sacrificing performance. This work exemplifies the ongoing efforts to make MLLMs more accessible and deployable in resource-constrained environments. Another important trend in MLLM research is the emphasis on enhancing multi-modal reasoning capabilities. [198] contributed to this area by introducing a new benchmark for evaluating complex reasoning tasks that require the integration

---

of visual and textual information. This work not only provides valuable tools for assessing MLLM performance but also highlights the increasing sophistication of tasks that these models are expected to handle. As MLLMs become more powerful and widely adopted, the research community has also turned its attention to addressing ethical concerns such as bias, fairness, and privacy. [60] made a significant contribution in this regard by proposing a framework for evaluating and mitigating biases in multi-modal models. This research underscores the importance of responsible AI development and the need to consider the societal implications of advancing MLLM technology.

In conclusion, the field of multi-modal large language models has seen rapid and multifaceted progress in recent years. From foundational architectures to state-of-the-art models with sophisticated reasoning capabilities, MLLMs have evolved to become powerful tools for understanding and generating content across multiple modalities. As research continues to advance, we can expect further improvements in model performance, efficiency, and ethical considerations. These developments promise to pave the way for more powerful, versatile, and responsible AI systems that can seamlessly integrate information from diverse sources and modalities.

### 3.5 Fine-tuning Large Language Models

Fine-tuning in deep learning adapts pre-trained models to specific tasks by adjusting their parameters, offering significant advantages over training from scratch. This technique not only improves model performance while reducing training time and data requirements, but also allows for specialization in targeted domains. There are two main approaches: full parameter fine-tuning, which adjusts all model parameters, and parameter-efficient fine-tuning, which modifies only a subset. Fine-tuning is particularly beneficial for specialized tasks, enabling models to learn domain-specific vocabulary and contextual nuances crucial in fields like medicine or law. It enhances the model's ability to produce structured outputs consistently, adhere to specific guidelines, and consider ethical implications relevant to particular applications. Whether through full or parameter-efficient methods, fine-tuning is essential for tailoring large language models to specialized use cases, significantly improving their accuracy, reliability, and practical applicability in targeted domains while balancing performance and efficiency.

Full parameter fine-tuning [209] is a comprehensive adaptation method for pre-trained models to new tasks. In this process, every layer and parameter of the model is updated and optimized, allowing the model to better learn the features of the new task and improve its performance. However, due to the large number of parameters to be adjusted, full parameter fine-tuning often requires significant computational resources and time. More crucially, it can lead to "catastrophic forgetting" [89], where the model forgets the knowledge it has previously learned when learning new tasks. This phenomenon is particularly prominent when the model is large and the tasks are highly dissimilar. Therefore, full parameter fine-tuning is generally suitable for scenarios with sufficient data, abundant computational resources, and a strong correlation between the new and old tasks. To mitigate catastrophic forgetting, techniques such as parameter-efficient fine-tuning and knowledge distillation can be considered.

Parameter-efficient fine-tuning [41] is a technique that adapts large pre-trained models to new tasks without significantly increasing the number of model parameters. The core idea is to achieve transfer learning across different tasks by fine-tuning specific parts of the model or introducing additional modules while maintaining model efficiency. Common parameter-efficient fine-tuning methods include partial fine-tuning, adapters, prompt tuning, prefix tuning, and LORA [53]. Partial fine-tuning only adjusts the top or a few layers of the model, preserving the general knowledge

---

acquired during pre-training; adapters insert small adapter modules into each layer of the model and only train the parameters of these modules; prompt tuning guides the model to generate the desired output by adding specific prompts to the input text; prefix tuning adds learnable prefixes to the beginning of the input sequence and controls the model’s output by adjusting these prefixes; LoRA performs low-rank decomposition of the model’s attention matrices and only trains the decomposed matrices. The advantages of parameter-efficient fine-tuning include reduced computational cost, faster training, alleviated overfitting, and improved generalization. The choice of parameter-efficient fine-tuning method depends on the specific task, dataset, and available computational resources.

### 3.6 Large Language Model Agent and Retrieval-Augmented Generation

Artificial Intelligence agents are autonomous entities capable of perceiving their environment, reasoning about it, and executing actions to achieve specific goals. They have made significant advancements in various fields, including robotics, finance, and healthcare. In robotics, AI agents have enabled autonomous navigation and manipulation, exemplified by the development of self-driving vehicles in recent years [6]. Advancements in deep reinforcement learning have led to AI agents capable of complex decision-making in dynamic environments [177]. These agents operate without continuous human intervention, leading to increased efficiency in tasks such as data analysis and real-time decision-making. They learn from interactions with their environment, allowing them to adapt to new situations and improve performance over time. The development of model-based reinforcement learning algorithms has been crucial in this regard [185]. However, designing AI agents involves complex algorithms and substantial computational resources, which can be cost-prohibitive. Many AI agents are task-specific and struggle to generalize knowledge to different domains. Additionally, autonomous decision-making raises issues related to accountability, transparency, and potential biases encoded in the agents. Recent studies have highlighted that AI systems can perpetuate existing biases, leading to unfair outcomes in applications such as hiring and lending [117]. In finance, AI agents are used in algorithmic trading, risk assessment, and fraud detection, improving efficiency and accuracy. Research has demonstrated that AI-driven high-frequency trading can influence market dynamics and liquidity [8]. In healthcare, AI agents assist in patient monitoring, diagnostic support, and personalized treatment planning. Deep learning models have been developed for medical image analysis, aiding in the early detection of diseases such as cancer [77].

Retrieval-augmented generation (RAG) combines large language models with information retrieval systems to generate responses grounded in external data sources. This approach enhances the factual accuracy and relevance of generated content. By accessing up-to-date information during generation, RAG models produce more accurate and contextually relevant outputs, overcoming the limitations of static training data. Lewis et al. [75] introduced the RAG model, which integrates a pre-trained language model with a neural retriever to improve performance on knowledge-intensive tasks. Subsequent research has built upon this framework to enhance the capabilities of language models. For example, Guu et al. [40] proposed REALM, which enables language models to retrieve and reason over large corpora. The effectiveness of RAG models heavily relies on the quality and reliability of the retrieved data. The retrieval process adds latency and requires additional computational resources. Integrating retrieval mechanisms with generative models introduces challenges in system design and optimization. Despite these challenges, RAG models have been successfully applied in various domains. In customer support, they power intelligent chatbots that provide accurate and context-specific assistance to users [31]. In education, they generate personalized learning content by retrieving relevant materials tailored to individual learner needs [29].

Despite the advancements brought by AI agents and RAG, limitations persist. AI agents often lack

---

the ability to generalize across tasks, and RAG models can be constrained by the quality of their data sources and computational demands. The new o1-preview model emerges as a response to these challenges, aiming to combine the autonomy of AI agents with the knowledge integration of RAG while mitigating their respective drawbacks. The o1-preview model is anticipated to enhance generalization by leveraging advanced learning architectures, enabling effective performance across diverse tasks without extensive retraining. It is designed to optimize performance by reducing computational overhead, allowing faster inference even when accessing external data. Additionally, it aims to improve reliability by incorporating robust mechanisms for data verification to ensure that outputs are accurate and trustworthy. The o1-preview model holds promise in advancing the field of AI by addressing the limitations of existing approaches and providing a more versatile and reliable framework for autonomous agents and knowledge integration.

### 3.7 Large Language Models & Reasoning

In recent years, large language models have demonstrated remarkable capabilities in natural language understanding and generation. However, their proficiency in complex reasoning tasks remains limited. Challenges arise particularly when dealing with multi-step logical deductions, abstract reasoning, and the integration of knowledge across various domains. Enhancing the reasoning capabilities of LLMs has thus become a crucial objective in artificial intelligence research. Based on the training mechanisms and cognitive frameworks of LLMs, their deficiencies in reasoning are evident. Specifically, addressing intricate problems necessitates deliberate and meticulous contemplation. However, due to the limited guidance provided during the training phase, LLMs are often constrained by token-level, left-to-right decision-making processes during inference [11]. When faced with complex reasoning challenges, humans typically employ diverse cognitive abilities and interact with tools, external knowledge, and environmental information to accomplish tasks. Replicating this level of cognitive function in LLMs remains a significant challenge [65].

In response to these limitations, several approaches have been proposed to emulate human reasoning processes. One foundational method is the *Chain-of-Thought* (CoT) prompting [191], which involves providing LLMs with examples that include detailed intermediate reasoning steps. By explicitly modeling the reasoning process, CoT enables LLMs to tackle arithmetic, commonsense reasoning, and problem-solving tasks more effectively. Building upon CoT, the *Tree-of-Thought* (ToT) framework [200] models the reasoning process as a search through a tree of possible thought sequences. This approach allows the model to explore multiple reasoning paths simultaneously, enhancing its ability to handle ambiguous or complex tasks by considering a broader range of possibilities. The *Self-Consistency* technique [186] improves reasoning by generating multiple reasoning paths and selecting the most consistent answer across these paths. This method reduces the impact of reasoning errors in any single chain and improves overall answer accuracy.

Another significant advancement is the integration of external tools into the reasoning process. The *ReAct* framework [199] combines reasoning and acting by enabling LLMs to interact with external environments and tools, such as calculators or knowledge bases. This integration allows LLMs to perform computations, retrieve up-to-date information, and validate reasoning steps, thereby enhancing factual accuracy and reducing hallucinations. By leveraging the strengths of both internal reasoning processes and external resources, ReAct represents a promising approach to bridging the gap between LLMs and human-like reasoning capabilities.

The evolution of reasoning methods in LLMs reflects a shift from implicit, data-driven learning to approaches that incorporate explicit reasoning processes. Early LLMs relied on large-scale

---

pre-training, hoping that reasoning abilities would emerge from vast amounts of data. However, this proved insufficient for tasks requiring logical deductions and complex problem-solving. The introduction of prompt engineering techniques, such as CoT, marked a significant advancement by leveraging the models' capabilities through carefully designed prompts that guide the reasoning process. Subsequent methods, such as ToT and Self-Consistency, focused on exploring multiple reasoning paths and selecting the most consistent answers, further enhancing the models' ability to handle ambiguous or complex tasks.

The integration of external tools and environments, as exemplified by the ReAct framework, represents a crucial step towards emulating human-like reasoning in LLMs. By combining internal reasoning processes with the ability to interact with external resources, these approaches more closely mimic the diverse cognitive abilities employed by humans when faced with complex challenges. As research in this field continues to progress, we can expect further advancements in the development of LLMs that exhibit increasingly sophisticated reasoning capabilities, bringing us closer to the goal of achieving truly intelligent and versatile artificial intelligence systems.

### 3.8 Reinforcement Learning with Human Feedback

Generating responses that satisfy human expectations is a critical challenge in the development of large language models. Reinforcement Learning from Human Feedback [18, 218] has emerged as a promising solution to this challenge, enabling the creation of some of the most widely deployed AI models to date.

The development of RLHF can be traced back to the work of Schulman et al. [158], who proposed a novel reinforcement learning approach that indirectly learns reward functions based on human preferences. This method simplified human involvement by having humans qualitatively compare the quality of two trajectories, achieving good results in simulated robotics and Atari games. In the same year, OpenAI introduced the Proximal Policy Optimization (PPO) algorithm [18], which has been widely applied in the RLHF stage of LLMs. PPO optimizes the policy function to maximize cumulative returns while maintaining similarity between new and old policies, effectively avoiding instability and high variance problems during training.

Stiennon et al. [165] further advanced RLHF by replacing automated evaluation methods with human preferences as the training objective. Starting from an initial model fine-tuned with Supervised Fine-Tuning (SFT), they collected samples through human comparisons, built a reward model, and then optimized using reinforcement learning strategies, significantly improving the quality of summaries.

A significant milestone was reached with InstructGPT [130], which combined RLHF with GPT-3. This approach fine-tuned GPT-3 through human feedback to improve the model's ability to follow user intent, achieving performance improvements while reducing the number of parameters.

In the context of RLHF, we define:

- **States ( $s$ ):** The current context or prompt.
- **Actions ( $a$ ):** The model's generated responses.
- **Reward ( $r$ ):** Feedback based on human preferences.
- **Policy ( $\pi$ ):** The strategy the model uses to generate responses.

The standard RLHF fine-tuning process typically consists of three stages:

- **Supervised Fine-Tuning (SFT)**: This initial stage adapts the pre-trained language model to generate outputs that align with human-written examples of desired behavior.
- **Reward Modeling (RM)**: A reward model is trained on pairs of model outputs, where humans have indicated a preference for one output over the other. This model learns to predict these preferences, enabling it to generate reward signals for new, unseen model outputs.
- **Reinforcement Learning (RL)**: The language model is fine-tuned using reinforcement learning to maximize the rewards predicted by the reward model.

Within this framework, various algorithms have been developed to implement and optimize the reinforcement learning process. Two prominent approaches are Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO):

#### **Proximal Policy Optimization (PPO) [158]:**

- Implements the RM and RL stages separately.
- Uses an explicit reward model trained on human preferences.
- Updates the policy by optimizing a clipped surrogate objective:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad (2)$$

where  $r_t(\theta)$  is the probability ratio between the new and old policy, and  $\hat{A}_t$  is the estimated advantage.

PPO's key contribution is its ability to perform stable policy updates while maintaining good sample efficiency. The "proximal" aspect of PPO is particularly important in the context of language models, as it helps maintain semantic consistency by ensuring that policy updates do not deviate too far from the previous policy.

#### **Direct Preference Optimization (DPO) [147]:**

- Combines the RM and RL stages into a single optimization process.
- Bypasses the need for an explicit reward model.
- Directly optimizes the policy based on preferred responses:

$$L^{DPO}(\theta) = \mathbb{E}(x, y_w, y_l) \sim \mathcal{D}[\log \sigma(\beta(r\theta(y_w|x) - r\theta(y_l|x)))] \quad (3)$$

where  $y_w$  and  $y_l$  are the winning and losing responses, respectively, and  $r\theta$  represents the policy's preference between responses.

DPO's key contribution is simplifying the RLHF process by eliminating the need for a separate reward model and policy optimization step, potentially leading to more efficient training.

The application of RLHF has expanded beyond text-only models to include multi-modal large models and text-to-image generation models. GPT-4 [2] demonstrated the effectiveness of RLHF in multi-modal contexts, while numerous works have leveraged RLHF to enhance model performance in text-to-image generation [71, 30, 13], showing good results in meeting human preferences, mitigating biases, and improving text-image alignment.

### **3.9 Evaluation Complex Reasoning Tasks**

o1-preview, a recent entrant in the pantheon of LLMs, has emerged as a paradigm shift in AI, particularly in its adeptness at tackling complex reasoning tasks. This model's innovative approach,

---

which integrates a "chain-of-thought" mechanism [190], sets it apart from its predecessors by enabling a more nuanced and methodical problem-solving process. Unlike traditional LLMs that often excel in pattern recognition and natural language processing, o1-preview's iterative reasoning allows it to dissect complex problems, evaluate intermediate steps, and synthesize solutions with a level of precision and logical coherence that was previously unattainable. By explicitly modeling the reasoning process, o1-preview demonstrates a significant advancement in the field of AI, bringing us closer to achieving human-like cognitive capabilities in machines.

The o1-preview's capabilities are most pronounced in domains that demand sustained, multi-step reasoning. For instance, in scientific research, where the model is tasked with analyzing data, formulating hypotheses, and drawing conclusions, it has demonstrated an impressive ability to conduct detailed analyses that are both accurate and insightful. Similarly, in the realm of programming, the model's capacity to understand and manipulate syntactic structures has been instrumental in tasks such as code debugging and optimization. Its integrated fact-checking feature is particularly valuable in high-stakes domains like legal analysis and medical diagnostics, ensuring that the information it processes and the conclusions it draws are grounded in verified data. These capabilities highlight the potential of o1-preview to revolutionize various industries by automating complex decision-making processes and augmenting human expertise.

However, o1-preview is not without its limitations. In scenarios that require rapid responses or where the complexity of reasoning is not paramount, the model's meticulous approach can be a liability. The iterative nature of its reasoning process can lead to slower output times compared to models optimized for speed, such as GPT-4 [2], in tasks that do not necessitate deep analytical capabilities. This trade-off between depth and speed is a critical consideration in applications where time sensitivity is paramount. It is essential to recognize these limitations and to deploy o1-preview in contexts where its strengths can be fully leveraged, while relying on alternative models for tasks that prioritize speed over depth.

The selection of evaluation domains for o1-preview was strategic, aiming to encompass a broad spectrum of complex reasoning tasks. These domains, ranging from computational mathematics to legal analysis, serve as a crucible for the model's capabilities, testing its adaptability and generalizability across different types of cognitive challenges [100]. By subjecting o1-preview to these diverse and demanding tasks, we seek to gain a comprehensive understanding of its potential and its limitations, which is essential for guiding its future development and application. The insights gleaned from this evaluation will not only inform the refinement of o1-preview but also contribute to the broader field of AI research, providing valuable lessons for the development of future models.

The implications of this evaluation extend beyond academic interest into the practical realms of technology and industry [92]. By delineating the strengths and weaknesses of o1-preview, we provide a roadmap for the integration of advanced AI into fields that stand to benefit from enhanced reasoning capabilities. The insights gleaned from this study will inform the development of future LLMs, potentially revolutionizing sectors such as finance, healthcare, and education by automating complex decision-making processes and augmenting human expertise. Furthermore, this work contributes to the broader goal of advancing artificial general intelligence, pushing the boundaries of what machines can achieve and how they can collaborate with humans to solve some of the most pressing challenges of our time.

In conclusion, the evaluation of o1-preview in complex reasoning tasks is a critical step towards harnessing the full potential of AI in transforming various sectors of society. Through a rigorous and comprehensive assessment, we aim to unlock new possibilities for innovation and progress, while

---

also identifying areas for improvement and further research. The o1-preview's performance in these evaluations will not only inform its immediate applications but also shape the trajectory of AI development, ensuring that future systems are better equipped to tackle the complex challenges that define our era. As we continue to push the boundaries of what is possible with AI, the insights gained from studying models like o1-preview will be instrumental in guiding us towards a future where machines and humans can work together seamlessly to solve the most pressing problems facing our world.

---

## 4 Experiments and Observation

### 4.1 Test Procedure

In the testing phase, we conducted a rigorous evaluation of the o1-preview model, deploying it across a diverse array of domain-specific tasks to comprehensively assess its performance and reasoning capabilities. The primary goal was to determine the model's ability to generate coherent, contextually appropriate, and logically consistent responses when confronted with challenges across various specialized fields. This was achieved by using domain-specific prompts and comparing the model's responses to pre-established benchmarks, thereby evaluating how effectively o1-preview could adapt to the nuanced demands of different domains.

Our testing approach followed a unified, systematic procedure designed to evaluate o1-preview's performance across several fields, including medicine, education, robotics, mathematics, etc. Each domain was chosen to highlight different aspects of the model's capabilities, ranging from multi-step reasoning and technical problem-solving to real-time decision-making and knowledge application. For example, in the medical domain, we used structured prompts that simulated real-world clinical scenarios, testing the model's ability to interpret patient data and suggest plausible diagnoses or treatments. In robotics, the model was tasked with generating code or instructions for various robotic tasks, emphasizing its ability to apply logic to dynamic, real-world operations. Similarly, the mathematics prompts required the model to solve intricate problems that necessitated a deep understanding of formulas, theorems, and multi-step problem-solving processes.

For each domain, we meticulously curated the datasets to ensure that they were representative of the types of challenges encountered by professionals in those fields. The structured prompts were designed not only to test the model's ability to retrieve factual information but also to assess how well it could adapt its reasoning to domain-specific contexts. In each case, we analyzed o1-preview's ability to generalize across tasks while maintaining coherence, accuracy, and relevance.

The evaluation went beyond simply recording o1-preview's correctness. Each response was critically analyzed for logical coherence, contextual relevance, and domain appropriateness. In domains like medicine, where incorrect responses could have serious real-world implications, we carefully assessed the model's ability to synthesize data from complex scenarios and provide plausible, evidence-based solutions. The same level of scrutiny was applied in mathematics, where multi-step reasoning is essential, and in robotics, where real-time task generation demands precision.

Moreover, we ensured that the evaluation was exhaustive by including multi-layered analyses—ranging from granular error breakdowns to broader assessments of how o1-preview handled domain-specific intricacies. For instance, when assessing its performance in education, we examined how effectively the model could generate accurate, didactic responses that were pedagogically sound. The detailed analysis provided insights into the model's ability to not only answer questions but also facilitate learning.

One of the key outcomes of this testing process was the identification of areas where o1-preview excelled, as well as those where further refinement is required. Its capacity to adapt to the unique demands of each domain while producing coherent and logically sound responses highlights its potential as a tool for a wide range of professional applications. However, the testing also revealed certain limitations, particularly in tasks requiring more intricate domain-specific reasoning, suggesting areas for further improvement.

## 4.2 Code Generation

To evaluate the code generation capabilities of o1-preview, we assessed its performance based on the number of correct solutions and the total points it accumulated in each contest. Table 2 presents the detailed results of o1-preview’s performance in our coding evaluation. Each problem within the contest is assigned a score proportional to its difficulty, with more challenging problems receiving higher scores. The maximum total score achievable in each contest is 21 points.

Table 2: Number of passed questions and corresponding points o1-preview has obtained in our coding test. o1-preview is given three attempts. A problem is considered successfully solved by o1-preview if it passes Leetcode’s automated judgement system within any of the three allowed submission attempts.

|                               | Contest 414 | Contest 413 | Biweekly Contest 138 | Total |
|-------------------------------|-------------|-------------|----------------------|-------|
| Number of passed solution     | 4           | 3           | 3                    | 10    |
| Percentage of passed solution | 100%        | 75%         | 75%                  | 83.3% |
| Score of point                | 21          | 15          | 14                   | 50    |

As presented in Table 2, o1-preview successfully passed 10 out of 12 problems, yielding a high passing rate of 83.3%. The model only failed one problem in Weekly Contest 413 and another in Biweekly Contest 138. Notably, both of these problems were categorized as "hard" level. Despite taking several minutes to generate solutions, o1-preview was unable to pass these challenges within the three submission attempts.

For the failed problem in Weekly Contest 413, o1-preview initially generated a solution that was correct but exceeded the time limit. Subsequent submissions in the second and third attempts resulted in incorrect answers. Similarly, for the problem in Biweekly Contest 138, o1-preview provided incorrect solutions for all three attempts.

Despite these challenges, o1-preview demonstrated comparable or faster code generation times relative to top human competitors. This indicates that o1-preview possesses a strong reasoning capability and can effectively handle the majority of coding tasks. However, in particularly complex or computationally intensive cases, the model’s performance remains limited, as observed in these difficult problems.

Figures 3 and 4 illustrate two examples from our coding evaluation. In Figure 3, o1-preview demonstrates its ability to efficiently solve easy-level problems, completing the task with minimal computation time and successfully passing all test cases. However, as shown in Figure 4, the model encounters difficulties with a hard-level problem. In this failure case, o1-preview initially generated a correct solution, but the solution exceeded the time limit. Following the first submission, the model became trapped in a cycle of attempting to optimize the code’s time complexity, which resulted in incorrect solutions in subsequent attempts. This example highlights the model’s struggle with complex problems requiring significant optimization, where its efforts to improve efficiency led to repeated errors. Additional test cases are provided in the Appendix A.1

## Prompt

You are given a string date representing a Gregorian calendar date in the yyyy-mm-dd format. date can be written in its binary representation obtained by converting year, month, and day to their binary representations without any leading zeroes and writing them down in year-month-day format.

Return the **binary** representation of date.

### **Example 1:**

**Input:** date = "2080-02-29"

**Output:** "100000100000-10-11101"

### **Explanation:**

100000100000, 10, and 11101 are the binary representations of 2080, 02, and 29 respectively.

### **Example 2:**

**Input:** date = "1900-01-01"

**Output:** "11101101100-1-1"

### **Explanation:**

11101101100, 1, and 1 are the binary representations of 1900, 1, and 1 respectively.

### **Constraints:**

•date.length == 10

•date[4] == date[7] == '-', and all other date[i]'s are digits.

•The input is generated such that date represents a valid Gregorian calendar date between Jan 1<sup>st</sup>, 1900 and Dec 31<sup>st</sup>, 2100 (both inclusive).

## O1-preview

```
def date_to_binary(date):
    # Extract year, month, and day as strings
    year_str = date[0:4]
    month_str = date[5:7]
    day_str = date[8:10]

    # Convert to integers
    year = int(year_str)
    month = int(month_str)
    day = int(day_str)

    # Convert to binary and remove the '0b' prefix
    binary_year = bin(year)[2:]
    binary_month = bin(month)[2:]
    binary_day = bin(day)[2:]

    # Return the formatted binary date
    return f'{binary_year}-{binary_month}-{binary_day}'
```

Figure 3: **Coding: Case 1.** A demonstration case from coding testing of a easy level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

| Prompt  | O1-preview  |
|---|---|
| <p>You are given a 2D matrix grid consisting of positive integers.</p> <p>You have to select one or more cells from the matrix such that the following conditions are satisfied:</p> <ul style="list-style-type: none"> <li>No two selected cells are in the same row of the matrix.</li> <li>The values in the set of selected cells are unique.</li> <li>Your score will be the sum of the values of the selected cells.</li> <li>Return the maximum score you can achieve.</li> </ul> <p><b>Example 1:</b><br/> Input: grid = [[1,2,3],[4,3,2],[1,1,1]]<br/> Output: 8</p> <p>We can select the cells with values 1, 3, and 4 that are colored above.</p> <p><b>Example 2:</b><br/> Input: grid = [[8,7,6],[8,3,2]]<br/> Output: 15</p> <p>We can select the cells with values 7 and 8 that are colored above.</p> <p><b>Constraints:</b><br/> <math>1 \leq \text{grid.length}, \text{grid}[i].length \leq 10</math><br/> <math>1 \leq \text{grid}[i][j] \leq 100</math></p> | <pre> class Solution:     def maxScore(self, grid):         from collections import defaultdict          n = len(grid) # Number of rows         assigned_rows = set()         self.max_sum = 0          # Collect all unique values and the rows they appear in         value_to_rows = defaultdict(set)         for row_index, row in enumerate(grid):             for value in row:                 value_to_rows[value].add(row_index)          # Sort values in descending order         sorted_values = sorted(value_to_rows.keys(), reverse=True)          # Precompute the cumulative sum of remaining values for pruning         cumulative_values = [0] * len(sorted_values)         cumulative_values[-1] = sorted_values[-1]         for i in range(len(sorted_values) - 2, -1, -1):             cumulative_values[i] = cumulative_values[i + 1] + sorted_values[i]          def dfs(index, assigned_rows, current_sum):             # Base case: all rows are assigned or all values are considered             if len(assigned_rows) == n or index == len(sorted_values):                 if current_sum &gt; self.max_sum:                     self.max_sum = current_sum                 return              # Prune if the maximum possible sum is less than current max_sum             remaining_sum = cumulative_values[index]             if current_sum + remaining_sum &lt;= self.max_sum:                 return              value = sorted_values[index]             rows = value_to_rows[value]              assigned = False             for row in rows:                 if row not in assigned_rows:                     assigned_rows.add(row)                     dfs(index + 1, assigned_rows, current_sum + value)                     assigned_rows.remove(row)                     assigned = True             # Optionally skip the current value             dfs(index + 1, assigned_rows, current_sum)          dfs(0, set(), 0)         return self.max_sum </pre> |

Figure 4: **Coding: Case 2.** In this case, o1-preview generates the correct code. However, this code has exceeded the time limit. For the following second and third attempt, o1-preview keeps generating the code with wrong answer. Therefore, we regard this problem as a failure case.

### 4.3 Radiology Report Generation

We evaluated the medical report generation capability of the o1-preview by comparing its ROUGE metrics against baseline models such as gpt-4-turbo, gpt-4o, and others. Table 3 provides a detailed comparison of the performance of o1-preview with five other models, with the best-performing results highlighted. Additionally, we calculated the average generation time for medical reports produced by each model to further assess their overall efficiency and performance.

As shown in Table 3, the ROUGE scores for the reports generated by o1-preview compared to those written by doctors are: R-1: 0.3019, R-2: 0.0448, and R-L: 0.2841, ranking highest among the six models. Notably, o1-preview also has the longest average report generation time at 15.051 seconds, while gpt-3.5-turbo is the fastest at 2.024 seconds, though it has the lowest ROUGE scores. This difference can be attributed to o1-preview’s extensive inference verification processes aimed at ensuring result accuracy, which gpt-3.5-turbo lacks. In terms of generation time, both o1-preview and gpt-4-turbo took longer than the other models but delivered stronger ROUGE results.

Table 3: Comparing different models with R-1, R-2, and R-L scores and average time, including performance indicators.

| Models        | SXY           |               |               | Average Time(s) |
|---------------|---------------|---------------|---------------|-----------------|
|               | R-1           | R-2           | R-L           |                 |
| o1-preview    | <b>0.3019</b> | 0.0448        | <b>0.2841</b> | 15.051          |
| o1-mini       | 0.2582        | 0.0586        | 0.2449        | 5.793           |
| gpt-4o        | 0.2855        | 0.0615        | 0.2552        | 4.764           |
| gpt-4o-mini   | 0.1346        | 0.0154        | 0.1269        | <b>1.629</b>    |
| gpt-4-turbo   | 0.2999        | <b>0.0656</b> | 0.2667        | 13.942          |
| gpt-3.5-turbo | 0.0600        | 0.0154        | 0.0600        | 2.024           |

Figures 5 and 6 present two examples illustrating the evaluation of our radiology report generation. During the evaluation, o1-preview demonstrated the ability to generate reliable radiological reports with zero-shot samples in less time, highlighting its strong reasoning and knowledge transfer capabilities, as well as its potential in the medical field. We observed that o1-preview aligns well with human writing patterns, featuring clear organization and concise language. While other models did not achieve the highest similarity scores, most were able to follow instructions and complete the tasks. Although these large language models have the potential to reduce doctors’ workloads, alleviate the strain on limited medical resources, and foster collaboration between clinical practitioners and AI technology, substantial improvements and rigorous validation are necessary before clinical deployment. Additional test cases are provided in the Appendix A.2

---

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. A pure ground glass nodule (CT value -627HU) approximately 4mm in the posterior basal segment of the lower lobe of the left lung is considered as a small ground glass nodule, and regular follow-up observation is recommended.
2. A cord like high-density lesion in the posterior segment of the upper lobe of the right lung, with clear edges, is considered a fibrous cord shadow.

### **Reference Answer**

1. Bilateral ground glass nodules are roughly the same as before, LU-RADS class 2, and annual follow-up is recommended.
2. The left breast mass has shrunk compared to before, indicating effective treatment. It is recommended to have a follow-up examination.
3. Multiple gallstones in the gallbladder.

Figure 5: **Radiology: Case 1.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. \*\* Postoperative changes after nodule resection in the left upper lung and right middle lung \*\*: Patchy dense shadows can be seen in the left upper lung and right middle lung, which are consistent with the characteristics of postoperative changes and show no significant changes compared to the previous examination.
2. \*\* Stable pure ground glass nodule in the upper left lung \*\*: Pure ground glass nodules (image numbers lm75, 123) were observed in the upper left lung, with no significant changes compared to the examination results on May 19, 2020, indicating stable nodules.
3. \*\* No new pulmonary lesions were found \*\*: No new nodules or lesions were found in the lungs, and the remaining lung field conditions were roughly the same as the previous examination.

### **Reference Answer**

Changes after resection of nodules in the upper left and middle right lungs, with ground glass nodules at the apex of the left lung as before (Class 2), suggest annual follow-up.

Figure 6: **Radiology: Case 2.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

---

## 4.4 Robot Command Planning

The Robot Operating System (ROS) is widely recognized as one of the most commonly utilized frameworks for robotic control and development. It integrates key functionalities such as node communication, parallel control, and behavior trees—features that are essential in robotics but rarely found in traditional software systems. Due to the complexity and specialized nature of these control mechanisms, mastering ROS is considered highly challenging, even among seasoned programmers. This difficulty is further compounded by the fact that many large language models (LLMs) struggle to generate high-quality ROS control code efficiently, primarily because of the nuanced and domain-specific knowledge required to handle the intricate interactions between its components. Consequently, ROS development remains a specialized skill that is not easily automated or replicated by general-purpose AI models.

The results, as depicted in Figure 7, clearly illustrate o1-preview’s profound comprehension of robot command structures. The generated code not only aligns accurately with the given instructions but also demonstrates a high degree of precision and reliability. This suggests that o1-preview possesses an exceptional ability to interpret and synthesize complex commands within the context of robotic control. The accuracy of the output further underscores the model’s robustness in handling intricate tasks, which are critical in real-world applications where precise control and execution are paramount.

**Prompt**

Can you help me write the code of ROS for 4-wheels control?

**O1-preview**

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import rospy
from geometry_msgs.msg import Twist

def move():
    rospy.init_node('four_wheel_controller', anonymous=False)
    cmd_vel_pub = rospy.Publisher('/cmd_vel', Twist, queue_size=10)
    rate = rospy.Rate(10)
    move_cmd = Twist()
    move_cmd.linear.x = 0.5
    move_cmd.angular.z = 0.0
    while not rospy.is_shutdown():
        cmd_vel_pub.publish(move_cmd)
        rate.sleep()
    if __name__ == '__main__':
        try:
            move()
        except rospy.ROSInterruptException:
            pass
```

Figure 7: **Robot Command Planning: Case 1.** ROS code generated by o1-preview for controlling a four-wheel robot, featuring Twist message handling to control forward movement and angular velocity.

---

Moreover, this performance highlights o1-preview's potential to play a significant role in advancing automation and robotics, as it can efficiently translate abstract instructions into executable commands. The model's capabilities extend beyond simple task interpretation, offering a deeper integration of machine learning in robotics. As the field moves toward more sophisticated and autonomous systems, the ability of models like o1-preview to adapt and deliver highly accurate code suggests promising advancements in both industrial and research-based robotic applications.

This demonstration affirms that o1-preview is not only a powerful tool for natural language processing but also a versatile system that can be leveraged to enhance control mechanisms in robotics, paving the way for innovative approaches in machine learning-driven automation.

---

## 4.5 Nature Language Inference

We randomly selected two test samples from each dataset’s test set, conducting a total of 10 test cases. Table 4 presents an example for each dataset, where o1-preview accurately analyzed the logical relationships, demonstrating advanced reasoning ability and domain-specific knowledge. This underscores its potential for real-world applications in various complex scenarios. Additional examples can be found in Appendix A.4.

Table 4: Examples of o1-preview responses on NLI task.

For NLI task, the prompt is: Given a premise sentence and a hypothesis sentence, predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral).

For QNLI task, the prompt is: Determine whether the context sentence contains the answer to the question.

| Dataset | Premise / Context   | Hypothesis / Question                                      | o1-preview Response  | Label          |
|---------|---|--|--|----------------|
| MNLI    | Tom is the winner of a year's supply of Turtle Wax, and he will receive his prize just as soon as the Shopping Avenger figures out how much Turtle Wax actually constitutes a year's supply.  | There are no winners of the one year supply of Turtle wax. | The premise explicitly states that Tom is the winner, while the hypothesis claims there are no winners. These statements directly oppose each other, making the relationship a contradiction. Answer: <b>Contradiction</b>   | Contradiction  |
| ANLI    | The recent 14% hike in third class postage rates, accompanied by simultaneous double-digit paper price increases, has hit smaller catalogers especially hard.   | Smaller catalogers are hit harder than bigger ones.        | The premise states that smaller catalogers have been hit especially hard by the increase in postage rates and paper prices. The term "especially hard" implies that smaller catalogers are more adversely affected compared to others in the industry. Although the premise does not explicitly mention bigger catalogers, the use of "especially" suggests a comparison, indicating that smaller catalogers are hit harder than their larger counterparts. Therefore, the premise entails the hypothesis. Answer: <b>Entailment</b> | Entailment     |
| QNLI    | Stagecoach is the primary operator in the city proper, with cross-city services mainly between both the West and East ends via the city centre with some services extending out to the MetroCentre, Killingworth, Wallsend and Ponteland. | How many major bus stations are in the city of Newcastle?  | No, the context sentence does not provide the number of major bus stations in the city of Newcastle.   | Not entailment |
| MedNLI  | EKG showed T-wave depression in V3-5, with no prior EKG for comparison.   | Patient has a normal EKG.                                  | The premise states that the EKG revealed T-wave depression in leads V3-5, which is an abnormal finding indicating potential cardiac issues such as ischemia or other heart conditions. A normal EKG would not show such abnormalities. Therefore, the hypothesis that the patient has a normal EKG directly contradicts the information provided in the premise. Answer: <b>Contradiction</b>  | Contradiction  |
| RadQNLI | The heart size is probably normal, but the cardiac silhouette is partially obscured by a moderate left pleural effusion.  | Do we find any unilateral or bilateral pleural effusion?   | Yes, the context sentence contains the answer. It mentions a moderate left pleural effusion, which indicates a unilateral pleural effusion.  | Entailment     |

## 4.6 Quantitative Investing

o1-preview has solid financial knowledge and statistical modeling skills, and can deeply understand and clearly explain complex concepts such as mean reversion (from Figure 96), ARIMA models (from Figure 8), stochastic oscillators (from Figure 98), and the role of autocorrelation (from Figure 96) in market forecasting. In addition, o1-preview has demonstrated the ability to comprehensively apply trend, seasonality (from Figure 97), and residual analysis in market forecasting, and has the ability to think systematically and apply practically. Overall, o1-preview has performed well in the fields of financial analysis and market forecasting, and has the ability to conduct comprehensive analysis and application.

### Prompt

What is the ARIMA model and how does it help in predicting stock market trends?

### O1-preview

The ARIMA model, which stands for AutoRegressive Integrated Moving Average, is a widely used statistical method for analyzing and forecasting time series data. In the context of the stock market, ARIMA models are employed to predict future stock prices based on historical data patterns.

Components of the ARIMA Model.

An ARIMA model is characterized by three parameters: p, d, and q, and is generally denoted as ARIMA(p, d, q).

1. AutoRegressive (AR) Component (p): Definition: Reflects the relationship between an observation and a specified number of lagged observations. Function: Captures the momentum or trends in the data by regressing the variable on its own previous values.
2. Integrated (I) Component (d): Definition: Represents the number of times the data have been differenced to achieve stationarity. Function: Eliminates non-stationarity (trends or seasonality) in the time series by differencing the data.
3. Moving Average (MA) Component (q): Definition: Shows the relationship between an observation and residual errors from a moving average model applied to lagged observations. Function: Accounts for shocks or random fluctuations in the data.

### Reference Answer

ARIMA models combine autoregressive (AR) and moving average (MA) components with differencing to capture trends and seasonality in stock data for forecasting.

Figure 8: **Quantitative Investing: Case 1.** o1-preview successfully explained the concept of ARIMA, described its functionality, and detailed its working mechanism, demonstrating a thorough understanding of the statistical model. The dataset for this task is sourced from [203].

**Prompt**

How does autocorrelation play a role in time series analysis and its implications for forecasting market trends?

**O1-preview**

Definition:

- Autocorrelation, also known as serial correlation, refers to the correlation of a time series with its own past and future values.
- It quantifies the similarity between observations as a function of the time lag between them.

Role of Autocorrelation in Time Series Analysis

1. Identification of Patterns:

1. Trend Detection: Positive autocorrelation at low lags may indicate a trend in the data.
2. Seasonality Recognition: Significant autocorrelations at specific lags can reveal seasonal patterns.

2. Model Selection and Specification:

1. ARIMA Models:
  1. AR Component: Autoregressive (AR) models rely on autocorrelation by regressing the variable on its own lagged values.
  2. MA Component: Moving Average (MA) models use autocorrelation in the residuals to model the error terms.
2. Parameter Estimation:
  1. ACF and PACF Plots: Autocorrelation and Partial Autocorrelation Functions help determine the appropriate order of AR and MA terms.

3. Diagnostic Checking:

1. Residual Analysis: After fitting a model, autocorrelation in residuals indicates model inadequacy.
2. Ljung-Box Test: A statistical test to check for the presence of autocorrelation in residuals.

**Reference Answer**

Autocorrelation measures the correlation of a time series with its lagged values, helping in identifying patterns and dependencies that can be leveraged for making predictions about future market trends.

Figure 9: **Quantitative Investing: Case 2.** o1-preview successfully identified the role of autocorrelation in market forecasting, explained its working mechanism, and highlighted its importance, demonstrating an accurate understanding of the underlying process.

## 4.7 Low-Resource Language Translation

In the evaluation of low-resource language translations by the o1-preview model, two Cherokee sentence cases were examined. In the first scenario, as shown in Figure 10, the sentence "The boys are playing ball" was accurately translated. The model identified all key elements, including the plural form of "boys" and the action verb "playing," while also offering a detailed breakdown of the grammatical structure. Similarly, in the second case, as shown in Figure 11, the o1-preview model translated the sentence into "The chickens are going to their coop because of the foxes." While it correctly identified "chickens," it has some interpretations, for example, "foxes" and "coop", differed from the reference translation, which described "roosters crowing at dawn." These instances demonstrated the model's ability to handle word recognition and basic grammar in low-resource language translation. However, while the model accurately faced unknown words in each case, it struggled with contextual nuances and precision in more complex scenarios.

### Prompt

Help me translate: Θ DhјG ᐕTLedI ɬθnG&oðS.

### O1-preview

The sentence "Θ DhјG ᐕTLedI ɬθnG&oðS." translates to:

"**The boys are playing ball.**"

Here's a breakdown of the translation:

Θ DhјG (na anitsutsa): "The boys"

Θ (na): "the"

DhјG (anitsutsa): "boys"

ᐕTLedI (sgwatlesdi): "ball"

ɬθnG&oðS (dananelohv?sga): "they are playing"

ɬ- (da-): plural prefix "they"

θnG&oðS (nanelohv?sga): "are playing"

### Reference Answer

**The boys are playing ball.**

Figure 10: **Low-Resource Language Translation: Case 1.** In this translation sample, the model accurately translates this Cherokee sentence as "The boys are playing ball", which is exactly the same with reference answer. It also offers a detailed breakdown of each component of the sentence, demonstrating its ability to handle both grammar and words in this translation.

**Prompt**

Help me translate: IrWS DhCꝫ O'YIroᬁSᬁ DθBAT.

**O1-preview**

The sentence "IrWS DhCꝫ O'YIroᬁSᬁ DθBAT." translates to:

"The chickens are going to their coop because of the foxes."

Breakdown of the Translation:

IrWS (tsitaga): "chickens"

IrWS is the Cherokee word for "chickens."

DhCꝫ (anitsvya): "foxes"

Dh- (ani-): A plural prefix used for animate beings, meaning "they" or indicating multiple subjects.

Cꝫ (tsvya): "fox"

Together, anitsvya means "foxes."

O'YIroᬁSᬁ (ugitsisdeyi): "their coop" or "their roost"

O'Y- (ugi-): Third person plural possessive prefix, meaning "their."

IroᬁSᬁ (tsisdeyi): "coop," "roost," or "nesting place"

Combined, ugitsisdeyi means "their coop" or "their nesting place."

DθBAT (anayvgoi): "they are going" or "they are heading"

Dθ- (ana-): Plural prefix "they"

BAT (yvgoi): "are going," "are heading," or "are proceeding"

Together, anayvgoi means "they are going" or "they are heading."

Interpretation:

The sentence describes a scenario where the chickens are moving towards their coop because of the presence of foxes. The structure emphasizes the cause (foxes) and the action taken by the chickens (going to their coop).

**Reference Answer**

Roosters crow at dawn.

Figure 11: **Low-Resource Language Translation: Case 2.** This translation correctly translates the word chicken. But o1-preview model is unable to translate the rest words, like male, crow and dawn, in this sentence.

## 4.8 Educational Q&A

In this educational test using o1-preview, the model successfully identified the correct answers for two distinct scientific concepts. In the first Figure 12, o1-preview correctly recognized that the phenomenon responsible for global wind deflection is the Coriolis effect, which results from the Earth's rotation and affects wind patterns across hemispheres. In the second Figure 13, o1-preview accurately identified exothermic as the process where changes from a less-ordered state to a more-ordered state, such as a liquid turning into a solid, involve the release of heat. These results demonstrate o1-preview's capability to accurately interpret and respond to scientifically framed questions.

### **Prompt**

What phenomenon makes global winds blow northeast to southwest or the reverse in the northern hemisphere and northwest to southeast or the reverse in the southern hemisphere?

Must Choose from below options:

- muon effect
- tropical effect
- centrifugal effect
- Coriolis effect

### **O1-preview**

The phenomenon responsible for the deflection of global winds in the described patterns is the Coriolis effect.

Coriolis Effect: This effect arises from the Earth's rotation. It causes moving air (and other objects) to turn to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. As a result, global wind patterns are deflected, leading to winds blowing from northeast to southwest or the reverse in the Northern Hemisphere and from northwest to southeast or the reverse in the Southern Hemisphere.

Correct Answer: **Coriolis effect**

### **Reference Answer**

**Coriolis effect**

Figure 12: **Educational Q&A: Case 1.** o1-preview successfully identified the correct answer, Coriolis effect, for the phenomenon responsible for global wind deflection, demonstrating accurate understanding of the underlying process. The dataset for this task is sourced from [193].

---

**Prompt**

Changes from a less-ordered state to a more-ordered state (such as a liquid to a solid) are always what?

Must Choose from below options:

- endothermic
- unbalanced
- exothermic
- reactive

**O1-preview**

Changes from a less-ordered state to a more-ordered state, such as a liquid turning into a solid (freezing), involve the release of energy. This process releases heat into the surrounding environment, making it exothermic.

**Correct Answer:** exothermic

**Reference Answer**

exothermic

Figure 13: **Educational Q&A: Case 2** o1-preview successfully identified the correct answer, exothermic, for the process of changes from a less-ordered state to a more-ordered state, such as a liquid turning into a solid, demonstrating accurate understanding of the energy release involved. The dataset for this task is sourced from [193].

---

## 4.9 Student Writing Improvement in Higher Education

As educators explore new technologies to enhance student learning, o1-preview may serve as an important teaching and learning assistant in helping students improve their writing skills. Qualitative analysis of the o1-preview model's application in student writing in Figures 14, 15, 16 has shown promising results in several areas. For example, the above screenshots reveal that o1-preview appears to be able to provide effective writing support, including improving linguistic accuracy, enhancing writing coherence, and generating writing outlines and initial ideas. However, while these functions seem to contribute positively, the overall impact may vary depending on the specific needs of learners and the contexts in which the tool is used. The initial analysis found that the AI-generated responses sometimes follow a rigid, machine-like template and lack citations, which may lead students to passively adopt these structures, ultimately limiting the effectiveness and creativity in their writing. Therefore, it is important for educators to understand that while GenAI can be a valuable supplement, it cannot replace the essential human guidance in writing education. Most importantly, the current AI model is still under development, which means its limitations must be carefully considered in educational contexts [180, 68]. We hope that future research will further investigate and examine o1's functions and its effectiveness in higher education, providing deeper insights into how it can be best utilized across various learning contexts.

## Improving linguistic accuracy

### **Prompt**

Modify and improve linguistic accuracy for the text below:

"With the two years' studying in the Unite state, my ambition and inspiration are gradually accompanied with and replaced by the homesickness. Once having the free time, recalling my hometown become the habit that I adopted recently. The place which always flashes in my eyes may be my high school, <place> which is a school always with flowers and happiness. Let's begin creating my memories."

### **o1-preview**

After studying in the United States for two years, my ambition and inspiration have gradually been replaced by homesickness. Whenever I have free time, I find myself recalling my hometown; it has become a habit I have adopted recently. The place that always comes to mind is my high school, <place>, a school always filled with flowers and happiness. Let me begin sharing my memories.

## Enhancing writing coherence, fluency, and natural expression

### **Prompt**

Use appropriate coherence devices to enhance coherence and fluency of the text below:

"I met my best friend in the middle school, but she left our school in the second year and went to New Zealand. We had many great memories together and we spent our most of time together in school as well. We have spent a lot of precious time together, remembering when we were running in the rain and laughing crazy at the same time. We used to like to go to the garden and the sidewalks around the school to appreciate different plants after we finished the school. And we liked to sing out loud in the hallway and discovered the secret tunnels in our school. So it was really hard for me to accept the fact that she has to leave the school and I have to deal with my school life without her. I remembered I feel sad every time I go pass by the places that we were used like to go in school. Sometimes the memories that I had with her would keep flashing back when I was walking around the school, then I would stop walking and started to think about her. Colors began to fade away in the days without her, but I still need to keep moving on."

### **o1-preview**

I met my best friend in middle school, but she left during the second year and moved to New Zealand. During our time together, we created many great memories and spent most of our time together at school. I recall the precious moments we shared, such as when we ran in the rain, laughing like crazy. After school, we used to enjoy exploring the garden and strolling along the sidewalks around the school to appreciate different plants. Additionally, we loved to sing out loud in the hallways and discover the secret tunnels in our school. Therefore, it was really hard for me to accept the fact that she had to leave, and I had to face school life without her. I remember feeling sad every time I passed by the places we used to go in school. Sometimes, memories of her would flash back when I was walking around the school, causing me to stop and think about her. As a result, colors began to fade away in the days without her, but I knew I still needed to keep moving on.

Figure 14: Academic Writing Improvement: Case 1. Example of general improvements in higher education writing.

## Generating writing outline and initial ideas

### **Prompt**

Generate an outline for a research paper about Generative AI in higher education

### **o1-preview**

Title: The Role of Generative AI in Transforming Higher Education

Abstract: A brief summary of the research paper, highlighting the key findings and implications of incorporating generative AI in higher education.

1. Introduction
  - Background: Overview of artificial intelligence (AI) and its evolution. Introduction to generative AI and its capabilities.
  - Purpose of the Study: Exploring the impact of generative AI on higher education. Identifying opportunities and challenges.
  - Research Questions: How is generative AI currently being used in higher education? What are the potential benefits and risks?
2. Literature Review
  - AI in Education: Historical context of AI applications in educational settings.
  - Generative AI Models: Explanation of models like GPT-3 and GPT-4. Advances in natural language processing (NLP).
  - Current Implementations: Case studies of generative AI tools used in universities.
  - Theoretical Framework: Educational theories supporting AI integration.
3. Methodology
  - Research Design: Qualitative, quantitative, or mixed-method approach.
  - Data Collection: Surveys, interviews, and academic performance data.
  - Data Analysis: Statistical methods and software used.
4. Applications of Generative AI in Higher Education
  - Personalized Learning: Adaptive learning systems. Customized study plans.
  - Content Generation: Automated creation of lecture notes and summaries. AI-generated practice questions and assessments.
  - Language Translation and Accessibility: Real-time translation services. Support for students with disabilities.
  - Administrative Support: Chatbots for student services. Streamlining enrollment and scheduling.
5. Challenges and Ethical Considerations
  - Academic Integrity: Risks of plagiarism and cheating with AI tools.
  - Bias and Fairness: Addressing algorithmic biases in AI-generated content.
  - Data Privacy: Safeguarding student information.
  - Dependency on Technology: Potential loss of critical thinking skills.
  - Regulatory Compliance: Navigating legal frameworks and accreditation standards.
6. Case Studies
  - University Implementations: Detailed analysis of institutions successfully integrating generative AI.
  - Outcomes and Feedback: Student and faculty perspectives.
  - Lessons Learned: Best practices and pitfalls.
7. Future Perspectives
  - Advancements in AI Technology: Emerging trends and potential breakthroughs.
  - Long-Term Implications: How generative AI might redefine higher education.
  - Policy Recommendations: Guidelines for ethical and effective AI integration.
8. Conclusion
  - Summary of Findings: Recap of the key insights from the research.
  - Implications for Stakeholders: Impact on students, educators, institutions, and policymakers.
  - Final Thoughts: The transformative potential of generative AI in education.

Figure 15: Academic Writing Improvement: Case 2 Example of writing preparation support.

---

## Lack of citations

**Prompt**

Write a paragraph regarding educational theories supporting AI integration.

**o1-preview**

Educational theories such as constructivism, cognitivism, and connectivism support the integration of artificial intelligence (AI) in education. Constructivism emphasizes that learners build their own understanding through experiences; AI can facilitate this by providing personalized learning environments that adapt to individual needs. Cognitivism focuses on the mental processes involved in learning, and AI tools can enhance these processes by offering adaptive feedback and scaffolding to develop critical thinking skills. Connectivism highlights the importance of networks and the flow of information; AI can connect students to vast resources and expert systems, fostering collaborative and self-directed learning. By aligning with these theories, AI integration can enrich educational experiences and promote more effective learning outcomes.

## Limited Creativity and Personalization

**Prompt**

Write a beginning paragraph for the following topic:

Should universities allow students to use GenAI tools like ChatGPT for academic writing?

**o1-preview**

The rapid advancement of generative artificial intelligence (GenAI) tools like ChatGPT has sparked a transformative shift in how information is accessed and produced. As these technologies become increasingly sophisticated and accessible, universities are grappling with the question of whether to permit their use in academic writing. Proponents argue that GenAI can serve as a valuable resource for idea generation and learning enhancement, while critics raise concerns about academic integrity and the potential erosion of critical thinking skills. This dilemma invites a deeper exploration of the role that AI should play in higher education and its impact on the development of student competencies.

Figure 16: Academic Writing Improvement: Case 3 Example of limitations.

---

## 4.10 3D Layout Generation

In this part, we evaluated the performance of model *o1-preview* in generating room layouts. We provided the model with a carefully designed prompt aimed at guiding it to effectively arrange objects within a given room type. We use the prompt from LLplace [197]. The specific prompt is as follows:

You are a skilled room layout designer. Your task is to arrange [Objects] within a given [Room Type] effectively. Follow these guidance to complete your design:

1. Extract the [Room Type], [Objects], and [Bounding Box Size] from the provided JSON data.
2. Analyze the spatial relationships among [Objects] within the specified [Room Type]. Pay special attention to avoiding overlap and consider other spatial factors like accessibility and aesthetics.
3. Determine and design the precise location of all [Objects] ensuring that their bounding boxes do not overlap and that the layout is functional and visually appealing.
4. I prefer objects to be placed at the edge (the most important constraint) of the room if possible, which makes the room look more spacious.
5. The objects are usually **aligned**.
6. Chairs must be placed near to the table/desk and face to the table/desk.
7. The last design output token is the [/Task Output] and only one.
8. Report your design with detailed 3D space coordinates and rotation angles for each object in JSON format, as follows:

```
"object": "object", "coordinates": [ "x": x, "y": y, "z": z ], "rotate": [ "angle": r ]
```

The centroid of the room is "x": 0.00, "y": 0.00, "z": 0.00 .

First carefully read this example:

[Example Room Type] Bedroom [/Example Room Type]

[Example Objects and Bounding Box Size] /\* A fixed example is put here to show the input format \*/ [/Example Objects and Bounding Box Size]

[Example Output] /\* A fixed example is put here to show the output format \*/ [/Example Output]

Now, please proceed with the design task as outlined and provide only the JSON formatted output of your design:

[Task Room Type] /\* Input room type \*/ [/Task Room Type]

[Task Objects & Bounding Box Size] /\* The JSON format input of objects description and bounding box size \*/ [/Task Objects & Bounding Box Size]

Through experiments, we found that model *o1-preview* performed exceptionally well in the task of room layout generation. The model was able to accurately extract the room type, objects, and bounding box sizes from the provided JSON data, ensuring the correct interpretation of fundamental information. It conducted an in-depth spatial relationship analysis of the objects within the specified room type, successfully avoiding overlap between objects in most cases, while also considering spatial factors such as accessibility and aesthetics.

In the layout design task, the model strictly adhered to the guidelines, precisely determining the positions of all objects and ensuring that their bounding boxes did not overlap. The resulting layouts were highly functional and visually pleasing. Notably, the model prioritized placing objects at the edges of the room, a strategy that effectively increased the sense of spaciousness. Objects in the

---

layout were generally aligned, and chairs were correctly placed near tables or desks and faced them, meeting both functional and aesthetic requirements.

However, it is necessary to rigorously point out that in a small number of examples, there were instances of object overlap in three-dimensional space, indicating that the results were not entirely accurate. To comprehensively evaluate the model’s performance, we conducted three-dimensional visualizations of each generated layout; specific details are shown in Figures 17, 18, 19, 20. Additionally, we reported the three-dimensional spatial coordinates and rotation angles of each object in the required detailed JSON format.

In summary, *o1-preview* demonstrated strong capabilities in the task of room layout generation. Although it accurately understood the requirements of the prompt and reflected a high level of functionality and aesthetics in the generated layouts, there were still some minor inaccuracies, such as overlaps in certain examples, indicating room for improvement. The model has broad application prospects in fields such as interior design and virtual reality, but future work should focus on enhancing its accuracy in complex layouts.

### Prompt

Here is a bedroom with a brown wardrobe, two brown nightstands with a lid, a wooden bowl with a metal handle and a grey double bed with cover and pillows.

### O1-preview

```
[{"object": "grey double bed with cover and pillows",
  "bounding_box": {"width": 1.6, "depth": 2.0, "height": 0.5},
  "coordinates": [{"x": 0.00, "y": -1.50, "z": 0.00}],
  "rotate": [{"angle": 0}]}, {"object": "brown nightstand with a lid",
  "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6},
  "coordinates": [{"x": -1.05, "y": -1.50, "z": 0.00}],
  "rotate": [{"angle": 0}]}, {"object": "brown nightstand with a lid",
  "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6},
  "coordinates": [{"x": 1.05, "y": -1.50, "z": 0.00}],
  "rotate": [{"angle": 0}]}, {"object": "brown wardrobe",
  "bounding_box": {"width": 1.5, "depth": 0.6, "height": 2.0},
  "coordinates": [{"x": -2.20, "y": 0.00, "z": 0.00}],
  "rotate": [{"angle": 90}]}, {"object": "wooden bowl with a metal handle",
  "bounding_box": {"width": 0.3, "depth": 0.3, "height": 0.2},
  "coordinates": [{"x": -1.05, "y": -1.50, "z": 0.75}],
  "rotate": [{"angle": 0}]]}
```

### 3D Layout Visualization

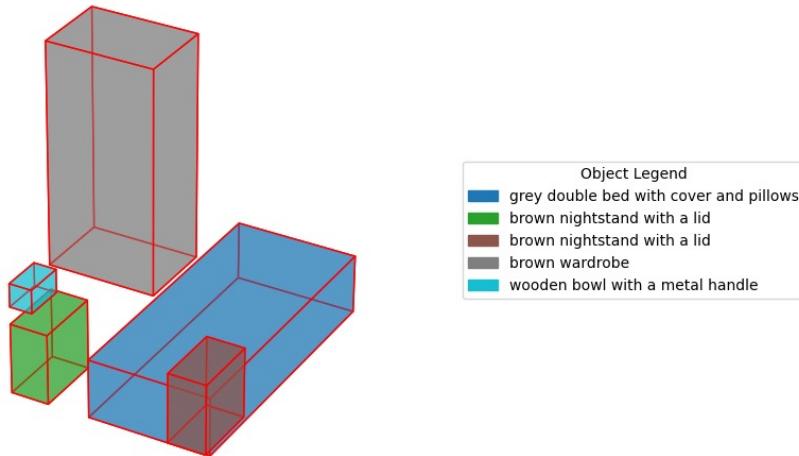


Figure 17: **3D Layout Generation: Case 1.** Generation of a bedroom layout.

### Prompt

I have a bedroom with a dressing table that has a drawer and a shelf, two black nightstands with a vase and books, a black and gray cabinet with two drawers, a green blanket stool, a black and white double bed, a black pendant lamp with metal rods, and a black wardrobe with doors and a drawer.

### O1-preview

```
[{"object": "black and white double bed",  
 "bounding_box": {"width": 1.6, "depth": 2.0, "height": 0.5},  
 "coordinates": [{"x": 0.00, "y": -3.50, "z": 0.00}],  
 "rotate": [{"angle": 0}],  
 {"object": "black nightstand with a vase and books",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6},  
 "coordinates": [{"x": -1.50, "y": -3.50, "z": 0.00}],  
 "rotate": [{"angle": 0}],  
 {"object": "black nightstand with a vase and books",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6},  
 "coordinates": [{"x": 1.50, "y": -3.50, "z": 0.00}],  
 "rotate": [{"angle": 0}],  
 {"object": "black wardrobe with doors and a drawer",  
 "bounding_box": {"width": 1.5, "depth": 0.6, "height": 2.0},  
 "coordinates": [{"x": -4.50, "y": 0.00, "z": 0.00}],  
 "rotate": [{"angle": 90}],  
 {"object": "dressing table with a drawer and a shelf",  
 "bounding_box": {"width": 1.2, "depth": 0.5, "height": 0.75},  
 "coordinates": [{"x": 0.00, "y": 4.50, "z": 0.00}],  
 "rotate": [{"angle": 180}],  
 {"object": "green blanket stool",  
 "bounding_box": {"width": 1.0, "depth": 0.5, "height": 0.45},  
 "coordinates": [{"x": 0.00, "y": 3.50, "z": 0.00}],  
 "rotate": [{"angle": 180}],  
 {"object": "black and gray cabinet with two drawers",  
 "bounding_box": {"width": 1.0, "depth": 0.5, "height": 1.0},  
 "coordinates": [{"x": 4.50, "y": 0.00, "z": 0.00}],  
 "rotate": [{"angle": -90}],  
 {"object": "black pendant lamp with metal rods",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.5},  
 "coordinates": [{"x": 0.00, "y": 0.00, "z": 2.50}],  
 "rotate": [{"angle": 0}]}]
```

### 3D Layout Visualization

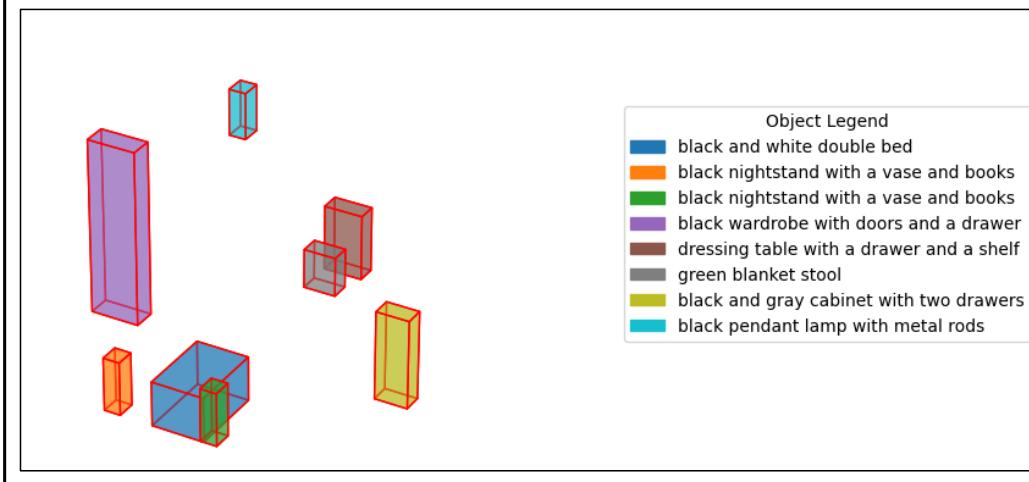


Figure 18: **3D Layout Generation: Case 2.** Generation of a bedroom layout.

### Prompt

This living room has a wooden coffee table, two wooden corner side tables with a pinecone on top, a metal cage pendant lamp, a wooden cabinet, and a wooden TV stand with a drawer. It also features a gray and yellow multi-seat sofa and a lounge chair with a pillow.

### O1-preview

```
[{"object": "grey double bed with cover and pillows", "bounding_box": {"width": 1.6, "depth": 2.0, "height": 0.5}, "coordinates": [{"x": 0.0, "y": -1.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "brown nightstand with a lid", "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6}, "coordinates": [{"x": -1.05, "y": -1.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "brown nightstand with a lid", "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.6}, "coordinates": [{"x": 1.05, "y": -1.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "brown wardrobe", "bounding_box": {"width": 1.5, "depth": 0.6, "height": 2.0}, "coordinates": [{"x": -2.2, "y": 0.0, "z": 0.0}], "rotate": [{"angle": 90}]}, {"object": "wooden bowl with a metal", "bounding_box": {"width": 2.5, "depth": 1.0, "height": 0.8}, "coordinates": [{"x": 0.0, "y": 3.5, "z": 0.0}], "rotate": [{"angle": 180}]}, {"object": "wooden coffee table", "bounding_box": {"width": 1.2, "depth": 0.6, "height": 0.45}, "coordinates": [{"x": 0.0, "y": 1.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "wooden TV stand with a drawer", "bounding_box": {"width": 1.5, "depth": 0.5, "height": 0.6}, "coordinates": [{"x": 0.0, "y": -3.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "lounge chair with a pillow", "bounding_box": {"width": 0.8, "depth": 0.9, "height": 0.9}, "coordinates": [{"x": -3.0, "y": 2.5, "z": 0.0}], "rotate": [{"angle": 45}]}, {"object": "wooden corner side table with a pinecone on top", "bounding_box": {"width": 0.6, "depth": 0.6, "height": 0.7}, "coordinates": [{"x": -4.5, "y": 3.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "wooden corner side table with a pinecone on top", "bounding_box": {"width": 0.6, "depth": 0.6, "height": 0.7}, "coordinates": [{"x": 4.5, "y": 3.5, "z": 0.0}], "rotate": [{"angle": 0}]}, {"object": "wooden cabinet", "bounding_box": {"width": 1.0, "depth": 0.5, "height": 2.0}, "coordinates": [{"x": 4.5, "y": 0.0, "z": 0.0}], "rotate": [{"angle": -90}]}, {"object": "metal cage pendant lamp", "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.5}, "coordinates": [{"x": 0.0, "y": 1.5, "z": 2.5}], "rotate": [{"angle": 0}]}, {"object": "gray and yellow multi-seat handle", "bounding_box": {"width": 0.3, "depth": 0.3, "height": 0.2}, "coordinates": [{"x": -1.05, "y": -1.5, "z": 0.75}], "rotate": [{"angle": 0}]]
```

### 3D Layout Visualization

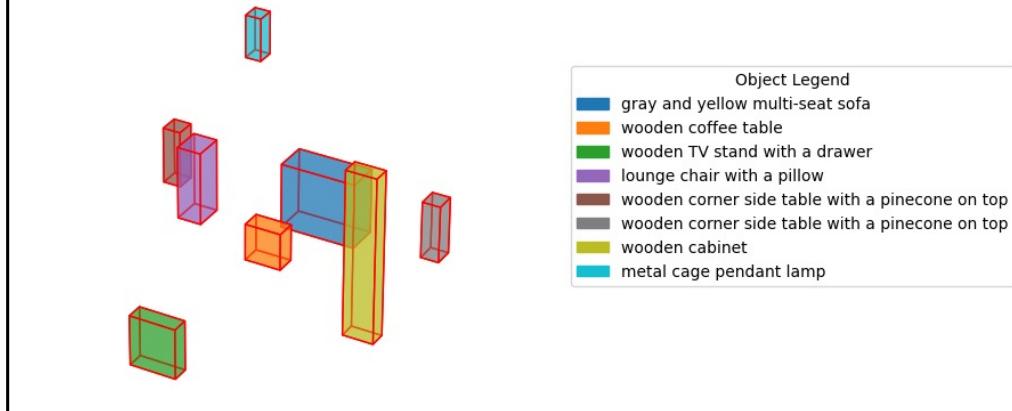


Figure 19: **3D Layout Generation: Case 3.** Generation of a living room layout.

### Prompt

In this living room, there are four black dining chairs with a frame, a black and gold TV stand with doors, a black and gold coffee table, a black pendant lamp with a shade, a black metal dining table, a purple and pink flower children cabinet, and a loveseat chesterfield sofa.

### O1-preview

```
[{"object": "black metal dining table",  
 "bounding_box": {"width": 1.5, "depth": 0.9, "height": 0.75},  
 "coordinates": [{"x": -2.50, "y": 0.00, "z": 0.00}], "rotate": [{"angle": 0}]},  
 {"object": "black dining chair with a frame",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.9},  
 "coordinates": [{"x": -2.50, "y": 1.00, "z": 0.00}], "rotate": [{"angle": 180}]},  
 {"object": "black dining chair with a frame",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.9},  
 "coordinates": [{"x": -2.50, "y": -1.00, "z": 0.00}], "rotate": [{"angle": 0}]},  
 {"object": "black dining chair with a frame",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.9},  
 "coordinates": [{"x": -3.25, "y": 0.00, "z": 0.00}], "rotate": [{"angle": 90}]},  
 {"object": "black dining chair with a frame",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.9},  
 "coordinates": [{"x": -1.75, "y": 0.00, "z": 0.00}], "rotate": [{"angle": -90}]},  
 {"object": "loveseat chesterfield sofa",  
 "bounding_box": {"width": 1.5, "depth": 0.9, "height": 0.9},  
 "coordinates": [{"x": 2.50, "y": 2.00, "z": 0.00}], "rotate": [{"angle": 180}]},  
 {"object": "black and gold coffee table",  
 "bounding_box": {"width": 1.0, "depth": 0.6, "height": 0.45},  
 "coordinates": [{"x": 2.50, "y": 0.50, "z": 0.00}], "rotate": [{"angle": 0}]},  
 {"object": "black and gold TV stand with doors",  
 "bounding_box": {"width": 1.5, "depth": 0.5, "height": 0.6},  
 "coordinates": [{"x": 2.50, "y": -2.00, "z": 0.00}], "rotate": [{"angle": 0}]},  
 {"object": "purple and pink flower children cabinet",  
 "bounding_box": {"width": 1.2, "depth": 0.5, "height": 1.0},  
 "coordinates": [{"x": 0.00, "y": -3.50, "z": 0.00}], "rotate": [{"angle": 0}]},  
 {"object": "black pendant lamp with a shade",  
 "bounding_box": {"width": 0.5, "depth": 0.5, "height": 0.5},  
 "coordinates": [{"x": 2.50, "y": 0.50, "z": 2.50}], "rotate": [{"angle": 0}]]
```

### 3D Layout Visualization

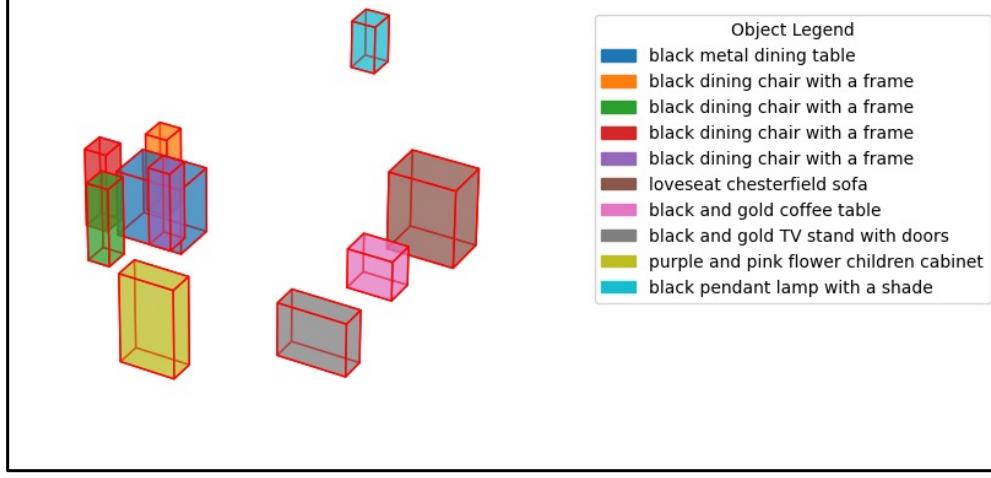


Figure 20: **3D Layout Generation: Case 4.** Generation of a living room layout.

---

## 4.11 Chip Design

In this study, we conducted a comprehensive evaluation of OpenAI’s o1-preview model within the context of chip design, a domain where LLMs show considerable promise but have yet to be fully explored. Our focus was on testing the model’s performance across three critical dimensions: Engineering Assistant Chatbot, Electronic Design Automation (EDA) Script Generation, and Bug Summary & Analysis. These dimensions represent pivotal tasks in semiconductor design and production, requiring deep technical expertise, problem-solving abilities, and precision—all areas where advanced LLMs could potentially offer transformative benefits.

Given the lack of standardized datasets specific to chip design, we adopted the ChipNeMo [90] test sample, a pioneering example of domain-adapted LLMs in this field, as a reference framework. This allowed us to rigorously assess o1-preview’s capabilities while making meaningful comparisons between its performance and that of ChipNeMo, a model specifically adapted for chip design.

### 4.11.1 Engineering Assistant Chatbot

In this section, we evaluate o1-preview’s performance as an engineering assistant chatbot. In the first example (Figure 21), the prompt asks how to run multiple seeds of random tests using the testgen framework. ChipNeMo’s response was functional but relatively basic, offering a clear, step-by-step approach to configuring the random test seeds. However, it lacked deeper insights into how this process could be optimized, especially when scaling for large simulations—a crucial factor in chip design where extensive random testing is essential for validating design robustness. On the other hand, o1-preview’s response was more comprehensive. In addition to covering the basic configuration steps, it offered a deeper explanation on how to parallelize the test process, ensuring that the tests were not only random but also scalable. This detailed approach reflects a higher level of expertise, as it anticipates potential challenges such as random number generation integrity and error handling during large-scale simulations, which are critical for efficient testing in high-stakes environments like semiconductor design. From a professional chip design perspective, o1-preview’s response demonstrates superior problem-solving skills, particularly in managing the intricacies of large-scale test simulations, making it the more valuable solution for real-world engineering scenarios.

In the second example (Figure 22), the task was to write a SystemVerilog assertion to check if signal A becomes true within five clock cycles of signal B being true. ChipNeMo’s response was technically correct, providing a solid and clear assertion that could be implemented effectively. However, while the explanation was accurate, it did not delve into any advanced verification challenges, such as handling asynchronous inputs or potential pitfalls that might arise in different clocking environments. o1-preview, in contrast, not only provided a correct assertion but also took the opportunity to explain how this assertion could behave in different scenarios, such as varying clock domains and frequency mismatches. Additionally, o1-preview discussed how engineers could modify the assertion for more complex verification setups, ensuring that it accounted for edge cases that are common in real-world chip designs. This level of detail and foresight is indicative of a deeper understanding of the verification process, making o1-preview’s response significantly more useful to engineers dealing with complex, time-sensitive verification tasks in chip design.

In the third example (Figure 23), the query was about retiming stages and their purpose in digital circuit design. ChipNeMo provided a thorough response, outlining the key benefits of retiming, such as reducing critical path delays, improving throughput through pipelining, and optimizing power consumption. While this explanation was comprehensive, it presented a list of advantages in a somewhat generic manner without exploring the more intricate trade-offs involved in retiming

---

decisions. o1-preview, however, offered a more nuanced response. In addition to explaining the same core benefits, it discussed the potential drawbacks of retiming, such as increased latency in certain paths or complications with meeting tight setup and hold time requirements. Furthermore, o1-preview offered advice on how to balance retiming with other optimization techniques, such as logic restructuring, to achieve the best overall design outcome. This consideration of trade-offs and broader optimization strategies demonstrates a more expert-level understanding of the circuit design process. From an engineering standpoint, o1-preview’s ability to not only highlight the advantages but also address the potential downsides of retiming provides a far more insightful and practical response.

#### 4.11.2 EDA Script Generation

In the context of EDA Script Generation, both o1-preview and ChipNeMo were tested on their ability to generate scripts for EDA tools, a critical task in chip design automation. These scripts are pivotal in automating various stages of chip development, such as logic verification and circuit optimization, which require both precise coding and a deep understanding of the chip design environment.

In Figure 24, one such prompt tasked the models with writing a TOOL1 code to count the number of flip-flop cells within a given rectangular boundary, a common operation in physical design when determining logic distribution and optimizing layout. ChipNeMo’s response provided a basic solution, covering the mechanics of how to traverse the design hierarchy and count the flip-flops in the specified region. While it delivered the correct structure and functionality, the response was somewhat limited in its flexibility. The code did not explore advanced features like handling edge cases (such as overlapping regions or boundary conditions) or optimizing the search algorithm for larger designs, both of which are essential considerations in production environments where efficiency is crucial.

In contrast, o1-preview’s response offered a more nuanced solution. The model not only provided the core code but also discussed potential optimization strategies, such as how to efficiently traverse large designs and how to handle boundary conditions more gracefully. Moreover, o1-preview included comments and suggestions for extending the functionality, such as adding the ability to count other types of cells or integrating the script into a larger automation flow for design rule checking. This response demonstrated a deeper understanding of the complexities involved in real-world EDA environments, where flexibility and scalability are paramount. o1-preview was able to anticipate potential issues that engineers might face when deploying this script in large-scale designs, making its response significantly more applicable for practical use in a professional setting.

Shown in Figure 25, another query asked the models to write a TOOL2 TCL procedure to find intersecting logic between two pins, a task that typically involves traversing the netlist and analyzing connections between logic elements. ChipNeMo’s response was straightforward, generating a functional TCL procedure that correctly identified the logic elements between two pins. However, the response lacked depth in terms of optimizing the logic traversal for large-scale designs or handling more complex netlist structures with multiple levels of hierarchy, which are often encountered in modern chip design.

o1-preview, on the other hand, delivered a more comprehensive response by not only generating the procedure but also discussing the nuances of handling different types of netlist structures, such as flat versus hierarchical designs. The model provided suggestions on optimizing the traversal algorithm to minimize runtime in large designs, an essential consideration in EDA where scripts often need to handle millions of logic elements. Furthermore, o1-preview anticipated potential edge cases, such as how to handle pins with multiple fan-outs or fan-ins, and proposed extensions to the procedure for

---

handling such scenarios. This level of detail illustrates a greater understanding of the complexities involved in chip design and the need for adaptable, efficient scripts in EDA environments.

#### 4.11.3 Bug Summary & Analysis

In the context of Bug Summary & Analysis, both o1-preview and ChipNeMo were tested on their ability to analyze bug reports related to silicon current measurement in chip fabrics, a crucial task in chip power optimization shown in Figure 26. Bug analysis in semiconductor design involves not only identifying the root cause of the problem but also summarizing the issue in a way that allows teams to prioritize and resolve it efficiently.

For the given bug, the report concerns a request to measure the power consumption of fabric macros in a chip under varying levels of activity and voltage conditions. This is a typical challenge during the design-for-test (DFT) phase, where engineers need to evaluate how closely the silicon's power consumption aligns with pre-silicon estimates. The goal is to isolate the power consumed by specific parts of the chip to improve the accuracy of power models and optimize power efficiency.

ChipNeMo's response was technically sound but somewhat formulaic in structure. It provided both a technical summary and a managerial summary, effectively breaking down the core components of the issue. The technical summary outlined the conditions for power measurement, such as test module activity percentages and the required voltage corners. However, the managerial summary simply relayed the back-and-forth discussion between users without delving into any deeper insights or offering a clear action plan. ChipNeMo's response focused heavily on summarizing the immediate facts of the bug report without providing much context or additional analysis on the implications of the power measurement for future designs or suggesting strategies for correlating silicon data with pre-silicon estimates.

In contrast, o1-preview's response offered a much more thorough analysis of the bug. While covering the same technical and managerial aspects as ChipNeMo, o1-preview also addressed the broader implications of the power measurement. It analyzed how the fabric power data could influence future chip designs, particularly in optimizing power density and improving area efficiency—factors that are critical in modern chip development. Furthermore, o1-preview provided a more in-depth interpretation of the request, offering suggestions on how to approach the power measurement with more refined isolation techniques and data breakdowns to ensure more accurate comparisons with pre-silicon estimates. It also provided a clearer task assignment framework, ensuring that the engineering teams could prioritize the tasks with timelines and specific milestones for isolating and measuring the fabric macros' power consumption.

From a bug analysis perspective, o1-preview not only responded to the immediate needs of the bug report but also demonstrated a deeper understanding of the challenges posed by such tasks in real-world chip development. Its response took into account both the technical and strategic dimensions of the issue, offering a more valuable solution for engineers tasked with optimizing chip power consumption. By anticipating potential challenges and offering concrete suggestions on data analysis and task delegation, o1-preview demonstrated superior problem-solving capabilities compared to ChipNeMo.

In conclusion, across the three critical tasks—Engineering Assistant Chatbot, EDA Script Generation, and Bug Summary & Analysis—o1-preview consistently demonstrated superior problem-solving capabilities, depth of analysis, and practical relevance compared to ChipNeMo. In the Engineering Assistant Chatbot task, o1-preview excelled by offering more comprehensive solutions, anticipating

---

edge cases, and providing optimization insights that reflected real-world engineering challenges. In EDA Script Generation, o1-preview again showed a greater understanding of scalability and flexibility, delivering solutions that could be adapted to complex, large-scale designs. Its ability to optimize code and handle intricate netlist structures demonstrated advanced reasoning that would be invaluable to engineers working on chip automation tasks. Finally, in Bug Summary & Analysis, o1-preview's thorough analysis and strategic foresight allowed it to offer not just a summary of the bug but also suggestions for power optimization and task prioritization. This comprehensive approach underscores o1-preview's capacity to engage with both the technical and managerial aspects of chip design, making it a powerful tool in handling complex semiconductor development tasks. Ultimately, o1-preview's consistent performance highlights its potential to play a pivotal role in advancing chip design automation and enhancing engineering workflows.

## **Prompt**

Write a System Verilog assert to check if signal A is true within 5 cycles of signal B being true

## **O1-preview**

To assert that signal **A** becomes true within 5 cycles after signal **B** becomes true in SystemVerilog, you can use the following assertion:

```
assert property (@(posedge clk) B |-> ##[1:5] A);
```

Explanation:

- @(posedge clk)**: This specifies that the assertion is clocked on the positive edge of the clock signal *clk*.
- B**: This is the antecedent of the implication; it checks when signal **B** is true.
- |->**: The non-overlapping implication operator, indicating that the consequent must follow after the antecedent is true.
- ##[1:5]**: This specifies a delayed sequence, meaning the consequent should occur 1 to 5 clock cycles after the antecedent.
- A**: This is the consequent; we're checking that signal **A** becomes true within the specified delay.

This assertion effectively checks that whenever **B** is true, **A** must become true at least once within the next 1 to 5 clock cycles.

**Note:** Ensure that you have a clock signal (*clk*) defined in your simulation environment for this assertion to work correctly.

## **Final Assertion:**

```
assert property (@(posedge clk) B |-> ##[1:5] A);
```

## **Reference Answer**

Sure, here is the System Verilog assert to check if signal A is true → within 5 cycles of signal B being true:

```
assert property (@(posedge clk) disable iff
  ~rst_n)
  (signal_B && !signal_A) |-> ##[1:5] signal_A
);
```

This assert checks that if signal B is true and signal A is false, then signal A will become true within the next 5 cycles. The '##[1:5]' syntax is used to specify a delay range from 1 to 5 cycles. The 'disable iff (~rst n)' clause ensures that the assert is not triggered during reset.

Figure 21: **Chip Design-Engineering Assistant Chatbot: Case 1.** The comparison between ChipNeMo and o1-preview in generating a SystemVerilog assertion to verify signal A within 5 cycles of signal B. o1-preview's response offered additional context, addressing clocking environments and real-world scenarios, making it more adaptable to complex design verification processes.

## **Prompt**

Why do we add retiming stages?

## **O1-preview**

Adding retiming stages in digital circuit design is a technique used to optimize synchronous circuits by repositioning the storage elements (such as flip-flops or registers) without altering the circuit's functional behavior. This process, known as **retiming**, is instrumental in enhancing the performance and efficiency of digital systems.

Here are the primary reasons why retiming stages are added:

### **1. Increase Clock Speed by Reducing Critical Path Delay:**

- Critical Path Minimization:** The critical path is the longest path that a signal must traverse between two sequential elements in a clock cycle. By repositioning registers through retiming, designers can shorten the critical path, allowing the circuit to operate at a higher clock frequency.
- Balanced Delay Distribution:** Retiming helps in distributing combinational delays more evenly across the circuit, preventing any single stage from becoming a bottleneck.

### **2. Improve Throughput via Pipelining:**

- Enhanced Parallelism:** Adding retiming stages effectively pipelines the circuit. This means multiple data inputs can be processed simultaneously at different stages, increasing the overall throughput.
- Latency Management:** While pipelining may introduce some latency, the increased data processing rate often outweighs this drawback in high-throughput applications.

### **3. Optimize Power Consumption:**

- Dynamic Power Reduction:** By enabling the circuit to run at a higher clock frequency with lower voltage levels, retiming can contribute to reduced dynamic power consumption.
- Glitch Minimization:** Shorter combinational paths reduce transient glitches, which can cause unnecessary switching and power wastage.

### **4. Enhance Signal Integrity and Reliability:**

- Reduced Signal Degradation:** Breaking long combinational paths with registers can mitigate issues like signal degradation and crosstalk, improving overall signal integrity.
- Improved Noise Margins:** Retimed circuits tend to have better noise tolerance due to reduced signal propagation times.

### **5. Meet Timing Constraints and Design Specifications:**

- Setup and Hold Time Compliance:** Adjusting the placement of registers helps ensure that setup and hold time requirements are met, which is crucial for reliable circuit operation.
- Clock Skew Management:** Retiming can alleviate the effects of clock skew by strategically placing registers to synchronize data paths.

### **6. Facilitate Resource Optimization:**

- Area Efficiency:** By balancing the computational load across the circuit, retiming can lead to a more efficient utilization of logic resources, potentially reducing the overall chip area.
- Cost Reduction:** Optimized circuits may require fewer components or simpler interconnects, leading to cost savings in manufacturing.

### **7. Simplify Design and Maintenance:**

- Modularity:** Retiming can make the circuit more modular, simplifying debugging, testing, and future modifications.
- Scalability:** A retimed circuit is often more scalable, as it can be more easily adjusted to meet new performance requirements.

## **Reference Answer**

Retiming stages are added to improve the timing of the design and to account for the physical distance between different parts of the chip. By adding retiming stages, we can reduce the amount of time it takes for signals to travel between different parts of the chip, which can help to improve the overall performance of the design. Additionally, retiming stages can also help to reduce the amount of noise and interference that can occur on the signals, which can further improve the performance of the design.

Figure 22: **Chip Design-Engineering Assistant Chatbot: Case 2.** This figure illustrates the responses from ChipNeMo and o1-preview on the purpose of adding retiming stages. o1-preview provided a comprehensive discussion on the trade-offs and broader context, offering a more expert-level analysis suitable for optimizing performance and resource utilization in high-frequency digital designs.

## Prompt

How do I run multiple seeds of random tests using testgen framework?

## O1-preview

To run multiple seeds of random tests using the TestGen framework, you can automate the test generation process by iterating over a range of seed values. This allows you to generate a diverse set of test cases with different random inputs, which is particularly useful for robust testing of algorithms that should behave correctly under various conditions.

Here's how you can do it:

1. **Identify the Seed Option:** Determine how the TestGen framework accepts a seed value. This is typically done via a command-line option like `--seed`, `-s`, or an environment variable.
2. **Create a Script to Iterate Over Seeds:** Write a script in your preferred scripting language (e.g., Bash, Python) that runs the test generation command multiple times, each time with a different seed value.

### Example Using Bash Script

```
#!/bin/bash
# Specify the number of seeds you want to run
NUM_SEEDS=10
# Loop over the seed values
for ((seed=1; seed<=NUM_SEEDS; seed++))
do
    echo "Generating test case with seed $seed"
    testgen --seed $seed --output test_case_${seed}.txt
done
```

- Replace `testgen` with the actual command used to invoke the TestGen framework.
- Ensure that `--seed` (or the correct option) is used to set the seed.
- The `--output` option directs the output to a file named uniquely for each seed.

### Example Using Python Script

```
import subprocess
NUM_SEEDS = 10
for seed in range(1, NUM_SEEDS + 1):
    print(f"Generating test case with seed {seed}")
    subprocess.run([
        'testgen',
        '--seed', str(seed),
        '--output', f'test_case_{seed}.txt'
    ])
```

3. **Parallel Execution (Optional):** If you want to speed up the process, you can run the test generations in parallel. Tools like GNU `parallel` can help with this.

### Using GNU Parallel

```
seq 1 10 | parallel -j+0 testgen --seed {} --output test_case_{}.txt
```

- Generate seeds from 1 to 10.
- Use `parallel` to run multiple jobs simultaneously (the `-j+0` option allows it to use as many cores as available).

- 4. Check TestGen Documentation:** Some TestGen frameworks may have built-in support for generating multiple tests with different seeds in a single command. For example:

```
testgen --seeds 1 2 3 4 5 --output-dir tests/
```

- This hypothetical command would generate tests for seeds 1 through 5 and store them in the *tests/* directory.
- 5. Automate and Integrate with CI/CD:** If you're using Continuous Integration/Continuous Deployment (CI/CD) pipelines, integrate your script into the pipeline so that tests are generated and run automatically with each build.
- 6. Handle Randomness in Tests:** Ensure that your tests are deterministic per seed. This means that running the test generation with the same seed should always produce the same test case. This is crucial for reproducibility when a test fails and you need to debug.

#### Tips

- **Logging:** Keep logs of which seeds have been used and any associated test results. This makes it easier to trace back any issues.
- **Resource Management:** Generating a large number of tests can consume significant resources. Monitor the CPU and memory usage and adjust the number of parallel jobs accordingly.
- **Incremental Seeds:** If you need a large number of seeds, consider using a more sophisticated method to generate seed values, such as random selection or based on timestamps.

#### Example Scenario

Suppose you're testing a sorting algorithm implementation, and you want to ensure it works correctly for various input sizes and patterns. By running:

```
for seed in {1..100}
do
  testgen --seed $seed --generate-sorting-input > input_$seed.txt
  my_sorting_algorithm < input_$seed.txt > output_$seed.txt
  verify_output output_$seed.txt
done
```

- Generates 100 different input files for the sorting algorithm using seeds 1 to 100.
- Runs the sorting algorithm on each input.
- Verifies the output of each run.

#### Conclusion

Running multiple seeds in the TestGen framework involves:

- Identifying how to set the random seed in your specific framework.
- Scripting the execution over multiple seeds.
- Optionally, optimizing the process with parallel execution.

By automating this process, you enhance your testing coverage and increase the likelihood of catching edge cases and random failures.

#### Reference Answer

You can run multiple seeds of random tests using the Testgen framework by using the `rand_tgen` target. This target allows you to specify a variety of options for the stopping criteria of random runs using the same Testgen testlist(s).

```
nvmk -c . rand_tgen TESTLIST=my_testlist MON=1
→ TGEN LOOP=50
```

This command will run 50 iterations of the *my\_testlist*, each with a different seed. You can specify the number of iterations and the seed mode using the `TGEN LOOP` and `SEED` options, respectively.

Figure 23: **Chip Design-Engineering Assistant Chatbot: Case 3.** The comparison between ChipNeMo and o1-preview in addressing how to run multiple seeds of random tests using the testgen framework. o1-preview offered a detailed, scalable solution, highlighting its ability to handle large-scale random testing, a critical requirement in design verification.

## **Prompt**

Write a code in TOOL1 to print the number of flop cells in the given rectangle (0,0,100,100)

## **O1-preview**

To solve this task, we'll use a script that interacts with TOOL1's design database to analyze the physical placement of cells in a given design. Since Verilog is primarily used for describing hardware behavior and structure rather than scripting or querying design databases, EDA tools often provide scripting capabilities using languages like Tcl or their proprietary scripting languages.

Below is an example of a Tcl script that could be used in an EDA tool like Synopsys IC Compiler II or Cadence Innovus to count the number of flip-flop cells within the specified rectangle (0,0,100,100). This script assumes that the tool uses Tcl as its scripting language and that the design database can be queried for cell instances, their types, and locations.

```
# Define the rectangle coordinates
set lower_left_x 0
set lower_left_y 0
set upper_right_x 100
set upper_right_y 100
# Initialize the flip-flop count
set flop_count 0
# Get all sequential cells (flip-flops) in the design
set flop_cells [get_cells -filter {is_sequential == true}]
# Iterate over each flip-flop cell
foreach flop $flop_cells {
    # Get the location of the flip-flop
    set location [get_attribute $flop location]
    set x_pos [lindex $location 0]
    set y_pos [lindex $location 1]
    # Check if the flip-flop is within the rectangle
    if { $x_pos >= $lower_left_x && $x_pos <= $upper_right_x &&
        $y_pos >= $lower_left_y && $y_pos <= $upper_right_y }{
        # Increment the flip-flop count
        incr flop_count
    }
}
# Print the total number of flip-flops found in the rectangle
puts "Number of flip-flops in rectangle (0,0)-(100,100): $flop_count"
```

## **Explanation:**

1. **Define Rectangle Coordinates:**
  - We specify the lower-left and upper-right coordinates of the rectangle.
2. **Initialize Flip-Flop Count:**
  - A variable flop\_count is initialized to keep track of the number of flip-flops found within the rectangle.
3. **Retrieve Flip-Flop Cells:**
  - The script retrieves all sequential cells (which are typically flip-flops) in the design using a filter that checks for is\_sequential == true.

4. **Iterate Over Flip-Flops:**
  - For each flip-flop cell, the script retrieves its location (x and y coordinates).
5. **Check Location Within Rectangle:**
  - The script checks if the flip-flop's location falls within the specified rectangle boundaries.
6. **Increment Count:**
  - If the flip-flop is within the rectangle, the flop\_count is incremented.
7. **Output Result:**
  - After iterating through all flip-flops, the script prints out the total count.

**Notes:**

- **Tool-Specific Functions:**
  - The functions `get_cells`, `get_attribute`, and `puts` are commonly available in EDA tool scripting environments but may vary slightly depending on the specific tool (TOOL1). You might need to adjust the script to match the exact syntax and functions provided by TOOL1.
- **Sequential Cells Identification:**
  - The property `is_sequential` is used to identify flip-flop cells. Ensure that this property or an equivalent is available in your tool to filter flip-flops accurately.
- **Coordinate Units:**
  - The units for the coordinates (0,0,100,100) should match the units used in the design database (e.g., micrometers).
- **Error Handling:**
  - For production scripts, consider adding error handling to manage cases where cells might not have location attributes or other unexpected conditions.

**Conclusion:**

This script provides a way to count the number of flip-flop cells within a specific rectangular area in a design using TOOL1's scripting capabilities. Adjust the script as necessary to fit the exact syntax and functions of TOOL1.

**Reference Answer**

```
# Sure, here is a TOOL1 code to print the number of
→ flop cells in the given rectangle
→ (0,0,100,100):

# This code gets the top_ref of the design
top_ref = env().get_top_reference()
# This code gets the cells of the rectangle in the
→ top_ref
cells = env().layout().cells_in_rectangle( top_ref,
→ rectangle(0, 0, 100, 100), hierarchical)
# This code gets the cells that are flop cells
flop_cells = cells.type("is_flop_cell")
# This code prints the number of flop cells
print(flop_cells.size())
```

Figure 24: **Chip Design-EDA Script Generation: Case 1.** This figure demonstrates the responses of ChipNeMo and o1-preview when tasked with generating a TOOL1 code to count flip-flop cells in a defined region. o1-preview extended the solution with performance optimization techniques for large-scale designs, offering deeper insight into real-world physical design challenges.

## **Prompt**

Write a TOOL2 tcl proc to find intersecting logic from pins "\$pinA"" to pin "\$pinB"

## **O1-preview**

To find the intersecting logic between two pins \$pinA and \$pinB in TOOL2 using Tcl scripting, you can write a Tcl procedure that traverses the downstream logic cone of \$pinA (fanout cone) and the upstream logic cone of \$pinB (fanin cone), and then computes the intersection of these two sets of logic elements.

Below is an example Tcl code that accomplishes this. The code includes helper procedures to traverse the fanout and fanin cones recursively, collect the logic elements, and compute their intersection.

```
# Helper procedure to compute the intersection of two lists
proc list_intersection {list1 list2} {
    set result {}
    foreach item $list1 {
        if {[lsearch -exact $list2 $item] != -1} {
            lappend result $item
        }
    }
    return $result
}
# Helper procedure to traverse fanout recursively and collect logic elements
proc traverse_fanout {pin visited} {
    if {[lsearch -exact $visited $pin] == -1} {
        lappend visited $pin
        # Get nets driven by this pin
        set nets [get_nets -of_objects $pin -driving]
        foreach net $nets {
            # Get load pins (input pins of downstream cells) connected to this net
            set load_pins [get_pins -of_objects $net -filter {direction == input}]
            foreach load_pin $load_pins {
                # Recursively traverse downstream pins
                traverse_fanout $load_pin visited
            }
        }
    }
}
# Helper procedure to traverse fanin recursively and collect logic elements
proc traverse_fanin {pin visited} {
    if {[lsearch -exact $visited $pin] == -1} {
        lappend visited $pin
        # Get nets connected to this pin (driven by)
        set nets [get_nets -of_objects $pin -driven_by]
        foreach net $nets {
            # Get driver pins (output pins of upstream cells) connected to this net
            set driver_pins [get_pins -of_objects $net -filter {direction == output}]
            foreach driver_pin $driver_pins {
                # Recursively traverse upstream pins
                traverse_fanin $driver_pin visited
            }
        }
    }
}
```

```

# Main procedure to find intersecting logic between pinA and pinB
proc find_intersecting_logic {pinA pinB} {
    # Initialize visited lists for fanout and fanin traversal
    set visited_fanout [list]
    set visited_fanin [list]

    # Traverse fanout cone of pinA
    traverse_fanout $pinA visited_fanout

    # Traverse fanin cone of pinB
    traverse_fanin $pinB visited_fanin

    # Compute intersection of the two visited lists
    set intersecting_logic [list_intersection $visited_fanout $visited_fanin]

    # Return the intersecting logic elements
    return $intersecting_logic
}

# Example usage:
# Replace these with your actual pin names
set pinA "top_module/inst1/output_pin"
set pinB "top_module/inst2/input_pin"

# Call the procedure to find intersecting logic
set intersecting_logic [find_intersecting_logic $pinA $pinB]

# Print the intersecting logic
puts "The intersecting logic between $pinA and $pinB is:"
foreach logic_element $intersecting_logic {
    puts $logic_element
}

```

#### Explanation:

1. **list\_intersection**: This helper procedure takes two lists as input and returns a list of items that are present in both lists.
2. **traverse\_fanout**: This recursive procedure starts from a given pin and traverses all downstream pins by following the nets driven by the current pin. It collects all visited pins in the *visited* list to avoid revisiting the same pin and to accumulate the logic elements.
3. **traverse\_fanin**: Similar to *traverse\_fanout*, this recursive procedure traverses all upstream pins starting from a given pin by following the nets that drive the current pin.
4. **find\_intersecting\_logic**: This is the main procedure that uses *traverse\_fanout* and *traverse\_fanin* to get the downstream logic cone of *\$pinA* and the upstream logic cone of *\$pinB*, respectively. It then computes the intersection of these two logic cones to find the intersecting logic elements.
5. **Example Usage**: Replace "*top\_module/inst1/output\_pin*" and "*top\_module/inst2/input\_pin*" with the actual hierarchical names of your pins *\$pinA* and *\$pinB*. Then, call *find\_intersecting\_logic* with these pins, and the script will output the intersecting logic.

**Notes:**

- The script assumes that your EDA tool (TOOL2) supports standard Tcl commands and provides the following commands:
  - `get_nets`: Retrieves nets associated with specified pins or cells.
  - `get_pins`: Retrieves pins associated with specified nets or cells.
  - `-of_objects`: Specifies the objects (pins, nets, cells) to query.
  - `-driving`: Option to get nets driven by the specified pin.
  - `-driven_by`: Option to get nets that drive the specified pin.
  - `-filter`: Used to filter pins based on their direction (input or output).
- If your EDA tool uses different commands or options, you will need to adjust the script accordingly.
- Be cautious of potentially large logic cones that can lead to long run times or excessive memory usage. For large designs, you might want to add limits or optimization to the traversal procedures.
- This script traverses the netlist at the pin level. Depending on your needs, you might want to modify it to work at the cell level or include additional information about the logic elements.

**Conclusion:**

By using this Tcl script within TOOL2, you can effectively find the intersecting logic between two pins in your Verilog design. This can be particularly useful for debugging, analysis, or optimization tasks in digital design workflows.

**Reference Answer**

```
#Sure, here is the code:
proc myfunc {pinA pinB} {
    # Get all the pins in the fanout of the given pin
    set fanout_pins [fanout -from $pinA]
    # Get all the pins in the fanout of the given pin
    set fanin_pins [fanin -to $pinB]
    # Get the intersection of the two collections
    set intersection [common_logic $fanout_pins
        ↳ $fanin_pins]
    # Return the intersection
    return $intersection
}
```

Figure 25: **Chip Design-EDA Script Generation: Case 2.** The figure shows the generated TCL procedures from ChipNeMo and o1-preview for finding intersecting logic between two pins. o1-preview provided an optimized approach for handling large designs and discussed edge cases, making it more applicable to real-world circuit analysis tasks.

## **Prompt**

BugId : 002

Synopsis : Request power/current measurement on fabrics using fabric test module

Module : Silicon- DFT

Description :

Hi all, We want to evaluate silicon current (power) consumed by fabric macros in a chip. This will help us correlate this power to pre-Si power estimates.

Here are the conditions we want to look at for this data:

Fabric test module activity: 0/25/50/75/100%

Corners:

typical low voltage

typical high voltage

We want to isolate fabric current (power) as much as possible, so we want to enable only the logic needed for fabric test modules and the fabrics themselves.

If this information can be broken down per module (there are 6 in the chip), that will help.

Severity : 5-Performance

Priority : 5-Normal

Comments :

#1 : Author : User A

Is this a requirement only for current or for future chips as well ?

#2 : Author : User B

There will likely be interest in future chips to perform this exercise as we push the boundaries of fabric power density for better area.

#3 : Author : User A

What is needed in this bug ?

@User B It is not clear.

Can you list the specific needs with timelines and priority

## **O1-preview**

Hi @User A,

Thank you for reaching out. I appreciate the opportunity to clarify our needs. Below is a detailed list of our specific requirements, along with proposed timelines and priorities.

### **Specific Needs**

#### **1. Measurement of Fabric Macro Power Consumption**

- Objective: Obtain accurate silicon current (power) measurements consumed by the fabric macros in the chip.
- Purpose: To correlate these measurements with our pre-silicon power estimates, enhancing the accuracy of our power modeling and estimation methodologies.

#### **2. Test Conditions**

- **Fabric Test Module Activity Levels:** We need to measure power consumption at varying levels of activity to understand how power scales with workload. The specific activity levels are:

- **0% Activity:** Idle state with no activity.
- **25% Activity**
- **50% Activity**
- **75% Activity**
- **100% Activity:** Full operational capacity.

- **Voltage Corners:**

- **Typical Low Voltage (Vmin):** The lowest voltage level within the typical operating range of the chip.
- **Typical High Voltage (Vmax):** The highest voltage level within the typical operating range.

### 3. Isolation of Fabric Power

- **Requirement:** Isolate the power consumption of the fabric macros as much as possible.
  - Only enable the necessary logic for the fabric test modules and the fabrics themselves during measurements.
  - Disable or power gate other non-essential logic blocks to minimize their impact on the measurements.

### 4. Breakdown

- **Details:** There are **6 fabric modules** within the chip.
- **Requirement: If feasible, provide power consumption data broken down per individual fabric module.**
  - This granularity will help in pinpointing modules that may deviate from expected power consumption and assist in targeted optimization efforts.

## Timelines

### 1. Preparation Phase

- **By: November 10, 2023**
- Activities:
  - Finalize test setups and procedures.
  - Ensure that the environments (test benches, measurement tools) are calibrated and ready.
  - Coordinate with the DFT and silicon validation teams to align on test methodologies.

### 2. Measurement Phase

- **Start: November 13, 2023**
- **End: November 24, 2023**
- Activities:
  - Conduct power measurements at all specified activity levels and voltage corners.
  - Repeat measurements to ensure data reliability and account for variability.

### 3. Data Analysis and Correlation

- **Start: November 27, 2023**
- **End: December 1, 2023**
- Activities:
  - Analyze the collected data.
  - Correlate silicon measurements with pre-silicon power estimates.
  - Identify discrepancies and areas for improvement.

### 4. Reporting

- **By: December 5, 2023**
- Deliverables:
  - A detailed report summarizing the findings.
  - Graphs and charts illustrating power consumption across different activity levels and voltage corners.
  - Recommendations for adjustments to power estimation models.

## Priority

- **Severity:** Initially marked as **5 - Performance**, but given the impact on power modeling accuracy and future chip designs, I recommend adjusting to **3 - Significant**.
- **Priority:** Initially marked as **5 - Normal**. Considering the importance for ongoing and future projects, I propose raising it to **2 - High**.

## Justification for Priority Adjustment

- **Impact on Future Designs:** The insights gained will be crucial as we develop chips with higher fabric power density for better area efficiency.
- **Optimization:** Early identification of power discrepancies allows for timely optimization, potentially reducing development time and costs.

- **Strategic Importance:** Accurate power modeling is essential for meeting performance and efficiency targets in competitive markets.

#### **Additional Notes**

- **Collaboration:** We are ready to collaborate closely with the relevant teams to facilitate this effort.
- **Resource Allocation:** Please let us know if additional resources or support are needed from our side to meet these timelines.
- **Future Considerations:** While this request is specific to the current chip, establishing this measurement and correlation process will be beneficial for all future chip designs.

Please let me know if this clarifies our needs or if further information is required. We are keen to proceed and align on the next steps.

Best regards,

User B

### **Reference Answer**

#### **Technical Summary:**

In the internal bug management system, a bug has been filed regarding the power and current measurements of the fabrics in the latest chip. The request is to evaluate the silicon current (power) consumed by the fabric macros in the latest chip and to correlate this power to pre-Si power estimates. The conditions for the measurements are the fabric test module activity at 0/25/50/75/100% and the corners are typical\_low\_voltage and typical\_high\_voltage. The data needs to be broken down per module (there are 6 in the latest chip). Only the logic needed for fabric test modules and the fabrics themselves need to be enabled.

#### **Managerial Summary:**

User A asks User B if the need for evaluating the silicon current of fabric macros in a chip is only for that specific chip or for future ones as well. User B replies that there will likely be interest in future chips to perform this exercise as they push the boundaries of fabric power density for better area. User A then asks what is needed in this bug and asks for specific needs with timelines and priority.

#### **Task Assignment:**

User B

Figure 26: **Bug Summary & Analysis: Case 1.** Bug report analysis comparing ChipNeMo and o1-preview for evaluating power consumed by fabric macros. While ChipNeMo provided a standard summary, o1-preview offered deeper analysis with strategic recommendations for power isolation and task prioritization, essential for power optimization in silicon designs.

## 4.12 Logical Reasoning

o1-preview has demonstrated powerful capabilities in the tests of logical reasoning problems. It has abundant knowledge related to logical reasoning. Judging from the test results, whether it is a regular logical reasoning question or a complex logical reasoning problem with a certain degree of difficulty, such as those involving multi - level nested reasoning or multi - condition comprehensive judgment, it can provide very accurate and well - organized analysis processes and answers, and its analysis ability is extremely outstanding.

In the process of in-depth research and testing on the reasoning ability of o1-preview, we have discovered a noteworthy phenomenon. The reasoning of o1-preview cannot be completely mapped accurately to the given standard answer. This implies that in some specific problem scenarios, o1-preview may not be able to provide results that are entirely consistent with the traditional standard answers, such as those in Figures 28, 107, 110, 111, 112. However, despite this shortcoming, the thinking process demonstrated by o1-preview is astonishing. When confronted with various complex reasoning tasks, o1-preview can rapidly analyze and break down problems. It will attempt to examine problems from different angles, just like a perceptive thinker exploring unknown realms. Whether dealing with logical puzzles, making situational inferences, or analyzing complex relationship networks, o1-preview can display its unique way of thinking. For instance, when facing a complex logical reasoning problem, o1-preview may not directly give the conclusion in the standard answer. Instead, it will analyze each element of the problem step by step and propose some novel thinking paths and possibilities. This exploratory thinking process provides us with a new perspective and prompts us to re-examine the essence and solution methods of problems.

It can analyze complex problems, break them down into smaller components, and systematically work through them to arrive at accurate conclusions. Whether it's solving mathematical puzzles, deducing the next step in a sequence, or evaluating the validity of an argument, o1-preview demonstrates a high level of intelligence and precision. See Figures 27, 105, 106, 108, 109. Its ability to handle multiple variables and conditions simultaneously is particularly impressive. It can consider different scenarios, weigh the pros and cons of each option, and make informed decisions. This not only showcases its logical acumen but also its adaptability to various types of problems.

The thinking process of o1-preview also reflects its strong learning and adaptability. Although it may deviate from the standard answer on some problems, by continuously receiving new data and feedback, o1-preview can gradually optimize its reasoning model and continuously approach the correct answer. This continuous improvement ability brings great potential for its future applications. It can present its thought process in a way that is easy to understand, making it a valuable tool for education, research, and problem-solving in various fields. However, although its error rate is relatively low, there are still some cases of mistakes. For example, on some problems that require the exploration of very subtle logical relationships, it may occasionally have deviations in reasoning. For most logical reasoning tests, only the basic information of the problem needs to be provided without particularly detailed guidance, and o1-preview can successfully complete the reasoning analysis. Moreover, it can quickly process a large number of different types of logical reasoning problems. Nevertheless, when facing extremely complex problems involving the intersection of multiple different logical systems, o1-preview may encounter bottlenecks. This is a limitation of o1-preview in handling complex logical reasoning tasks.

What we would like to point out is that the official standard answer is not necessarily accurate. Instead, the analysis and reasoning process offered by o1-preview is truly convincing. From our perspective, this is far more gratifying than relying on the standard answer.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

The dust removal advantage of broadleaf trees is obvious, and the adsorption effect of PM2.5 is the best, and the annual average dust retention of a broadleaf tree is 3.16 kg. Conifers have a small area, and the adsorption effect of PM2.5 is weak. On average throughout the year, broadleaf trees are much better at vacuuming than conifers. Broad-leaved trees are also much better than shrubs and grasses in vacuuming effect. Taking the Chinese sophora, a common broad-leaved tree in Beijing, as an example, the vacuuming effect of the Chinese sophora tree is about 30% higher than that of the common grassland of the same area. According to this, some people believe that in order to reduce dust, Beijing should vigorously promote broadleaf trees and minimize the area of coniferous forests.

Which of the following, if true, would most undermine the views of those concerned?

- A. Although conifers do not shed leaves in winter, they are basically in A "dormant" state and have poor biological activity.
- B. The proportion of broadleaf trees and conifers is unbalanced, which is not only easy to break out diseases, pests, fires, etc., but also affects the growth and health of trees.
- C. Plant afforestation not only needs to control PM2.5, but also needs to control other pollutants, which needs to be rationally distributed.
- D. broad-leaved trees shed their leaves in winter, and in cold winter, their maintenance costs are much higher than conifers.

### **O1-preview**

The passage suggests that broadleaf trees are superior to conifers in removing dust and PM2.5 particles, leading some to recommend promoting broadleaf trees and minimizing coniferous forests in Beijing to reduce dust. To undermine this view, we need to introduce information that challenges the effectiveness or practicality of this recommendation.

Option A mentions that conifers are dormant in winter and have poor biological activity. This doesn't undermine the recommendation because it aligns with the argument that conifers are less effective than broadleaf trees in dust removal.

Option D states that broadleaf trees have higher maintenance costs in winter due to shedding leaves. While this suggests a downside to broadleaf trees, it doesn't necessarily outweigh their dust removal benefits nor directly challenge the conclusion.

Option C suggests that afforestation needs to address pollutants other than PM2.5 and requires rational distribution. This implies a need for varied plant types but doesn't directly refute the superiority of broadleaf trees in dust removal.

Option B, however, introduces significant drawbacks to minimizing coniferous forests and over-promoting broadleaf trees. It states that an imbalance in the proportion of broadleaf trees and conifers can lead to increased risks of diseases, pests, fires, and negatively affect tree health. This directly undermines the recommendation by highlighting serious ecological and environmental risks associated with reducing coniferous forests, suggesting that a balanced mix of tree species is necessary for a healthy ecosystem.

**\*\*Answer: Option B\*\***

Figure 27: **Logical Reasoning: Case 1.** In this case, In this sample, the o1-preview model demonstrated excellent logical analysis performance and obtained the answer consistent with the expectation.

## **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

There are three types of robots in a laboratory. Type A can recognize colors, type B can recognize shapes, and type C can neither recognize colors nor shapes. The laboratory experimented with robots 1 and 2 using red balls, basketballs, red squares, and blue squares, and ordered them to pick up the red balls, but number 1 picked up the red squares, and number 2 picked up the basketball.

Based on the above experiment, which of the following must be true?

- A. Both No.1 and No.2 must be type C.
- B. Both No.1 and No.2 must not be type C.
- C. No.1 is A and No.2 is Type B.
- D. No.1 is not B and No.2 is not Type A.

## **O1-preview**

Based on the given information, we need to determine which statement must be true about Robots 1 and 2.

**Summary of the Situation:**

- Type A robots can recognize colors.
- Type B robots can recognize shapes.
- Type C robots can recognize neither colors nor shapes.
- The lab has four items: red balls, basketballs, red squares, and blue squares.
- Robots are ordered to pick up red balls.
- Robot 1 picks up the red squares.
- Robot 2 picks up the basketball.

**Analyzing Robot 1:**

- Picked up red squares when told to pick up red balls.
- The object picked is red but the wrong shape.
- This suggests Robot 1 recognizes colors but not shapes.
- Therefore, Robot 1 must be Type A.

**Analyzing Robot 2:**

- Picked up the basketball when told to pick up red balls.
- The object picked is the correct shape (ball) but the wrong color.
- This suggests Robot 2 recognizes shapes but not colors.
- Therefore, Robot 2 must be Type B.

**Considering Type C Robots:**

- Type C robots cannot recognize colors or shapes.
- If they cannot recognize either, their selection would be random.
- It is highly unlikely for both robots to coincidentally pick items that match part of the instruction if they were selecting randomly.
- Therefore, neither Robot 1 nor Robot 2 is Type C.

**Evaluating the Options:**

- Option A: Both robots are Type C (False).
- Option B: Both robots are not Type C (True).
- Option C: Robot 1 is Type A, and Robot 2 is Type B (True).
- Option D: Robot 1 is not Type B, and Robot 2 is not Type A (True).

While options B, C, and D are all true, Option C provides the most specific and direct conclusion based on the evidence. **Answer: C.No.1 is A and No.2 is Type B.**

Figure 28: **Logical Reasoning: Case 2.** There are also test results that are different from the answers provided on the official website. From the perspective of the standard answers of the test questions, these are failed samples.

---

### 4.13 Table-to-Text Generation

The o1-preview model has demonstrated strong performance on the table-to-text generation task using the ADNI dataset, particularly in transforming clinical measurements into coherent and accurate narratives. The examples of table-to-text generation for the ADNI datasets are shown in (Figure 29) and (Figure 30). Its ability to interpret and present complex clinical data—such as patient demographics, cognitive scores, and imaging biomarkers—has proven both reliable and precise, offering valuable insights for medical research and reporting. The model’s output not only captures the nuances of the data but also maintains the integrity of the clinical information, making it a promising tool for enhancing data-driven storytelling in healthcare. Based on the experimental results, the model provides well-organized and accurate analyses, demonstrating exceptional capabilities in summarizing and interpreting clinical data.

During our in-depth testing of o1-preview’s table-to-text abilities, we observed an intriguing phenomenon. In cases where certain clinical measurements, such as a patient’s data, were marked as "NaN" (Not a Number), o1-preview didn’t simply output this label. Instead, it intelligently inferred that the data was missing or unavailable, offering a more meaningful interpretation. This highlights the model’s ability to not only reflect the raw data but also apply a deeper understanding of its context. Such features suggest o1-preview could be instrumental in automating comprehensive and insightful reports, further elevating the potential of AI in medical documentation and analysis.

### 4.14 High School Level Math Competition

To effectively assess the logical reasoning capability of the o1-preview, we conducted 10 test cases, 5 algebra problems and 5 counting and probability problems, spanning difficulty levels from 1 to 5. The o1-preview demonstrated outstanding performance, achieving 100% accuracy across all 10 test cases. This consistent performance on problems of varying difficulty demonstrates its ability to handle both simple and more advanced mathematics problems. Beyond just providing correct final answers, the o1-preview’s solutions have also demonstrated detailed, step-by-step explanations which include the key steps found in the reference solutions. This level of detailed solutions suggest that the model is approaching a human-like capability in logical reasoning within the domain of mathematics.

In case 1 (Figure 31), a level 5 algebra problem was presented, asking the model to determine the maximum difference between the radii of two circles given a constraint on the difference in their areas. The problem involved both geometric and algebraic reasoning, as the model needed to consider the difference of squares in the area equation and apply the given constraints to determine the maximum difference between the length of the radii. The o1-preview accurately identified the essential steps of utilizing the area difference formula and solving the inequality, eventually reaching the final answer of 0.5. The solution was detailed and the key steps aligned with the reference answer, highlighting the model’s capability to handle complex geometric problems with logical steps.

In case 2 (Figure 32), the problem involved calculating the positive difference between 120% of 30 and 130% of 20. Although this was only a level 1 algebra problem involving basic percentage calculations, it could still display the o1-preview’s logical approach if it could clearly present the essential steps. In the o1-preview’s solution, it converted percentages into decimals, multiplied them by the appropriate values, and computed the absolute difference between the results. The o1-preview calculated the correct answer of 10, with the key steps shared with the reference answer.

### **Prompt**

According to given examples. Please translate the tabular data into readable paragraphs.

Table:

Sub\_id": "002\_S\_0295", "Age": 84.8, "Sex": "Male", "Edu": 18, "Ethnic": "Not Hisp/Latino", "Race": "White", "Marital": "Married", "EXAMDATE": "2006-11-02", "BI\_DX": "CN", "DX": "CN", "APOE4": 1.0, "FDG": NaN, "ABETA": NaN, "TAU": NaN, "PTAU": NaN, "CDRSB": 0.0, "ADAS11": 5.33, "ADAS13": 6.33, "MMSE": 28.0, "ADASQ4": 1.0, "RAVLT\_immediate": 50.0, "RAVLT\_learning": 7.0, "RAVLT\_forgetting": 2.0, "RAVLT\_perc\_forgetting": 16.6667, "LDELTOTAL": NaN, "DIGITSCOR": 43.0, "TRABSCOR": 68.0, "mPACCdigit": -0.294118, "mPACCTrailsB": 0.367504, "FAQ": 0.0, "FLDSTRENG": "1.5 Tesla MRI", "FSVERSION": "Cross-Sectional FreeSurfer (FreeSurfer Version 4.3)", "Ventricles": 39101.0, "Hippocampus": 6896.0, "WholeBrain": 1089680.0, "Entorhinal": 3662.0, "Fusiform": 18041.0, "MidTemp": 19825.0, "ICV": 1645300.0.

### **O1-preview**

Basic Personal Information:

The subject is an 84.8-year-old male who has completed 18 years of education. He identifies as Not Hispanic/Latino and White. Marital status is married.

Biomarker Measurements:

The subject's genetic profile includes an ApoE4 status of 1.0. **FDG-PET data is unavailable**. Cerebrospinal fluid analysis data is unavailable for A $\beta$ 42, total tau protein, and phosphorylated tau levels.

Cognitive and Neurofunctional Test Results:

The Mini-Mental State Examination score is 28.0. The Clinical Dementia Rating, sum of boxes, is 0.0. **ADAS11 and ADAS13 scores are 5.33 and 6.33** respectively, with a delayed word recall score of 1.0. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 50.0, learning score at 7.0, forgetting score at 2.0, and percent forgetting at 16.67%. The Logical Memory test for delayed recall is unavailable. The Digit Symbol Substitution test resulted in a score of 43.0. Trails B testing showed a score of 68.0. The modified Preclinical Alzheimer's Cognitive Composite scores using Digit Symbol Substitution and Trails B are **-0.294118 and 0.367504** respectively. The Functional Activities Questionnaire rated the subject at 0.0.

Imaging Data Points:

Under MRI conditions at a field strength of 1.5 Tesla, using Cross-Sectional FreeSurfer (Version 4.3), the imaging data recorded includes ventricles volume at **39,101.0**, hippocampus volume at 6,896.0, whole brain volume at 1,089,680.0, entorhinal cortex volume at **3,662.0**, fusiform gyrus volume at 18,041.0, and middle temporal area volume at 19,825.0. The intracranial volume measured is **1,645,300.0**.

### **Reference Answer**

Basic Personal Information: Subject 002\_S\_0295 is a 84.8-year-old Male. He/She has completed 18 years of education. The ethnicity is Not Hisp/Latino and race is White. Marital status is Married. Biomarker

Measurements: The subject's genetic profile includes an ApoE4 status of 1.0. **Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being nan**. Cerebrospinal fluid analysis reveals A $\beta$ 42 levels at nan, total tau protein levels at nan, and phosphorylated tau levels at nan. Cognitive

and Neurofunctional Test Results: The Mini-Mental State Examination score stands at 28.0. The Clinical Dementia Rating, sum of boxes, is 1.0. **ADAS 11 and 13 scores are 5.33 and 6.33** respectively, with a score of 1.0 in delayed word recall. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 50.0, learning score at 7.0, forgetting score at 2.0, and percent forgetting at 16.67. Logical Memory test for delayed recall scored nan. The Digit Symbol Substitution test resulted in a score of 43.0. Trails B testing showed a score of 68.0. The modified Preclinical Alzheimer's Cognitive Composite scores using Digit Symbol Substitution and Trails B are **-0.294118 and 0.367504** respectively. The Functional Activities Questionnaire rated the subject as 0.0.

Imaging Data Points: Under MRI conditions at a field strength of 1.5 Tesla MRI Tesla, using Cross-Sectional FreeSurfer (FreeSurfer Version 4.3), the imaging data recorded includes ventricles volume at **39101.0**, hippocampus volume at 6896.0, whole brain volume at 1089680.0, entorhinal cortex volume at **3662.0**, fusiform gyrus volume at 18041.0, and middle temporal area volume at 19825.0. The intracranial volume measured is **1645300.0**.

Figure 29: **Table-to-Text: Case 1.** A demonstration case of ADNI tabular data and its corresponding generated report. Green highlights indicate that o1-preview precisely repeats the measurements, while yellow suggests that o1-preview may have its own interpretation of the tabular data during text generation.

### **Prompt**

According to given examples. Please translate the tabular data into readable paragraphs.

Table:

Sub\_id": "002\_S\_0295", "Age": 84.8, "Sex": "Male", "Edu": 18, "Ethnic": "Not Hisp/Latino", "Race": "White", "Marital": "Married", "EXAMDATE": "2007-05-25", "BI\_DX": "CN", "DX": "CN", "APOE4": 1.0, "FDG": NaN, "ABETA": 858.3, "TAU": 399.5, "PTAU": 39.29, "CDRSB": 0.0, "ADAS11": 4.67, "ADAS13": 5.67, "MMSE": 30.0, "ADASQ4": 1.0, "RAVLT\_immediate": 53.0, "RAVLT\_learning": 10.0, "RAVLT\_forgetting": 2.0, "RAVLT\_perc\_forgetting": 14.2857, "LDELTOTAL": 15.0, "DIGITSCOR": 54.0, "TRABSCOR": 72.0, "mPACCdigit": 3.12291, "mPACCTrailsB": 2.40573, "FAQ": 0.0, "FLDSTRENG": "1.5 Tesla MRI", "FSVERSION": "Cross-Sectional FreeSurfer (FreeSurfer Version 4.3)", "Ventricles": 40554.0, "Hippocampus": 6894.0, "WholeBrain": 1085210.0, "Entorhinal": 3572.0, "Fusiform": 17386.0, "MidTemp": 19917.0, "ICV": 1651500.0

### **O1-preview**

Basic Personal Information:

The subject is an 84.8-year-old male who has completed 18 years of education. He identifies as Not Hispanic/Latino and White. Marital status is married.

Biomarker Measurements:

The subject's genetic profile includes an ApoE4 status of 1.0. FDG-PET data is unavailable. Cerebrospinal fluid analysis reveals A $\beta$ 42 levels at 858.3, total tau protein levels at 399.5, and phosphorylated tau levels at 39.29.

Cognitive and Neurofunctional Test Results:

The Mini-Mental State Examination score is 30.0. The Clinical Dementia Rating, sum of boxes, is 0.0. ADAS11 and ADAS13 scores are 4.67 and 5.67 respectively, with a delayed word recall score of 1.0. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 53.0, learning score at 10.0, forgetting score at 2.0, and percent forgetting at 14.29%. The Logical Memory test for delayed recall scored 15.0. The Digit Symbol Substitution test resulted in a score of 54.0. Trails B testing showed a score of 72.0. The modified Preclinical Alzheimer's Cognitive Composite scores using Digit Symbol Substitution and Trails B are 3.12291 and 2.40573 respectively. The Functional Activities Questionnaire rated the subject at 0.0.

Imaging Data Points:

Under MRI conditions at a field strength of 1.5 Tesla, using Cross-Sectional FreeSurfer (Version 4.3), the imaging data recorded includes ventricles volume at 40,554.0, hippocampus volume at 6,894.0, whole brain volume at 1,085,210.0, entorhinal cortex volume at 3,572.0, fusiform gyrus volume at 17,386.0, and middle temporal area volume at 19,917.0. The intracranial volume measured is 1,651,500.0.

### **Reference Answer**

Basic Personal Information: Subject 002\_S\_0295 is a 84.8-year-old Male. He/She has completed 18 years of education. The ethnicity is Not Hisp/Latino and race is White. Marital status is Married. Biomarker

Measurements: The subject's genetic profile includes an ApoE4 status of 1.0. Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being nan. Cerebrospinal fluid analysis reveals A $\beta$ 42 levels at 858.3, total tau protein levels at 399.5, and phosphorylated tau levels at 39.29.

Cognitive and Neurofunctional Test Results: The Mini-Mental State Examination score stands at 30.0. The Clinical Dementia Rating, sum of boxes, is 1.0. ADAS 11 and 13 scores are 4.67 and 5.67 respectively, with a score of 1.0 in delayed word recall. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 53.0, learning score at 10.0, forgetting score at 2.0, and percent forgetting at 2.0. Logical Memory test for delayed recall scored 15.0. The Digit Symbol Substitution test resulted in a score of 54.0. Trails B testing showed a score of 72.0.

The modified Preclinical Alzheimer's Cognitive Composite scores using Digit Symbol Substitution and Trails B are 3.12291 and 2.40573 respectively. The Functional Activities Questionnaire rated the subject as 0.0. Imaging Data Points: Under MRI conditions at a field strength of 1.5 Tesla MRI Tesla, using Cross-Sectional FreeSurfer (FreeSurfer Version 4.3), the imaging data recorded includes ventricles volume at 40554.0, hippocampus volume at 6894.0, whole brain volume at 1085210.0, entorhinal cortex volume at 3572.0, fusiform gyrus volume at 17386.0, and middle temporal area volume at 19917.0. The intracranial volume measured is 1651500.0.

Figure 30: **Table-to-Text: Case 2.** A demonstration case of ADNI tabular data and its corresponding generated report.

---

Case 3 (Figure 33) presented a combinatorial problem involving the principle of counting and probability to determine how many students at a school played on exactly two sports teams. The o1-preview adeptly handled this problem by first defining variables for each team and calculating the team sizes using the given ratios. The o1-preview then applied the inclusion-exclusion principle to account for the overlaps between teams, particularly focusing on the number of students playing on exactly two teams. The final solution of the o1-preview matched the reference answer of 14 and the o1-preview's solution demonstrated its capability in handling the counting problems, such as overcounting and managing complex combinatorial reasoning tasks.

In case 4 (Figure 34), the o1-preview was asked to solve a algebraic problem that involved determining how many non-congruent isosceles triangles exist with integer side lengths and a perimeter of 10. The problem required the use of the triangle inequality theorem to validate potential side lengths while ensuring that no two triangles were congruent. The o1-preview systematically considered all possible combinations of side lengths, verified each combination using the triangle inequality, and correctly identified that two such triangles satisfied the conditions. The key steps in o1-preview's solution mirrored those in the reference answer, with the green-highlighted final answer and yellow-highlighted steps indicating the shared reasoning between the model's solution and the reference. This case further illustrates the o1-preview's strength in handling algebraic and combinatorial problems, showcasing its ability to systematically integrate algebraic and geometric reasoning.

Another notable observation is the speed at which the o1-preview generated solutions. For most cases, the model produced correct answers in just a few seconds, reflecting its efficiency in processing and reasoning. This instant response time is promising for real-time applications in mathematical problem-solving. However, for some of the more challenging level 5 problems, the o1-preview took up to 50 seconds to generate solutions. This prolonged thinking time may indicate that the o1-preview is engaging in more intricate reasoning, similar to how humans approach difficult problems. While this observation suggests the model may exhibit human-like thought processes under higher difficulty, further investigation is necessary to draw definitive conclusions about its human-like thinking process in these scenarios.

### **Prompt**

You have two circles, one with radius  $r$  and the other with radius  $R$ . You wish for the difference in the areas of these two circles to be less than or equal to  $5\pi$ . If  $r+R=10$ , what is the maximum difference in the lengths of the radii?

### **O1-preview**

To find the maximum difference in the lengths of the radii  $R$  and  $r$  given the constraints, we'll follow these steps:

**Given:**

- $R + r = 10$
- The difference in the areas  $\leq 5\pi$

#### **Step 1: Write the Difference in Areas**

The area of a circle is  $A=\pi r^2$ . The difference in areas between the two circles is:

$$\Delta A = \pi R^2 - \pi r^2 = \pi(R^2 - r^2)$$

#### **Step 2: Factor the Difference of Squares**

Factor  $R^2 - r^2$ :

$$\Delta A = \pi(R-r)(R+r)$$

#### **Step 3: Apply the Area Constraint**

Set the difference in areas less than or equal to  $5\pi$ :

$$\pi(R-r)(R+r) \leq 5\pi$$

Divide both sides by  $\pi$ :

$$(R-r)(R+r) \leq 5$$

#### **Step 4: Use the Sum of Radii**

From  $R+r=10$ , substitute into the inequality:

$$(R-r)(10) \leq 5$$

#### **Step 5: Solve for $R-r$**

Divide both sides by 10:

$$R-r \leq 0.5$$

Simplify:

$$R-r \leq 0.5$$

#### **Conclusion:**

The maximum difference in the lengths of the radii is **0.5 units**.

**Answer: 0.5**

### **Reference Answer**

We want to solve the inequality  $\pi R^2 - \pi r^2 \leq 5\pi$ . Dividing both sides by  $\pi$ , we get:

$$R^2 - r^2 \leq 5$$

Next, we factor the left-hand side:

$$(R+r)(R-r) \leq 5$$

Now, substituting  $R+r=10$  into the inequality, we have:

$$10(R-r) \leq 5$$

Dividing both sides by 10:

$$R-r \leq 0.5$$

Thus, the maximum difference between the lengths of the radii is **0.5**.

Figure 31: **High School Level Math Competition: Case 1: Algebra.** This level 5 algebra problem involves determining the maximum difference between the radii of two circles given an area constraint. The yellow highlights indicate shared steps in both the o1-preview output and the reference answer, while the green highlights mark the shared final answers.

### **Prompt**

What is the positive difference between 120% of 30 and 130% of 20?

### **O1-preview**

To find the positive difference between 120% of 30 and 130% of 20, we'll follow these steps:

#### **Step 1: Calculate 120% of 30**

- Convert the percentage to a decimal:

$$120\% = 120/100 = 1.20$$

#### **Multiply by 30:**

$$1.20 \times 30 = 36$$

**Result:** 120% of 30 is 36.

#### **Step 2: Calculate 130% of 20**

- Convert the percentage to a decimal:

$$130\% = 130/100 = 1.30$$

#### **Multiply by 20:**

$$1.30 \times 20 = 26$$

**Result:** 130% of 20 is 26.

#### **Step 3: Find the Positive Difference**

- Subtract the two results:

$$\text{Difference} = |36 - 26| = 10$$

(We take the absolute value to ensure the difference is positive.)

#### **Conclusion:**

The positive difference between 120% of 30 and 130% of 20 is 10.

**Answer: 10**

### **Reference Answer**

One hundred twenty percent of 30 is  $120 * 30 * (1/100) = 36$ , and 130% of 20 is  $130 * 20 * (1/100) = 26$ . The difference between 36 and 26 is 10

Figure 32: **High School Level Math Competition: Case 2: Algebra.** This level 1 algebra problem involves calculating the positive difference between two percentage values. The yellow highlights mark the shared intermediate steps shared in both the o1-preview's output and the reference answer, while the green highlights emphasize the correct final answers. The o1-preview efficiently completed this problem in 4 seconds.

### **Prompt**

At a school, all 60 students play on at least one of three teams: Basketball, Soccer, and Mathletics. 8 students play all three sports, half the students play basketball, and the ratio of the size of the math team to the size of the basketball team to the size of the soccer team is \$4:3:2\$. How many students at the school play on exactly two teams?

### **O1-preview**

To determine how many students play on exactly two teams, we'll use the information given and apply the principle of inclusion-exclusion.

#### **Step 1: Assign Variables and Given Values**

Let:

- M = Number of students on the **Mathletics** team
- B = Number of students on the **Basketball** team
- S = Number of students on the **Soccer** team
- $n(M \cap B)$  = Number of students on **both Mathletics and Basketball**
- $n(M \cap S)$  = Number of students on **both Mathletics and Soccer**
- $n(B \cap S)$  = Number of students on **both Basketball and Soccer**
- $n(M \cap B \cap S)$  = Number of students on **all three teams** (given as 8)

Given:

- Total students, N=60
- Half the students play basketball:  $B=1/2 \times 60=30$
- Ratio of team sizes:  $M:B:S=4:3:2$
- Number of students who play all three sports:  $n(M \cap B \cap S)=8$

#### **Step 2: Determine Team Sizes Using Ratios**

Given  $B=30$  and the ratio  $M:B:S=4:3:2$ :

- Find the ratio multiplier k:  $B=3k \Rightarrow 3k=30 \Rightarrow k=10$
- Calculate M and S:  $M=4k=4 \times 10=40 \quad S=2k=2 \times 10=20$

#### **Step 3: Apply the Principle of Inclusion-Exclusion**

The total number of students is:

$$N=M+B+S-n(M \cap B)-n(M \cap S)-n(B \cap S)+n(M \cap B \cap S)$$

Substitute known values:

$$60=40+30+20-n(M \cap B)-n(M \cap S)-n(B \cap S)+8$$

Simplify:

$$60=90-[n(M \cap B)+n(M \cap S)+n(B \cap S)]+8$$

$$[n(M \cap B)+n(M \cap S)+n(B \cap S)]=90+8-60=38$$

#### **Step 4: Account for the Triple Overlap**

Each pairwise intersection includes the students who play on exactly two teams **and** those who play on all three teams.

Therefore:

$$n(M \cap B)=\text{Students on exactly Mathletics and Basketball}+8$$

Similarly for  $n(M \cap S)$  and  $n(B \cap S)$ .

Let:

- x = Number of students on exactly Mathletics and Basketball
- y = Number of students on exactly Mathletics and Soccer
- z = Number of students on exactly Basketball and Soccer

So:

$$x+y+z=[n(M \cap B)+n(M \cap S)+n(B \cap S)]-3 \times 8=38-24=14$$

#### **Step 5: Conclude the Number of Students on Exactly Two Teams**

The total number of students playing on exactly two teams is:

$$x+y+z=14$$

**Answer: 14**

### **Reference Answer**

We have enough information to solve for the size of each team. There are **60/2 = 30 members of the basketball team**, **4/3 \* 30 = 40 members of the math team**, and **2/3 \* 30 = 20 members of the soccer team**. Adding these up gives us 90, so clearly we're overcounting since there are only 60 students. The number of times that each student is counted in this total is equal to the number of teams that the student plays on. This means that all 60 students will be counted at least once, all students who play exactly two sports will be counted one extra time, and all students who play three sports will be counted two extra times. Letting x be the number of students who play two sports and y be the number who play all three gives us  $60 + x + 2y = 90$ . However, we know that  $y = 8$ , so we can substitute that in and get **x = 14**.

Figure 33: High School Level Math Competition: Case 3: Counting and Probability. This level 5 counting and probability problem involves determining how many students play on exactly two sports teams. The yellow highlights mark shared steps in both o1-preview's output and the reference answer, showing the team size calculations. The green highlights indicate the shared final answer. The o1-preview completed this high level problem in an extended time of 52 seconds.

### **Prompt**

How many (non-congruent) isosceles triangles exist which have a perimeter of 10 and integer side lengths?

### **O1-preview**

To determine the number of non-congruent isosceles triangles with integer side lengths and a perimeter of 10, we need to find all sets of integer sides  $a$ ,  $a$ , and  $b$  (or  $b$ ,  $b$ , and  $a$ ) that satisfy the triangle inequality conditions.

#### **Step 1: Understand the Problem Constraints**

- **Perimeter Constraint:**  $2a+b=10$  or  $2b+a=10$
- **Integer Side Lengths:**  $a$  and  $b$  are positive integers.
- **Triangle Inequality:** The sum of any two sides must be greater than the third side.

#### **Step 2: Find Possible Values for $a$ and $b$**

##### **Case 1: Isosceles with Two Equal Sides $a$ , Base $b$**

- **Perimeter Equation:**  $2a+b=10 \Rightarrow b=10-2a$
- **Possible Values for  $a$ :**
  - Since  $b>0$ ,  $10-2a>0 \Rightarrow a<5$ .
  - Triangle inequality  $2a>b$  becomes:  
 $2a>10-2a \Rightarrow 4a>10 \Rightarrow a>2.5$ .
- **Possible Integer Values for  $a$ :**  $a=3, 4$
- **Corresponding  $b$  Values:**
  - For  $a=3$ :  $b=10-6=4$
  - For  $a=4$ :  $b=10-8=2$

##### **Case 2: Isosceles with Two Equal Sides $b$ , Base $a$**

- **Perimeter Equation:**  $2b+a=10 \Rightarrow a=10-2b$
- **Possible Values for  $b$ :**
  - Since  $a>0$ ,  $10-2b>0 \Rightarrow b<5$ .
  - Triangle inequality  $2b>a$  becomes:  $2b>10-2b \Rightarrow 4b>10 \Rightarrow b>2.5$   
So  $b \geq 3$ .
- **Possible Integer Values for  $b$ :**  $b=3, 4$
- **Corresponding  $a$  Values:**
  - For  $b=3$ :  $a=10-6=4$
  - For  $b=4$ :  $a=10-8=2$

#### **Step 3: List the Possible Triangles**

- **Triangle 1:** Sides 3,3,4
- **Triangle 2:** Sides 4,4,2

#### **Step 4: Ensure Triangles are Non-Congruent**

These two triangles are not congruent because they have different side lengths.

- **First Triangle:** Sides 3,3,4
- **Second Triangle:** Sides 4,4,2

#### **Step 5: Conclusion**

There are **two** non-congruent isosceles triangles with integer sides and a perimeter of 10.

**Answer:** 2

### **Reference Answer**

Let  $x$  be the measure of each of the equal sides. Since the perimeter is 10 units, the side lengths measure  $x$ ,  $x$ , and  $10-2x$  units. Since the third side length must be positive, we have  $10-2x>0$  which implies  $x<5$ . By the triangle inequality, the sum of the two equal sides must exceed the third side. Solving  $x+x>10-2x$  gives  $x>2.5$ . There are  $\boxed{2}$  integers strictly between 2.5 and 5.

Figure 34: High School Level Math Competition: Case 4: Counting and Probability. This level 3 counting and probability problem involves determining the number of non-congruent isosceles triangles with integer side lengths and a perimeter of 10. The yellow highlights indicate key steps shared in both the o1-preview output and the reference answer, while the green highlights mark the shared final answer.

## 4.15 College-level Math Problems

To evaluate o1-preview’s ability to solve college-level math problems, we presented the model with 12 manually chosen problems. These problems cover topics including discrete math, Calculus and proofs of advanced math theorems. The answers generated by the o1-preview model were manually evaluated by the authors and categorized as correct, questionable, or incorrect. Due to complexity of the problems, we do not provide reference answers in our Figures. The test results are summarized in Table 5.

Table 5: Number of correct, questionable and incorrect answers generated by the o1-preview model for each type of math problem. The correctness of the answers was manually determined by the authors.

|                      | Basic discrete | Advanced discrete | Calculus | Proofs | Total |
|----------------------|----------------|-------------------|----------|--------|-------|
| correct answers      | 2              | 0                 | 1        | 4      | 7     |
| questionable answers | 0              | 0                 | 1        | 0      | 1     |
| incorrect answers    | 0              | 3                 | 0        | 1      | 4     |
| Total                | 2              | 3                 | 2        | 5      | 12    |

Now, we provide a detailed analysis of each problem and the answer generated by o1-preview. In particular, we provide in-depth analysis for each of the questionable or incorrect answers.

Question 1 (Figure 35) and 2 (Figure 119) are basic discrete math problems from the field of Combinatorics. Both are presented as counting problems involving variables, asking for explicit formulas expressed in terms of those variables. o1-preview solved these successfully, indicating that the model can understand basic discrete patterns and perform symbolic computation effectively.

Question 3 (Figure 120), 4 (Figure 121), 5 (Figure 36) are advanced discrete math problems from the field of elementary number theory. The model struggled the most with this set of problems, failing to solve all three.

To take a closer look into these problems, Question 3 involves proving an identity about two sequences  $\{a_i\}_{i=1}^n$  and  $\{b_i\}_{i=1}^n$  of positive integers provided that they satisfy certain conditions. During test time, after first noticing some easy identities satisfied by the sequences, the model introduced a creative step (highlighted in yellow in Figure 120), proposing that if a specific relation between  $a_i$  and  $b_i$  for each  $i$  could be proved, then the problem would be easily solved. This suggestion is incorrect in the first place, as the suggested relation is not true in general. What is more concerning is the model’s subsequent steps to follow through on its own suggestion. Instead of proving that the relation holds universally for any valid sequences  $\{a_i\}_{i=1}^n$  and  $\{b_i\}_{i=1}^n$ , the model constructed a specific pair of sequences that satisfies the relation (also highlighted in yellow). For this particular pair, the required identity trivially holds. The model then erroneously generalized from this specific case, concluding that the relation must hold in all cases (highlighted in red) and, therefore, claiming to have proven the identity. This approach clearly violates logical principles, as a single example cannot be used to prove a general statement. This error, known as “proof by example”, which is a type of inappropriate generalization, indicates a gap in the model’s reasoning abilities: while the model may generate creative insights, it lacks the rigor required to validate general claims through proper logical reasoning and proof construction. The experiment was repeated a few more times, but the o1-preview model was not able to generate a correct solution.

Question 4 asks for a proof that there always exists a particular ordering of the first  $n$  integers so that

---

when concatenated into a single integer, the resulting number is divisible by 7. One example ( $n = 2$ ) is included in the prompt to help the model understand the question. Initially, the model proposed several ideas, none of which were particularly useful for solving the problem. It eventually resorted to using the pigeonhole principle. This approach was incorrect, as counterexamples can easily be constructed. The authors followed up by highlighting this mistake and providing counterexamples. Although the model acknowledged its mistakes, it ultimately responded that it was unable to solve the problem.

Question 5 is the famous Problem 6 from the 1988 International Mathematical Olympiad (IMO), which requires showing that whenever an integer  $n$  can be written in a specific form, it must be a perfect square (i.e.  $n = k^2$  for some  $k$ ). Given the popularity of this problem, the authors initially expected the model to solve it with ease, but it did not. The model did demonstrate some insight by recognizing that proving the discriminant of specific quadratic is a perfect square would solve the problem. In the subsequent steps, however, the model claimed to observe a pattern by testing small integer values. However, misled by specific test values, the model committed another case of inappropriate generalization, assuming that the observed pattern holds in general. Moreover, this observed pattern is of the form “ $n = b^2$ ”, which overlaps with what it was asked to prove (highlighted in red in Figure 36). This constitutes “circular reasoning”, where the model’s assumption essentially repeats the problem’s conclusion. After a few more attempts, the model kept making similar mistakes and was unable to generate a correct proof. These logical flaws suggest that while the o1-preview model can identify patterns from small examples, it is prone to being misled and making inappropriate generalizations. Additionally, in cases requiring longer reasoning chains, the model may lose track of its initial assumptions and fall into circular reasoning.

Question 6 (Figure 122) and 7 (Figure 37) are advanced Calculus problems. o1-preview performed reasonably well on both. Specifically, Question 6 requires proving a bound for a continuously differentiable function defined on  $[0, 1]$ . Solving this problem requires applying the Fundamental Theorem of Calculus and the Cauchy-Schwartz inequality. The model solved it correctly and efficiently, suggesting that it possesses adequate knowledge of classical theorems and is able to apply them when the clues are obvious.

Question 7 involves testing the convergence of an infinite series. Since such problems are classical and many examples are available on the internet, the prompt specifically asked for two different methods to ensure that the model would not rely on a memorized solution. The model provided two methods, but neither was completely satisfactory. The first method was simply incorrect. The model attempted to compare the given series to a specific subseries of the harmonic series  $\sum_{n=1}^{\infty} \frac{1}{n}$ , concluding the divergence of the given series based on the divergence of the harmonic series (highlighted in red in Figure 37). This reasoning is flawed, as a subseries of a divergent series can still converge. The second method, while correct, relied on advanced techniques from graduate-level Real Analysis (highlighted in yellow). Given the overall inadequacy of the answers, particularly the incorrect first method and the overly advanced second method, we deemed the model’s responses to be questionable.

The last five problems are all advanced theorem proof problems.

Question 8 (Figure 123) and 8.5 (Figure 124) come from the field of Real Analysis, with Question 8.5 being a follow-up to Question 8. Both involve verifying certain topological properties of infinite dimensional spaces. The model successfully solved both problems.

Question 9 (Figure 125), which was picked from a textbook on Functional Analysis [67], proved to be challenging to the model. o1-preview failed to solve it particularly due to a misunderstanding of

---

linear independence in the context of infinite dimensional spaces.

Question 10 (Figure 126) and 10.5 (Figure 127) are on Hilbert’s Nullstellensatz, one of the most famous theorems from the field of Algebraic Geometry. The two questions are essentially the same, but each allows the model to assume a different intermediate result. This design ensures that the model pursues distinct approaches, rather than relying on a memorized solution. The model performed well on both questions, providing two different proofs that align with their respective intermediate results.

To close this subsection, we provide a summary of the strengths and limitations of o1-preview on college-level mathematical reasoning tasks.

Regarding the strengths, we have empirically observed that:

- Overall, o1-preview possesses comprehensive knowledge of college-level mathematical terminologies and classical theorems. Even when problems involve advanced concepts like the function space  $C[0, 1]$  or topological interior and closure, the model had no difficulty understanding the problem statements. Moreover, when the problem can be solved by directly applying one or two theorems, no matter how advanced they are, the model can effectively identify and apply the appropriate theorems. This in particular applies to Question 6, Question 7 (second method), Question 8, and Question 10 and 10.5.
- o1-preview shows a spark of creative reasoning ability, which involves making insightful observations, and sometimes constructing small examples to aid observation-making, as in Question 3 and Question 5. However, this ability seems to be in its early stages and is incomplete.
- Some of the mistakes made by the o1-preview model are remarkably human-like. Unlike earlier models, such as GPT-4o, whose incorrect responses were often vague, repetitive, and filled with irrelevant details, the errors made by the new model are more relatable and understandable. In fact, the types of mistakes resemble those a college student might make, particularly in Question 4, Question 7 (first method), and Question 9. For comparison purposes, the response to Question 7 generated by GPT-4o is illustrated in Figure 38. Note the repetition as highlighted in yellow, and two occurrences vague reasoning as highlighted in red.

For the limitations, our empirical observations indicate that:

- While o1-preview demonstrates the ability to make insightful observations and identify patterns from examples it constructs, it can also be misled by its own observations, leading to inappropriate generalizations—mistakenly assuming that coincidences apply more broadly. This applies to Question 3 and Question 5. In general, there is no evidence to support that the model possesses a genuine understanding of logical principles. When faced with extended reasoning chains, where it may even fall into circular reasoning.
- It seems the o1-preview model lacks an awareness of problem difficulty and often prioritizes efficiency over appropriateness. For instance, in Question 7, it applied overly advanced techniques to solve a moderately challenging problem.
- Overall, the model tends to perform better on problems with a high level of abstraction that require advanced theorems but involve shorter reasoning chains, such as the last five problems, rather than on problems with straightforward statements that require less advanced knowledge but involve much longer reasoning processes, such as the three advanced discrete math problems.

|   |   |
|---|---|
| <p><b>Prompt</b></p> <p>Consider the set of integers <math>\{1, 2, \dots, n\}</math>. How many subsets of this set contain exactly <math>k</math> integers but do not contain consecutive integers? Please deduce an explicit formula step by step.</p> <p><b>O1-preview</b></p> <p>To find the number of (<math>k</math>)-element subsets of the set (<math>\{1, 2, \dots, n\}</math>) that contain no consecutive integers, we'll derive an explicit formula step by step.</p> <p><b>Step 1: Understanding the Problem</b></p> <p>We need to choose (<math>k</math>) integers from (<math>\{1, 2, \dots, n\}</math>) such that no two selected integers are consecutive. This means that between any two chosen numbers, there must be at least one number that is not chosen.</p> <p><b>Step 2: Mapping the Problem</b></p> <p>We can model this problem by considering positions where we can place the chosen integers:</p> <ul style="list-style-type: none"> <li>- Imagine (<math>n</math>) positions corresponding to the integers (1) through (<math>n</math>).</li> <li>- We need to select (<math>k</math>) positions such that no two are adjacent.</li> </ul> <p>To ensure no two selected positions are adjacent, we can think of "gaps" that must exist between the chosen positions.</p> <p><b>Step 3: Adjusting the Number of Available Positions</b></p> <p>Because we need at least one unchosen position between any two chosen positions, the effective number of available positions is reduced. Specifically:</p> <ul style="list-style-type: none"> <li>- The minimum length required to place (<math>k</math>) non-consecutive elements is (<math>2k - 1</math>).</li> </ul> | <ul style="list-style-type: none"> <li>- The total number of positions we have is (<math>n</math>). This means we need to select (<math>k</math>) positions out of (<math>n - (k - 1)</math>) available positions.</li> </ul> <p><b>Step 4: Formulating the Problem Mathematically</b></p> <p>Let's define a mapping to simplify the problem:</p> <ul style="list-style-type: none"> <li>- Let (<math>d_j = i_j - (j - 1)</math>), where (<math>i_j</math>) is the position of the (<math>j</math>)-th chosen integer.</li> <li>- The (<math>d_j</math>) values are strictly increasing and satisfy (<math>1 \leq d_1 &lt; d_2 &lt; \dots &lt; d_k \leq n - k + 1</math>).</li> </ul> <p>This mapping transforms the problem into selecting (<math>k</math>) distinct integers from (<math>\{1, 2, \dots, n - k + 1\}</math>).</p> <p><b>Step 5: Calculating the Number of Ways</b></p> <p>The number of ways to choose (<math>k</math>) non-consecutive integers is thus the number of (<math>k</math>)-element subsets of (<math>\{1, 2, \dots, n - k + 1\}</math>), which is given by the combination formula:</p> $[\text{Number of ways} = \binom{n - k + 1}{k}]$ <p><b>Step 6: Final Answer</b></p> <p>Therefore, the explicit formula for the number of (<math>k</math>)-element subsets of (<math>\{1, 2, \dots, n\}</math>) with no consecutive integers is:</p> $[\boxed{\binom{n - k + 1}{k}}]$ <p>An explicit formula: number of subsets = <math>C(n - k + 1, k)</math>; that is, <math>\boxed{\binom{n-k+1}{k}}</math></p> |
|---|---|

Figure 35: **College-Level Math: Case 1.** A demonstration case of solving a basic discrete math problem. The generated answer is manually evaluated as correct.

|  |  |
|--|--|
| <p><b>Prompt</b></p> <p>Let \$a, b\$ be positive integers. Show that whenever <math>\frac{a^2+b^2}{ab+1}</math> is an integer, it must be a perfect square.</p> <p><b>O1-preview</b></p> <p>To show that whenever <math>\frac{a^2+b^2}{ab+1}</math> is an integer, it must be a perfect square, we can proceed by exploring the relationship between (a), (b), and <math>(n = \frac{a^2+b^2}{ab+1})</math>.</p> <p><b>Step 1: Set Up the Equation</b></p> <p>First, let's denote:</p> $[ n = \frac{a^2+b^2}{ab+1} ]$ <p>Assuming (n) is an integer, we can rewrite the equation:</p> $[ a^2 + b^2 = n(ab + 1) ]$ <p>This simplifies to:</p> $[ a^2 - n ab + b^2 - n = 0 ]$ <p><b>Step 2: Consider (a) as a Function of (b)</b></p> <p>Treat this as a quadratic in (a):</p> $[ a^2 - (n b)a + (b^2 - n) = 0 ]$ <p>For (a) to be an integer, the discriminant (D) of this quadratic equation must be a perfect square:</p> $[ D = [-n b]^2 - 4 \cdot 1 \cdot (b^2 - n) = n^2 b^2 - 4(b^2 - n) ]$ <p>Simplify the discriminant:</p> $[ D = n^2 b^2 - 4b^2 + 4n = b^2(n^2 - 4) + 4n ]$ <p><b>Step 3: Set the Discriminant to Be a Perfect Square</b></p> <p>Let (D) be a perfect square (<math>k^2</math>):</p> $[ D = k^2 = b^2(n^2 - 4) + 4n ]$ | <p><b>Step 4: Solve for (n) and (b)</b></p> <p>Rewriting:</p> $[ k^2 = b^2(n^2 - 4) + 4n ]$ <p>We can attempt to find integer values of (n) and (b) such that (D) is a perfect square. Through testing small integer values, we observe a pattern:</p> <ul style="list-style-type: none"> <li>- When <math>(n = b^2)</math>, the discriminant (D) becomes:</li> </ul> $[ D = b^2(b^4 - 4) + 4b^2 = b^6 - 4b^2 + 4b^2 = b^6 ]$ <p>So <math>(D = b^6)</math>, which is a perfect square.</p> <p><b>Step 5: Show That <math>(n = b^2)</math> and <math>(a = b^3)</math></b></p> <p>Assuming <math>(n = b^2)</math>, the quadratic equation becomes:</p> $[ a^2 - b^3 a + b^2 - b^2 = 0 \implies a^2 - b^3 a = 0 ]$ <p>This simplifies to:</p> $[ a(a - b^3) = 0 ]$ <p>Since <math>(a &gt; 0)</math>, we have <math>(a = b^3)</math>.</p> <p><b>Step 6: Verify That <math>(n)</math> Is a Perfect Square</b></p> <p>Substituting <math>(a = b^3)</math> and <math>(n = b^2)</math> back into the original equation:</p> $[ n = \frac{(b^3)^2 + b^2}{b^3 \cdot b + 1} = \frac{b^6 + b^2}{b^4 + 1} ]$ <p>Simplify the numerator and denominator:</p> $[ n = \frac{b^2(b^4 + 1)}{b^4 + 1} = b^2 ]$ <p>Thus, <math>(n = b^2)</math>, which is a perfect square.</p> <p><b>Conclusion:</b></p> <p>Whenever <math>\frac{a^2+b^2}{ab+1}</math> is an integer, it must be of the form <math>(n = b^2)</math>, where (n) is a perfect square. Therefore, (n) is always a perfect square under the given conditions.</p> <p><b>Answer:</b> It must be a perfect square—such n is always a square number whenever the expression is integral.</p> |
|--|--|

Figure 36: College-Level Math: Case 5. A demonstration case of solving a advanced discrete math problem. The generated answer is manually evaluated as incorrect.

|   |   |
|---|---|
| <p><b>Prompt</b></p> <p>Does the series <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> converge or diverge? Prove with two different methods.</p> <p><b>O1-preview</b></p> <p>To determine whether the series <math>\sum_{n=1}^{\infty} \frac{ \sin n }{n}</math> converges or diverges, we'll employ two different methods: comparison with a divergent series over an infinite subsequence and estimating the average value of <math> \sin n </math>.</p> <p><b>Method 1: Comparison with the Harmonic Series over an Infinite Subsequence</b></p> <p><b>Step 1: Identify an Infinite Subsequence</b></p> <p>We observe that <math> \sin n </math> attains values arbitrarily close to 1 infinitely often because <math>(\sin n)</math> is dense in the interval <math>([-1, 1])</math>. In particular, there are infinitely many integers <math>(n)</math> such that:</p> $ \sin n  \geq \frac{1}{2}$ <p><b>Step 2: Establish a Lower Bound</b></p> <p>For these integers, we have:</p> $\frac{ \sin n }{n} \geq \frac{1}{2n}$ <p><b>Step 3: Compare with the Harmonic Series</b></p> <p>The series <math>\sum_{n=1}^{\infty} \frac{1}{n}</math> is the harmonic series, which diverges. Therefore, the sum over our infinite subsequence diverges:</p> $\sum_{\substack{n=1 \\  \sin n  \geq \frac{1}{2}}}^{\infty} \frac{1}{n} \geq \sum_{\substack{n=1 \\  \sin n  \geq \frac{1}{2}}}^{\infty} \frac{1}{2n}$ <p>Since this sum diverges, the original series diverges by the Comparison Test.</p> | <p><b>Method 2: Estimating the Average Value of <math> \sin n </math></b></p> <p><b>Step 1: Recognize the Equidistribution</b></p> <p>The sequence <math>(n \bmod 2\pi)</math> is equidistributed in <math>([0, 2\pi))</math> because <math>(2\pi)</math> is irrational with respect to 1. This implies that <math>(\sin n)</math> takes on values in <math>([-1, 1])</math> in a way that averages out over the interval.</p> <p><b>Step 2: Compute the Expected Value</b></p> <p>The average value of <math> \sin x </math> over the interval <math>([0, 2\pi))</math> is:</p> $\left[ \frac{1}{2\pi} \int_0^{2\pi}  \sin x  dx = \frac{2}{\pi} \right]$ <p><b>Step 3: Estimate the Partial Sums</b></p> <p>Using this average, we can estimate the partial sums:</p> $\left[ S_N = \sum_{n=1}^N \frac{ \sin n }{n} \approx \frac{2}{\pi} \sum_{n=1}^N \frac{1}{n} \right]$ <p><b>Step 4: Recognize the Divergence</b></p> <p>Since <math>\sum_{n=1}^{\infty} \frac{1}{n}</math> diverges (the harmonic series), it follows that:</p> $\left[ \lim_{N \rightarrow \infty} S_N \approx \frac{2}{\pi} (\ln N + \gamma) \rightarrow \infty \right]$ <p>where <math>(\gamma)</math> is the Euler-Mascheroni constant. Therefore, the original series diverges.</p> <p><b>Conclusion:</b></p> <p>By both methods, we conclude that the series diverges.</p> <p><b>Answer: It diverges; two proofs below show that the series diverges by comparison with the harmonic series and by estimating its divergent partial sums.</b></p> |
|---|---|

Figure 37: College-Level Math: Case 7. A demonstration case of solving a Calculus problem. The generated answer is manually evaluated as questionable.

|   |   |
|---|---|
| <p><b>Prompt</b></p> <p>Does the series <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> converge or diverge? Prove with two different methods.</p> <p><b>GPT-4o</b></p> <p>We are tasked with determining whether the series <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> converges or diverges. We'll approach this problem using two different methods.</p> <p><b>Method 1: Comparison Test</b></p> <p>To use the Comparison Test, let's first analyze the behavior of <math> \sin(n) </math>. The function <math> \sin(n) </math> oscillates between (-1) and (1), so <math> \sin(n) </math> is bounded between (0) and (1). In particular:</p> $0 \leq  \sin(n)  \leq 1 \quad \text{for all } n.$ <p>This suggests we compare <math>\frac{ \sin(n) }{n}</math> with the simpler series <math>\sum_{n=1}^{\infty} \frac{1}{n}</math>, which is a harmonic series. Recall that the harmonic series diverges:</p> $\sum_{n=1}^{\infty} \frac{1}{n} = \infty.$ <p>Since <math>0 \leq \frac{ \sin(n) }{n} \leq \frac{1}{n}</math> for all <math>n</math>, and because the harmonic series <math>\sum_{n=1}^{\infty} \frac{1}{n}</math> diverges, by the Comparison Test, the series <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> also diverges.</p> | <p><b>Method 2: Limit Comparison Test</b></p> <p>Now let's apply the Limit Comparison Test. The idea is to compare <math>\frac{ \sin(n) }{n}</math> to <math>\frac{1}{n}</math>, but instead of a simple comparison, we use the following limit:</p> $\lim_{n \rightarrow \infty} \frac{ \sin(n) }{n} = \lim_{n \rightarrow \infty}  \sin(n) .$ <p>Since <math> \sin(n) </math> oscillates between 0 and 1, the limit does not exist in a strict sense, but we can still gain insight from the fact that <math> \sin(n) </math> does not approach 0. In fact, for infinitely many <math>n</math>, <math> \sin(n) </math> is bounded away from zero. This implies that the terms <math>\frac{ \sin(n) }{n}</math> behave similarly to <math>\frac{1}{n}</math> for large <math>n</math>.</p> <p>Because <math>\sum_{n=1}^{\infty} \frac{1}{n}</math> diverges and <math>\frac{ \sin(n) }{n}</math> behaves similarly to <math>\frac{1}{n}</math>, the Limit Comparison Test tells us that <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> also diverges.</p> <p><b>Conclusion</b></p> <p>Both methods show that the series <math>\sum_{n=1}^{\infty} \frac{ \sin(n) }{n}</math> diverges.</p> |
|---|---|

Figure 38: College-Level Math: Case 7 (GPT-4o). Answer to the same Calculus problem as in Figure 37 generated by GPT-4o.

## 4.16 Electronic Health Record Diagnosis

To evaluate the diagnostic capabilities of the o1-preview model using EHRs, we tested it with 10 manually selected cases, each representing a different cardiovascular disease category. The model's responses were manually reviewed by the authors and categorized as either correct (marked in green) or incorrect (marked in red). In addition to identifying key medical patterns, the o1-preview model also provided diagnostic explanations and reasoning based on the information provided, demonstrating strong inferential capabilities.

o1-preview has demonstrated notable potential in supporting medical diagnosis through its ability to interpret EHR data and provide relevant diagnoses. While the model performs exceptionally well with shorter, more concise texts (Figure 40, Figure 41, Figure 42), challenges persist when dealing with longer and more complex records, which may impact its accuracy (Figure 39). Nevertheless, the model's ability to offer reasoning and diagnostic justifications indicates that with future fine-tuning, particularly through reinforcement learning tailored to EHR data, its diagnostic capabilities could be significantly enhanced. Additional examples can be found in (Appendix A.11).

Overall, the o1-preview's capabilities to understand and analyze EHRs show great potential for the development of AI-driven healthcare tools, though further refinement is needed to improve its effectiveness in real-world medical applications.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Haemophilus influenzae meningitis with prolonged hospital course.

A retrospective evaluation of Haemophilus influenzae type b meningitis observed over a 2-year period documented 86 cases.

Eight of these patients demonstrated an unusual clinical course characterized by persistent fever (duration: greater than 10 days), cerebrospinal fluid pleocytosis, profound meningeal enhancement on computed tomography, significant morbidity, and a prolonged hospital course.

The mean age of these 8 patients was 6 months, in contrast to a mean age of 14 months for the entire group. Two patients had clinical evidence of relapse.

Four of the 8 patients tested for latex particle agglutination in the cerebrospinal fluid remained positive after 10 days.

All patients received antimicrobial therapy until they were afebrile for a minimum of 5 days.

Subsequent neurologic examination revealed a persistent seizure disorder in 5 patients (62.5%), moderate-to-profound hearing loss in 2 (25%), mild ataxia in 1 (12.5%), and developmental delay with hydrocephalus which required shunting in 1 (12.5%).

One patient had no sequelae.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Nervous System Diseases**.

Explanation:

Haemophilus influenzae Meningitis:

The abstract discusses 86 cases of Haemophilus influenzae type b meningitis observed over a two-year period.

Meningitis is an inflammation of the meninges, the protective membranes covering the brain and spinal cord.

The infection leads to neurological symptoms such as persistent fever, cerebrospinal fluid pleocytosis (increased white blood cells in the cerebrospinal fluid), profound meningeal enhancement on computed tomography, seizures, hearing loss, ataxia, developmental delays, and hydrocephalus requiring shunting.

Relevance to the Category:

While Haemophilus influenzae is a bacterium, the primary system affected by this infection is the central nervous system.

The significant neurological complications and sequelae emphasize the impact on the nervous system.

Therefore, the condition is best classified under Nervous System Diseases rather than solely as a bacterial infection.

Conclusion:

Given that the patient's primary condition is meningitis affecting the central nervous system, the most appropriate category is **Nervous System Diseases**.

### **Reference Answer**

**Bacterial Infections and Mycoses**

Figure 39: **Electronic Health Record Diagnosis: Case 1.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Haemophilus influenzae meningitis with prolonged hospital course.

A retrospective evaluation of Haemophilus influenzae type b meningitis observed over a 2-year period documented 86 cases.

Eight of these patients demonstrated an unusual clinical course characterized by persistent fever (duration: greater than 10 days), cerebrospinal fluid pleocytosis, profound meningeal enhancement on computed tomography, significant morbidity, and a prolonged hospital course.

The mean age of these 8 patients was 6 months, in contrast to a mean age of 14 months for the entire group. Two patients had clinical evidence of relapse.

Four of the 8 patients tested for latex particle agglutination in the cerebrospinal fluid remained positive after 10 days.

All patients received antimicrobial therapy until they were afebrile for a minimum of 5 days.

Subsequent neurologic examination revealed a persistent seizure disorder in 5 patients (62.5%), moderate-to-profound hearing loss in 2 (25%), mild ataxia in 1 (12.5%), and developmental delay with hydrocephalus which required shunting in 1 (12.5%).

One patient had no sequelae.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient category in question is associated with **Cardiovascular Diseases**.

Explanation:

The abstract discusses causes of death in an emergency department, focusing on nontrauma deaths.

The most common nontrauma causes of death are:

Sudden death of uncertain cause (34%)

Coronary artery disease (34%)

Other heart disease (4%)

These causes are directly related to the cardiovascular system. Specifically:

Coronary artery disease is a well-known cardiovascular condition.

Other heart diseases also fall under cardiovascular disorders.

Sudden death of uncertain cause often involves cardiac arrhythmias or undiagnosed heart conditions, which are cardiovascular in nature.

Given the options listed, **Cardiovascular Diseases** is the most appropriate category for this patient group.

### **Reference Answer**

**Cardiovascular Diseases**

Figure 40: **Electronic Health Record Diagnosis: Case 2.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Disseminated strongyloidiasis in AIDS and non-AIDS immunocompromised hosts: diagnosis by sputum and bronchoalveolar lavage.

In conclusion, disseminated strongyloidiasis is a fatal disease that commonly affects the lungs.

The disease should be suspected in an immunocompromised host who came from an area endemic for *S. stercoralis* even years before the onset of symptoms or in patients with unexplained gram-negative bacteremia or meningitis.

Treatment should be started promptly and should be maintained for a long time.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Parasitic Diseases**.

Explanation:

Disseminated Strongyloidiasis:

The abstract discusses disseminated strongyloidiasis in immunocompromised hosts, including those with AIDS. Strongyloidiasis is an infection caused by the parasitic nematode *Strongyloides stercoralis*.

This parasite can cause severe and potentially fatal infections, especially when the immune system is compromised.

Clinical Features:

The disease commonly affects the lungs, as mentioned in the abstract.

Symptoms may include unexplained gram-negative bacteremia or meningitis.

It is important to suspect this disease in patients from endemic areas, even if exposure occurred years prior.

Relevance to the Category:

Since *Strongyloides stercoralis* is a parasite, infections caused by it are classified under Parasitic Diseases. The condition involves systemic dissemination of the parasite, affecting multiple organ systems.

Conclusion:

Given that the patient's condition is a parasitic infection caused by *Strongyloides stercoralis*, the appropriate category is **Parasitic Diseases**.

### **Reference Answer**

**Parasitic Diseases.**

Figure 41: **Electronic Health Record Diagnosis: Case 3.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Cerebellopontine angle lipoma in a teenager.

Lipomas of the cerebellopontine angle are very rare lesions.

To date, 18 patients have been reported, 17 of whom were adults.

A second child is described with cerebellopontine angle lipoma.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Neoplasms**.

Explanation:

Cerebellopontine Angle Lipoma:

The abstract discusses a lipoma located in the cerebellopontine angle, which is a region at the junction of the cerebellum and the pons in the brain.

A lipoma is a benign tumor composed of adipose (fat) tissue.

Neoplasms:

Neoplasms are abnormal growths of tissue, which can be benign or malignant (cancerous).

Lipomas are classified as benign neoplasms because they are non-cancerous tumors arising from fat cells.

Relevance to the Category:

While the lipoma is located in the nervous system, its classification is based on its nature as a tumor.

Therefore, it fits best under Neoplasms rather than Nervous System Diseases, which typically refer to functional disorders of the nervous system.

Conclusion:

Given that the patient has a benign tumor (lipoma) in the cerebellopontine angle, the most appropriate category is **Neoplasms**.

### **Reference Answer**

**Neoplasms.**

Figure 42: **Electronic Health Record Diagnosis: Case 4.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

---

## 4.17 Stochastic Processes in Statistics

The performance of the o1-preview on stochastic process problems is summarized as follows. Overall, the model exhibited a strong ability to not only understand the majority of stochastic process problems but also to provide step-by-step, detailed solutions. In many cases, the explanations were thorough, though at times overly detailed, which occasionally led to somewhat redundant or verbose answers. This tendency, while not detrimental, suggests that the model's solution clarity could benefit from a more concise presentation in certain situations.

Despite its success with most problems, the o1-preview encountered difficulties with more complex Markov chain problems, particularly in accurately determining the steady state of the Markov chain, which resulted in errors. This indicates that while the model performs well on simpler tasks, there is room for improvement in its deep understanding and handling of more intricate, advanced topics.

As shown in Appendix A.12, the model achieved an 80% accuracy rate on basic problems. However, as evidenced in Figure 44, it struggled with more challenging cases. Additionally, the model was evaluated on some classic mathematical problems that large language models typically find challenging, such as comparing 9.11 and 9.8. In these instances, the o1-preview performed exceptionally well, as shown in Figure 137.

## Prompt

Here are some questions about stochastic process, please answer step by step.

A certain calculating machine uses only the digits 0 and 1. It is supposed to transmit one of these digits through several stages. However, at every stage, there is a probability  $p$  that the digit that enters this stage will be changed when it leaves and a probability  $q = 1 - p$  that it won't.

1. **Form a Markov chain** to represent the process of transmission by taking as states the digits 0 and 1. What is the matrix of transition probabilities?
2. **Draw a tree** and assign probabilities assuming that the process begins in state 0 and moves through two stages of transmission. What is the probability that the machine, after two stages, produces the digit 0 (i.e., the correct digit)?

## O1-preview

Answer to Problem 1

### Transition Probability Matrix:

Let  $P$  be the transition matrix, where  $P_{ij}$  represents the probability of transitioning from state  $i$  to state  $j$ :

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \begin{bmatrix} \text{Probability from 0 to 0} & \text{Probability from 0 to 1} \\ \text{Probability from 1 to 0} & \text{Probability from 1 to 1} \end{bmatrix} = \begin{bmatrix} q & p \\ p & q \end{bmatrix}$$

Answer to Problem 2

### Tree Diagram:

Here's the conceptual tree diagram with assigned probabilities:

Start at State 0

First Stage:

- Remain in State 0 (probability  $q$ )
- Second Stage:
  - Remain in State 0 (probability  $q$ )
  - Transition to State 1 (probability  $p$ )
- Transition to State 1 (probability  $p$ )
- Second Stage:
  - Transition to State 0 (probability  $p$ )
  - Remain in State 1 (probability  $q$ )

$$\text{Total Probability} = q^2 + p^2$$

## Reference Answer

$$P = \begin{bmatrix} q & p \\ p & q \end{bmatrix}$$

$$q^2 + p^2$$

Figure 43: Stochastic Processes in Statistics: Case 1. A demonstration case of Stochastic Process problem solving. The o1-preview successfully solved the problem.

## Prompt

Here are some questions about stochastic process, please answer step by step.

Assume that a man's profession can be classified as **professional**, **skilled labourer**, or **unskilled labourer**. Assume that, of the sons of professional men, 80 percent are professional, 10 percent are skilled labourers, and 10 percent are unskilled labourers. In the case of sons of skilled labourers, 60 percent are skilled labourers, 20 percent are professional, and 20 percent are unskilled. Finally, in the case of unskilled labourers, 50 percent of the sons are unskilled labourers, and 25 percent each are in the other two categories.

Assume that every man has at least one son, and form a Markov chain by following the profession of a randomly chosen son of a given family through several generations.

1. **Set up the matrix of transition probabilities.**
2. **Find the probability** that a randomly chosen grandson of an unskilled labourer is a professional man.

## O1-preview

To solve this problem, we'll proceed step by step.

### 1. Set up the matrix of transition probabilities.

Transition Matrix  $P$ :

$$P = \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.20 & 0.60 & 0.20 \\ 0.25 & 0.25 & 0.50 \end{pmatrix}$$

### 2. Find the probability that a randomly chosen grandson of an unskilled labourer is a professional man.

$$P_{3,1}^{(2)} = \sum_{k=1}^3 P_{3,k} \times P_{k,1}$$

Answer:

The probability that a randomly chosen grandson of an unskilled labourer is a professional man is **0.375** or **37.5%**.

## Reference Answer

$$P = \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.20 & 0.60 & 0.20 \\ 0.25 & 0.25 & 0.50 \end{pmatrix}$$

**0.375**

Figure 44: Stochastic Processes in Statistics: Case 2. A demonstration case of Stochastic Process problem solving. The o1-preview failed to solve the problem.

## 4.18 Medical Text Anonymization

To assess o1-preview’s effectiveness in medical text anonymization, we crafted three customized prompts with varying degrees of specificity. These prompts were structured from broad to detailed, allowing us to gauge the model’s performance across different levels of granularity. The first, the coarse prompt, provides minimal guidance, simply directing the model to identify private information without offering further specifics. The medium prompt introduces a more comprehensive list of privacy-related elements, while the fine prompt supplies detailed instructions, specifying the properties of each privacy feature. Table 6 outlines this experiment’s complete set of prompts.

Table 6: Prompt Table for Medical Text Anonymization.

| Implicit Prompt   | General Prompt  | Explicit Prompt   |
|---|---|---|
| Please anonymize the following text. Replace all the private information with the term “[redacted]”.: | What are the names? Who are the professions? What are the locations? What are ages? What are dates? What are contacts? What are IDs? What are the phone numbers? Replace the matching characters with [redacted]: | Please anonymize the following clinical note. Specifically, replace all the following information with the term “[redacted]”: redact any strings that might be a name or acronym or initials, redact any strings separated by the \ / symbol, redact patients’ names, doctors’ names and the strings in front of M.D. or after Dr., redact pager names and medical staff names, redact any strings that look like something years old or age 37, redact any dates and IDs and numbers and record dates, redact locations and addresses and clinic names, redact professions and ages and contacts, redact any acronyms and initials.: |

As the results (Figures 45, 46, 47) present, o1-preview is capable of detecting nearly all sensitive information, even when processing lengthy content. It can efficiently handle long texts, analyze the material from multiple perspectives, and generate results in a short amount of time. Although different words are redacted across the three prompts, each effectively removed essential privacy information, demonstrating the impressive general capabilities of o1-preview. The results have proved its potential to be a powerful automatic tool to anonymize professional medical texts.

o1-preview accurately detects personal names and timestamps, but slight variations can be observed when using different prompts. With implicit prompts, o1-preview focuses primarily on the overall structure of the content, identifying sensitive information in fixed positions while preserving abbreviations that aid in localization. For general prompts, due to the strong alignment between the annotation features and the prompts, o1-preview detects all sensitive information without any additional effort. However, with explicit prompts, the abundance of instructions causes the o1-preview to evaluate each word individually. While it successfully identifies all sensitive information, it occasionally removes more meaningful words than expected. In summary, the effectiveness of o1-preview largely depends on the appropriateness of the prompt.

### **Prompt**

Please anonymize the following text. Replace all the private information with the term “[redacted]”:

### **O1-preview**

[redacted] [redacted] is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis—possibly RA—with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking 1 aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow up with me shortly after the ultrasound study. [redacted] [redacted] DD: [redacted] DT: [redacted] DV: [redacted]

### **Reference Answer**

Record date: 2069-04-07 Mr. Villegas is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis - possibly RA - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking 1 aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow-up with me shortly after the ultrasound study. Xzavian G. Tavares, M.D. XGT:holmes DD: 04/07/69 DT: 04/15/69 DV: 04/07/69

Figure 45: **Medical Text Anonymization: Case 1.** An example showing the capability of o1-preview to anonymize medical texts with the implicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### Prompt

What are the names? Who are the professions? What are the locations? What are ages? What are dates? What are contacts? What are IDs? What are the phone numbers? Replace the matching characters with [redacted]:

### O1-preview

Record date: [redacted] [redacted] is seen today. I have not seen him since [redacted]. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis—possibly RA—with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking 1 aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow up with me shortly after the ultrasound study. [redacted] [redacted] DD: [redacted] DT: [redacted] DV: [redacted]

### Reference Answer

Record date: 2069-04-07 Mr. Villegas is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis - possibly RA - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking 1 aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow-up with me shortly after the ultrasound study. Xzavian G. Tavares, M.D. XGT.holmes DD: 04/07/69 DT: 04/15/69 DV: 04/07/69

Figure 46: **Medical Text Anonymization: Case 2.** An example showing the capability of o1-preview to anonymize medical texts with the general prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates additional modifications beyond the human-marked annotation.

### **Prompt**

Please anonymize the following clinical note. Specifically, replace all the following information with the term "[redacted]": redact any strings that might be a name or acronym or initial, redact any strings separated by the "V" symbol, redact patients' names, doctors' names and the strings in front of M.D. or after Dr., redact pager names and medical staff names, redact any strings that look like something years old or age 37, redact any dates and IDs and numbers and record dates, redact locations and addresses and clinic names, redact professions and ages and contacts, redact any acronyms and initials.:

### **O1-preview**

[redacted] [redacted] is seen today. I have not seen him since [redacted]. About [redacted] he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and [redacted] Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, [redacted] [redacted]. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better [redacted] of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis - possibly [redacted] - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking [redacted] aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow-up with me shortly after the ultrasound study. [redacted], M.D. [redacted] [redacted]

### **Reference Answer**

Record date: 2069-04-07 Mr. Villegas is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the carotid. He has no tenderness in this region either. He has good peripheral pulses at the arms. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression: #1: Inflammatory arthritis - possibly RA - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. #2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. He has been taking 1 aspirin per day prophylaxis long-term, and I stressed that he continue to do so. He will follow-up with me shortly after the ultrasound study. Xzavian G. Tavares, M.D. XGT;holmes DD: 04/07/69 DT: 04/15/69 DV: 04/07/69

Figure 47: **Medical Text Anonymization: Case 3.** An example showing the capability of o1-preview to anonymize medical texts with the explicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

## 4.19 Social Media Analysis

To evaluate the capabilities of o1-preview in social media analysis, we used four common tweet datasets and reported results across 11 cases. The prompt templates used throughout the testing process are presented in Table 7. Among these 11 cases, three cases are sentiment analysis (Figure 48), two cases are offensive language detection (Figure 49), four are emotion recognition (Figure 144), and two are irony identification (Figure 145).

Table 7: Prompt table for social media analysis.

| Sentiment analysis  | Emotion detection   | Offensiveness  | Irony detection  |
|---|---|--|--|
| Is the sentiment of this text positive, neutral, or negative? Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative. Here is the text: [Tweet text] | Which of these [number of] emotions—[list of emotions]—best represents the mental state of the person writing the following (Indonesian) text? Answer only with a number: 1 if [emotion1], 2 if [emotion2], [...]. Here is the text: [Tweet text] | Is the following post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post: [Tweet text] | Is the following post ironic? Please answer with numbers only: 1 for irony and 0 for non-irony. Here is the post: [Tweet text] |

During the reasoning tests, we found that o1-preview performed very well in most cases, with its predictions aligning closely with the ground truth. However, there were some instances where it produced incorrect results. For example, such errors were observed in both sentiment analysis and emotion recognition tasks. In the case of sentiment analysis, o1-preview mistakenly judged a tweet with neutral sentiment as containing negative sentiment. When performing sentiment analysis on this example, o1-preview output was: *I'm mapping out the sentiment expressed in the text, where it condemns a massacre and bombings, and advocates action for Aleppo, indicating both negative and positive sentiments.* Although o1-preview ultimately gave the wrong answer, it was actually unsure during its reasoning whether the text had more positive or negative sentiments. Despite this uncertainty, it did not choose the neutral option that lies between positive and negative sentiments. Similarly, this phenomenon was also observed in another error case on the emotion recognition task.

Although our test results are not comprehensive due to OpenAI's policy restrictions, we have roughly summarized some reasons that may cause o1-preview to analyze incorrect results in some cases. First, these tasks require deep semantic understanding and the ability to interpret subtle emotional cues, such as sarcasm or tone, which models may struggle to capture accurately. Emotions often rely on implicit context, cultural background, or even non-verbal cues, which are difficult for text-based models to grasp. This suggests that while large language models excel in many areas, they still face significant limitations when it comes to understanding and interpreting human emotions. Overall, the performance results of o1-preview on social media analysis tasks are positive. It captures complex expressions such as irony, sarcasm, and humor on most tasks.

**Prompt**

*Is the sentiment of this text positive, neutral, or negative?  
Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative.  
Here is the text:*

@user we all know on JANUARY 20th trump gonna push oil, fossil fuels, fracking down our throats. Fml... We must fight back. GOP sucks.

**O1-preview**

3

**Reference Answer**

3 (negative)

**Prompt**

*Is the sentiment of this text positive, neutral, or negative?  
Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative.  
Here is the text:*

Rally in #Toronto condemns the massacre in #Aleppo & urges #Canada to #StandwithAleppo & act to stop the bombings

**O1-preview**

3

**Reference Answer**

2 (neutral)

**Prompt**

*Is the sentiment of this text positive, neutral, or negative?  
Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative.  
Here is the text:*

Omg I love Jill Scott why am I just now watching #ThankYouObama 🙏

**O1-preview**

1

**Reference Answer**

1 (positive)

Figure 48: **Social Media Analysis: Case 1-3.** Demonstration cases for sentiment analysis. Red indicates that the output of the o1-preview model is completely inconsistent with the reference answer, while green signifies full consistency.

**Prompt**

*Is the following post ironic? Please answer with numbers only: 1 for irony and 0 for non-irony. Here is the post:*

@user #Not #Long To Go #Before #MainStream #Media #Pitch. #Surgery #follows #with #me #using #Prototype #demonstrating #real #situation.Phil

**O1-preview**

0

**Reference Answer**

0 (non-irony)

**Prompt**

*Is the following post ironic? Please answer with numbers only: 1 for irony and 0 for non-irony. Here is the post:*

30 years on, people still haven't got justice they deserve. #BhopalGasTragedy Yet another brilliant job done by the Americans #sarcasm

**O1-preview**

1

**Reference Answer**

1 (irony)

Figure 49: **Social Media Analysis: Case 4-5.** Demonstration cases for offensive language identification.

---

## 4.20 Analogical Reasoning

To evaluate the analogical reasoning ability of o1-preview, we designed a set of experiments focused on semantic structure mapping, where the model needs to transfer semantic relations from one domain to another. These tasks were created to test various aspects of analogical reasoning. The complexity of these tasks varies, from basic relational analogies to more intricate challenges involving multiple layers of semantic content.

The experimental tasks were divided into two main categories:

1. **Mapping of Semantic Structure:** These tasks involve identifying and transferring relational structures between semantically meaningful symbols. The model must align abstract relationships such as those between "dog" and "puppy" with non-linguistic representations.
2. **Mapping of Semantic Content:** The model needed to transfer specific content-related attributes from one domain to another. For example, it must infer that the "dog" symbol typically has four legs, converting it into a non-semantic representation.

The prompt template used in this experiment was: "We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing. Question: [Question]".

o1-preview shows varying levels of success across different tasks. It demonstrates a strong ability to analyze and align underlying semantic structures in simple relational analogies. As the complexity of the tasks increases, especially those involving multiple layers of semantic content and structure, the model's performance remains strong, correctly answering 10 test questions.

Figure 50 and 51 illustrate this performance. Figure 50 shows examples of direct semantic structure mapping, both with and without distractors, while Figure 51 shows examples of complex analogies. In these tasks, o1-preview effectively eliminates distractions and extracts analogical features through reasoning analysis to obtain correct results.

The observations from these experiments indicate that o1-preview has robust analogical reasoning capabilities. It can derive correct reasoning outcomes in both direct semantic structure mapping and more complex analogies, demonstrating an ability to effectively identify and apply learned relational patterns.

**Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

square => E K C  
rectangle => E % K % C  
circle => C K E  
oval =>

**O1-preview**

C % K % E

**Reference Answer**

C % K % E

**Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

lime => U  
king => C C C  
queen => c c c  
man => C C  
woman =>

**O1-preview**

c c

**Reference Answer**

c c

**Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

puppy => Q Q Z Z I I  
kitten => Q Q Z Z I I Q Q Z Z I I  
cat => Q Z I Q Z I  
dog =>

**O1-preview**

Q Z I

**Reference Answer**

Q Z I

Figure 50: **Analogical Reasoning: Case 1-3.** A demonstration case for direct semantic structure mapping. Green denotes that the answers are correct.

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

feet => \*  
inches => \*\*\*\*\*  
quarter => \*\*\*  
dollar => \*  
dime => \*\*\*\*\*  
week =>

### **O1-preview**

\*\*\*\*\*

### **Reference Answer**

\*\*\*\*\*

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

pants => H # H  
glove => X # X  
torso => V  
foot => Z  
head => M  
shirt => V # V  
hat =>

### **O1-preview**

M # M

### **Reference Answer**

M # M

Figure 51: **Analogical Reasoning: Case 4-5.** A demonstration case for multiple layers of semantic content and structure. Green denotes that the answers are correct.

---

## 4.21 Sentiment Analysis

To evaluate the sentiment analysis capabilities of the o1-preview model, we designed a series of experiments focusing on testing the model's ability to classify and understand sentiments across different domains and text types. We employed three widely-used benchmark datasets: IMDB, SemEval-2014 Task 4, and SemEval-2015 Task 12. These datasets cover sentiment analysis tasks ranging from movie reviews to restaurants, electronics, and social media content, providing a comprehensive evaluation of the model's abilities.

The experimental tasks were divided into 2 main categories:

1. **Overall Sentiment Classification:** Using the IMDB dataset, the model was tasked with classifying long-form movie reviews as positive or negative. This dataset contains 50,000 user-generated reviews with diverse language styles, including sarcasm, metaphors, and complex linguistic phenomena.
2. **Aspect-Based Sentiment Analysis:** Utilizing the SemEval-2014 Task 4 dataset, the model needed to identify sentiments toward specific aspects within the text, such as the service quality of a restaurant or the battery life of a laptop. This required the model to not only understand the overall sentiment but also to analyze sentiment tendencies toward specific attributes.

In our experiments, we provided the model with appropriate prompt templates, such as: "You are a sentiment analysis expert. Analyze the sentiment of the following text and provide the sentiment category (Positive, Negative, Neutral). Text: [Text Content]." and "You are a sentiment analysis expert. Please analyze all the emotions expressed in the following text. For each emotion, provide the sentiment category, the relevant text snippet, and a brief explanation. Text:[Text Content]."

The o1-preivew model performed excellently across different tasks. In the overall sentiment classification task, it accurately handled complex and lengthy movie reviews, correctly classifying the majority of test samples. In aspect-based sentiment analysis, the model successfully identified sentiments toward specific aspects within the text and could distinguish different sentiments toward different aspects within the same text.

These results are illustrated in Figures 52, 53, 54, 55, 148, 149. From these experiments, it can be observed that the o1-preview model possesses strong sentiment analysis capabilities, accurately performing sentiment classification and analysis across different domains and text types. It excels in understanding complex linguistic structures and identifying implicit sentiments.

Overall, these experiments demonstrate that the o1-preview model has strong capabilities in sentiment analysis, achieving a high level of sentiment understanding across different domains and text types.

**Prompt**

You are a sentiment analysis expert. Analyze the sentiment of the following text and provide the sentiment category (Positive, Negative, Neutral).

Text:

Once again Mr. Costner has dragged out a movie for far longer than necessary. Aside from the terrific sea rescue sequences, of which there are very few I just did not care about any of the characters. Most of us have ghosts in the closet, and Costner's character are realized early on, and then forgotten until much later, by which time I did not care. The character we should really care about is a very cocky, overconfident Ashton Kutcher. The problem is he comes off as kid who thinks he's better than anyone else around him and shows no signs of a cluttered closet. His only obstacle appears to be winning over Costner. Finally when we are well past the half way point of this stinker, Costner tells us all about Kutcher's ghosts. We are told why Kutcher is driven to be the best with no prior inkling or foreshadowing. No magic here, it was all I could do to keep from turning it off an hour in.

**O1-preview**

Negative

**Reference Answer**

Negative

Figure 52: **Sentiment Analysis: Case 1.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

**Prompt**

You are a sentiment analysis expert. Analyze the sentiment of the following text and provide the sentiment category (Positive, Negative, Neutral).

Text:

I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.

**O1-preview**

Positive

**Reference Answer**

Positive

Figure 53: **Sentiment Analysis: Case 2.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

### **Prompt**

You are a sentiment analysis expert. Please analyze all the emotions expressed in the following text. Its input is an E#A pair and a sentence, requiring us to output the emotional polarity of this E#A pair.

For example:

Judging from previous posts this used to be a good place, but not any longer.

```
{target:"NULL" category:"RESTAURANT#GENERAL" polarity:"negative" from:"-" to="-"}
```

Text:

We, there were four of us, arrived at noon - the place was empty - and the staff acted like we were imposing on them and they were very rude.

### **O1-preview**

```
{target:"staff" category:"STAFF#ATTITUDE" polarity:"negative" from:"-" to="-"}
```

### **Reference Answer**

```
{target:"staff" category:"SERVICE#GENERAL" polarity:"negative" from:"75" to:"80"}
```

Figure 54: **Sentiment Analysis: Case 3.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

### **Prompt**

You are a sentiment analysis expert. Please analyze all the emotions expressed in the following text. Its input is an E#A pair and a sentence, requiring us to output the emotional polarity of this E#A pair.

For example:

Judging from previous posts this used to be a good place, but not any longer.

```
{target:"NULL" category:"RESTAURANT#GENERAL" polarity:"negative" from:"-" to="-"}
```

Text:

The food was lousy - too sweet or too salty and the portions tiny.

### **O1-preview**

```
{target:"food" category:"FOOD#TASTE" polarity:"negative" from:"-" to:"-"}
```

```
{target:"portions" category:"FOOD#SIZE" polarity:"negative" from:"-" to:"-"}
```

### **Reference Answer**

```
{target:"food" category:"FOOD#QUALITY" polarity="negative" from:"4" to:"8"}
```

```
{target:"portions" category:"FOOD#STYLE_OPTIONS" polarity:"negative" from:"52" to:"60"}
```

Figure 55: **Sentiment Analysis: Case 4.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

## 4.22 Anthropology and Geology

Testing has shown that o1-preview is currently capable of answering complex questions in paleoanthropology and geology. It can address detailed aspects of ancient genes in paleoanthropological research (Figure 56) and has a deep understanding of geological issues(Figure 60). As the questions become more specific, o1-preview is able to provide increasingly detailed and in-depth responses, offering researchers in these fields relevant information, current research status, and commonly used methods. Additionally, it can engage in discussions on specific topics, much like a peer. However, it lacks creative thinking and cannot delve into the finer details of particular issues. Nonetheless, it serves as a valuable resource for providing quick, introductory guidance to a given field.

### **Prompt**

If you are an anthropologist and have discovered ancient human hair at a certain location, how would you obtain key anthropological information from this hair?

### **O1-preview**

If you are an anthropologist who has discovered ancient human hair at a specific location, you can obtain key anthropological information through several methods:

1. **DNA Analysis:** Extracting DNA from the hair can provide valuable genetic information. This includes insights into genetic relationships, ancestry, population migration patterns, and evolutionary history. Both mitochondrial DNA (passed down maternally) and nuclear DNA can be analyzed to construct genetic lineages and connections with contemporary or ancient populations.
  2. **Isotopic Analysis:** Examining the isotopic composition of elements like carbon, nitrogen, and sulfur in the hair can reveal details about the individual's diet, such as whether they consumed primarily marine or terrestrial food sources. It can also provide information on the geographic region where the person lived, based on local environmental isotopic signatures.
  3. **Radiocarbon Dating:** Using radiocarbon dating techniques on the hair can help determine its age. This provides a chronological context for the findings, allowing you to place the individual within a specific historical period.
  4. **Microscopic Examination:** Analyzing the hair's microscopic structure can offer insights into the health and lifestyle of the individual. Signs of malnutrition, disease, or hair treatments (like dyeing or styling practices) can be detected through morphological studies.
  5. **Chemical Residue Analysis:** Testing for the presence of toxins, heavy metals, or other chemical compounds can indicate environmental exposures or dietary habits. For example, high levels of mercury might suggest certain medicinal practices or occupational hazards.
  6. **Paleopathological Assessment:** Looking for abnormalities or damage in the hair structure can provide evidence of illnesses or stressors the individual experienced, such as infections or hormonal imbalances.
  7. **Contextual Archaeological Analysis:** Understanding the location and manner in which the hair was found can shed light on cultural practices, such as burial rituals, ceremonial offerings, or personal grooming habits.
- By combining these methods, you can piece together a comprehensive picture of the individual's life, health, and the broader culture and environment in which they lived.

### **Reference Answer**

Very professional and knowledgeable, like a paleoanthropologist, 5 points.

Figure 56: **Anthropology Reasoning: Case 1.** An example illustrating the capability of o1-preview to respond to specialized questions in the domain of paleoanthropology. Green denotes that the answers are correct.

### **Prompt**

If you are a paleoanthropologist, what precautions should be taken regarding the use of mitochondrial DNA to avoid inaccurate data?

### **O1-preview**

When using mitochondrial DNA (mtDNA) in paleoanthropological research, it is crucial to take several precautions to avoid inaccurate data:

1. Preventing Contamination:

Sterile Handling: Work in clean rooms with filtered air systems to prevent contamination from modern DNA.

Protective Gear: Wear gloves, masks, and lab coats that are frequently changed to minimize the introduction of external DNA.

Dedicated Equipment: Use equipment and reagents dedicated exclusively to ancient DNA work.

Negative Controls: Include extraction and PCR blanks to monitor for contamination.

2. Authentication of Ancient DNA:

Replication: Replicate results independently, ideally in different laboratories, to confirm findings.

Damage Patterns: Look for characteristic patterns of DNA damage (e.g., increased deamination leading to C→T and G→A substitutions at molecule ends) that indicate ancient origin.

Molecular Copy Number: Assess the number of DNA copies to ensure that results are not skewed by a small number of contaminating molecules.

3. Avoiding Nuclear Mitochondrial Inserts (NumtS):

Primer Design: Use primers that specifically amplify mtDNA regions without amplifying nuclear pseudogenes.

Sequence Verification: Carefully analyze sequences to differentiate between true mtDNA and NumtS, which can be similar but may lead to erroneous conclusions.

4. Accounting for Post-Mortem Damage:

Damage Repair Enzymes: Consider using enzymes that can repair some types of DNA damage before amplification.

High-Fidelity Polymerases: Use DNA polymerases that are tolerant of damaged DNA but have high fidelity to reduce errors during amplification.

Bioinformatic Tools: Employ software that models and accounts for typical ancient DNA damage patterns during data analysis.

5. Optimizing DNA Recovery:

Sample Selection: Choose samples with the best preservation conditions, such as those from cooler, dryer environments.

Minimal Destruction: Use extraction methods that maximize DNA yield while minimizing destruction of valuable specimens.

Fragment Size Consideration: Target shorter DNA fragments, as ancient DNA is often highly fragmented.

6. Ethical and Legal Considerations:

Permissions: Obtain all necessary permits and follow legal requirements for working with human remains.

Cultural Sensitivity: Engage with descendant communities and consider the cultural implications of genetic research on human remains.

7. Accurate Data Interpretation:

Population Structure Awareness: Be cautious of oversimplifying genetic relationships; ancient populations may not align neatly with modern groupings.

Avoiding Reference Bias: Use diverse reference sequences to prevent skewed results due to reliance on limited modern DNA databases.

Statistical Validation: Apply rigorous statistical methods to validate findings and assess the significance of results.

8. Reporting Standards:

Transparent Methodology: Provide detailed descriptions of methods and precautions taken to allow for reproducibility and assessment by the scientific community.

Data Sharing: Share data in accordance with ethical guidelines, ensuring that sensitive information is protected.

9. Technological Updates:

Stay Informed: Keep abreast of the latest developments in ancient DNA extraction, sequencing technologies, and analytical methods.

Training: Ensure that all team members are properly trained in ancient DNA techniques and contamination prevention.

By meticulously applying these precautions, you can minimize errors and obtain more reliable and accurate mtDNA data, leading to meaningful insights into ancient human populations.

### **Reference Answer**

Very professional and knowledgeable, like a paleoanthropologist, 5 points.

Figure 57: **Anthropology Reasoning: Case 2.** An example illustrating the capability of o1-preview to respond to specialized questions in the domain of paleoanthropology. Green denotes that the answers are correct.

### **Prompt**

If you are a paleoanthropologist, and I, as a student, want to use ancient human genetic data to study the interactions between tribes and other groups during that time, could you offer me some advice?

### **O1-preview**

Certainly! Studying ancient human genetic data to understand interactions between tribes and other groups is a fascinating area of research in paleoanthropology. Here are some suggestions to guide your study:

1. Literature Review:

Familiarize with Previous Studies: Start by reading key research papers on **ancient DNA (aDNA) studies related to population interactions, migrations, and admixture events.**

Theoretical Frameworks: Understand the models and theories that **explain human migration patterns, such as the Out-of-Africa theory, multiregional hypothesis, and concepts of gene flow.**

2. Data Acquisition:

Accessing Genetic Data: Obtain genetic data from reputable databases like the **European Nucleotide Archive (ENA)** or the **National Center for Biotechnology Information (NCBI)**. Some studies also publish their raw data or provide access upon request.

Sample Selection: Focus on samples that **are geographically and temporally** relevant to your research question. Ensure the samples have sufficient coverage and quality for analysis.

3. Methodological Approaches:

**Mitochondrial DNA and Y-Chromosome Analysis:** Use mtDNA (maternal lineage) and Y-chromosome DNA (paternal lineage) to trace lineage-specific migrations and interactions.

**Autosomal DNA Analysis:** Examine autosomal DNA for a more comprehensive understanding of genetic relationships and admixture between populations.

**Population Genetics Tools:** Utilize software like ADMIXTURE, STRUCTURE, or PCA (Principal Component Analysis) to identify genetic clusters and infer population structure.

**Phylogenetic Analysis:** Construct phylogenetic trees to visualize evolutionary relationships between different groups.

4. Interpreting Genetic Signals:

**Gene Flow and Admixture:** Look for signs of genetic mixing, which may indicate interactions such as trade, intermarriage, or conquest.

**Isolation by Distance:** Consider how geographical barriers might have limited or facilitated interactions between groups.

**Temporal Dynamics:** Incorporate radiocarbon dating to correlate genetic data with specific time periods **helping to contextualize interactions historically.**

5. Integrating Multidisciplinary Data:

**Archaeological Evidence:** Combine genetic findings with archaeological records, such as **artifacts, settlement patterns, and burial practices, to strengthen interpretations.**

**Linguistic Studies:** Explore correlations between genetic data and linguistic similarities or differences among ancient populations.

**Environmental Context:** Consider how climate and environmental changes may have influenced migration and interaction patterns.

6. Ethical Considerations:

**Cultural Sensitivity:** Be mindful of the descendant communities related to the ancient populations you are studying. Engage with them respectfully and, if possible, involve them in your research.

**Legal Permissions:** Ensure that all genetic data usage complies with legal and ethical guidelines, including obtaining necessary permissions for data access and publication.

7. Collaborations:

**Mentorship:** Seek guidance from experienced researchers in ancient DNA studies or population genetics. **Interdisciplinary Teams:** Collaborate with archaeologists, geneticists, bioinformaticians, and anthropologists to enrich your study.

8. Technical Proficiency:

**Bioinformatics Skills:** Develop proficiency in bioinformatics tools and programming languages like Python or R for data analysis. **Statistical Analysis:** Strengthen your understanding of statistical methods used in genetic studies to accurately interpret results.

9. Potential Challenges:

**DNA Degradation:** Ancient DNA is often fragmented and degraded. Be prepared to address issues related to DNA preservation and contamination.

**Sample Bias:** Acknowledge and account for biases due to uneven sample distribution across regions or time periods.

**10. Communication of Findings:**

Academic Writing: Enhance your skills in writing clear and concise research papers or reports.

Presentations: Share your findings through conferences or seminars to receive feedback and engage with the academic community.

**11. Staying Updated:**

Current Research: Keep up with the latest developments in ancient DNA technologies, such as next-generation sequencing and improved computational methods.

Workshops and Courses: Attend workshops or enroll in courses specializing in ancient DNA analysis and population genetics.

**12. Formulating a Clear Research Question:**

Specificity: Define a focused research question or hypothesis about the interactions between specific tribes or groups.

Objectives: Outline clear objectives and expected outcomes for your study.

By carefully planning your research and utilizing the available resources and methodologies, you can gain valuable insights into the interactions between ancient human populations. Remember to approach your study with scientific rigor and ethical responsibility.

**Reference Answer**

Like a professor in paleoanthropology, 5 points.

Figure 58: **Anthropology Reasoning: Case 3.** An example illustrating the capability of o1-preview to respond to specialized questions in the domain of paleoanthropology. Green denotes that the answers are correct.

### **Prompt**

As a paleoanthropologist, how do you conduct genetic variation analysis?

### **O1-preview**

As a paleoanthropologist, conducting genetic variation analysis involves a multidisciplinary approach that combines fieldwork, laboratory techniques, and computational methods to study the genetic makeup of ancient human populations. The primary goal is to understand evolutionary relationships, migration patterns, adaptation, and interactions among early human groups. Here's how the process typically unfolds:

#### 1. Sample Collection:

Archaeological Excavation: Collect ancient human remains such as bones, teeth, or hair from archaeological sites with meticulous documentation of their context and stratigraphy.

Preservation Assessment: Evaluate the preservation state of the samples, as DNA degrades over time and is sensitive to environmental conditions.

#### 2. Ancient DNA (aDNA) Extraction:

Clean Laboratory Conditions: Perform DNA extraction in specialized clean rooms dedicated to aDNA work to prevent contamination from modern DNA.

Chemical Treatment: Use protocols optimized for ancient samples, which often involve demineralization and the use of enzymes to release DNA from the cells.

Contamination Controls: Include negative controls to monitor for contamination throughout the extraction process.

#### 3. DNA Sequencing:

Library Preparation: Convert extracted DNA into a format suitable for sequencing, often involving the addition of adapters to fragmented DNA molecules.

Next-Generation Sequencing (NGS): Utilize high-throughput sequencing platforms (e.g., Illumina, Ion Torrent) capable of reading millions of DNA fragments simultaneously.

Targeted Enrichment: Employ techniques like hybridization capture to enrich for specific genomic regions of interest, such as mitochondrial DNA or particular nuclear DNA segments.

#### 4. Data Processing and Quality Control:

Sequence Alignment: Map the sequenced DNA fragments to a reference genome to identify their origin within the genome.

Damage Pattern Analysis: Assess typical aDNA damage patterns (e.g., cytosine deamination) to authenticate the ancient origin of the DNA.

Contamination Assessment: Estimate the proportion of modern human DNA contamination using computational methods.

#### 5. Genetic Variation Analysis:

Mitochondrial DNA (mtDNA) and Y-Chromosome Analysis: Examine mtDNA and Y-chromosome markers to trace maternal and paternal lineages, respectively.

Autosomal DNA Analysis: Analyze nuclear DNA to investigate genetic relationships and diversity within and between populations.

**Single Nucleotide Polymorphisms (SNPs):** Identify and compare SNPs to study genetic variation.

Population Structure Analysis: Use statistical models (e.g., Principal Component Analysis, ADMIXTURE) to infer population structure and admixture events.

Phylogenetic Reconstruction: Build phylogenetic trees to depict evolutionary relationships among individuals or groups.

#### 6. Computational and Statistical Methods:

Bioinformatics Tools: Employ software like PLINK, BEAST, or STRUCTURE for genetic data analysis.

Demographic Modeling: Apply coalescent theory and models (e.g., Approximate Bayesian Computation) to infer past population sizes, divergence times, and migration rates.

Admixture Detection: Identify gene flow between populations using methods like f-statistics or D-statistics.

#### 7. Integration with Other Disciplines:

Archaeological Context: Correlate genetic findings with archaeological evidence such as artifacts, settlement patterns, and dietary remains.

Paleoclimatology: Consider environmental data to understand how climate changes may have influenced human migration and interaction.

Linguistics and Cultural Studies: Integrate linguistic data and cultural practices to provide a comprehensive view of past human societies.

#### 8. Ethical Considerations:

Permissions and Regulations: Obtain necessary permits and comply with legal and ethical guidelines for working with human remains.

Community Engagement: Collaborate with indigenous and local communities, respecting their perspectives and wishes regarding the research.

Data Sharing and Privacy: Handle genetic data responsibly, ensuring confidentiality and appropriate use.

**9. Interpretation and Reporting:**

Critical Analysis: Interpret results cautiously, acknowledging limitations such as sample size, preservation biases, and potential contamination.

Peer Review: Publish findings in scientific journals after rigorous peer review to validate the research.

Public Communication: Share results with the broader community through lectures, exhibitions, or accessible publications, while avoiding sensationalism.

**10. Continuous Learning and Collaboration:**

**Stay Updated:** Keep abreast of advancements in sequencing technologies, analytical methods, and theoretical frameworks.

Interdisciplinary Collaboration: Work with geneticists, bioinformaticians, archaeologists, and other specialists to enhance the quality and scope of the research.

**Challenges to Consider:**

DNA Degradation: Ancient DNA is often fragmented and chemically modified, requiring specialized techniques to retrieve and analyze it.

Contamination Risk: Modern DNA contamination is a significant concern, necessitating strict laboratory protocols and authentication methods.

Interpretational Complexity: Genetic data must be interpreted within a broader archaeological and environmental context to draw meaningful conclusions.

**Example Applications:**

Tracing Migration Routes: By analyzing genetic similarities and differences, it's possible to reconstruct ancient human migration patterns, such as the peopling of the Americas or the spread of Neolithic farming.

Studying Adaptations: Investigate genetic adaptations to environmental pressures, such as high-altitude hypoxia or lactose tolerance.

Understanding Social Structures: Genetic data can shed light on kinship patterns, marriage practices, and social organization within ancient communities.

By employing these methods, paleoanthropologists can unlock valuable insights into human history, evolution, and diversity, contributing to our understanding of how modern populations came to be.

**Reference Answer**

Addressed the key points, very professional, 5 points.

Figure 59: **Anthropology Reasoning: Case 4.** An example illustrating the capability of o1-preview to respond to specialized questions in the domain of paleoanthropology. Green denotes that the answers are correct.

### **Prompt**

As a geologist, the discovery of dropstones (also known as erratics) with low roundness and angular shapes, seemingly "suspended" in marine sedimentary rock layers like mudstone, can be explained by the process of glacial transport and ice-raftered debris (IRD) deposition

### **O1-preview**

Yes, that's correct. The presence of dropstones or erratics with low roundness and angular shapes embedded in fine-grained marine sedimentary rocks like mudstone is explained by glacial transport and ice-raftered debris (IRD) deposition. Here's how the process works:

1. Glacial Erosion and Transport: Glaciers move across the land, eroding the underlying rock and picking up a range of debris, from fine sediments to large boulders. This material becomes entrained within the ice.
2. Calving and Iceberg Formation: When glaciers reach the sea, chunks of ice break off in a process called calving, forming icebergs that float in the ocean. **These icebergs carry the embedded rocks and sediments with them.**
3. Ice-Rafted Debris Deposition: As icebergs drift away from their source, they gradually melt due to warmer ocean temperatures. **When the ice melts sufficiently, the entrained debris is released and drops to the ocean floor.** This material can range from tiny grains to large, angular boulders.
4. Sedimentation in Marine Environments: The ocean floor, especially in calm marine settings, **is often accumulating fine-grained sediments like mud and clay.** When the larger rocks from melting icebergs settle through the water column, they become embedded in these finer sediments.
5. Formation of Dropstones: Over time, the accumulation of sediments buries these rocks, and they become part of the **sedimentary record.** In the resulting sedimentary rock layers, these larger clasts appear **"suspended"** within the **finer matrix,** creating what geologists refer to as dropstones.

Significance in Geology:

**Paleoclimate Indicators:** The presence of dropstones in marine sedimentary rocks is a key indicator of past glacial activity and can provide evidence for ancient ice ages and the extent of glaciation.

**Reconstruction of Past Environments:** Studying the characteristics of dropstones (such as their composition and source) helps geologists understand past geological processes, including glacier movement and the transportation paths of icebergs.

**Stratigraphic Markers:** **Dropstones can serve as important markers within the stratigraphic record, helping to correlate rock layers across different geographic regions.**

In summary, the angular and poorly rounded nature of these dropstones suggests they have not been significantly eroded by water transport but were instead transported by ice and deposited directly onto the seafloor as the ice melted. This process explains their occurrence within fine-grained marine sedimentary layers like mudstone.

### **Reference Answer**

Not only answered the question correctly but also explained the underlying principles, 5 points.

Figure 60: **Geology Reasoning: Case 1.** An example illustrating the capability of o1-preview to respond to specialized questions in the domain of Geology. Green denotes that the answers are correct.

---

## 4.23 Educational Measurement and Psychometrics

In evaluating o1-preview's performance in educational measurement and psychometrics, we selected multiple representative quiz questions from an introduction-level course named *Measurement Theory* based on Bandalos' textbook [7] for first-year graduate students. Our goal is to evaluate the performance of the o1-preview in addressing key concepts in a relatively narrow domain from both theoretical and practical perspectives, including test score reliability, standard error of measurement, and interrater reliability. Each prompt was designed to test the o1-preview's efficacy in providing concrete explanations, applying appropriate statistical techniques, and offering solutions that align with established psychometric frameworks. The following sections provide detailed descriptions of all testing scenarios:

For the first task in Figure 61, which involved calculating a reliability coefficient using the Spearman-Brown Prophecy formula, o1-preview demonstrated impressive accuracy. The model not only applied the correct formula but also provided a detailed explanation of each step, making the process informative and easy to follow. This suggests that o1-preview is proficient in handling statistical calculations related to test reliability, which is an important area of educational psychometrics. Its ability to calculate and articulate the reasoning behind the steps reflects a strong understanding of this foundational area in the field.

In the second task in Figure 62, which assessed o1-preview understanding of the standard error of measurement and confidence intervals, the model once again delivered a solid response. It accurately calculated the correct range of scores, demonstrating its capability to identify and apply measurement concepts to real-world test scenarios. This further highlights o1-preview's ability to provide reliable and informed outputs in educational measurement tasks.

However, o1-preview's performance in the third task in Figure 63 exposed some limitations in disentangling similar concepts. The third task aims to solve a multiple choice question inculding more than one correct response that evaluates understanding concepts of interrater agreement and interrater reliability, which are two close concepts that measure the consistency of multiple raters' ratings on a performance assessment (e.g., an essay). The model incorrectly identified nominal agreement and Cohen's kappa as measures of interrater reliability, when the appropriate measures should have been Intraclass Correlation (ICC) and Coefficient Alpha. However, the o1-preview offered the correct definition of the four measures together with the practical application scenarios. This response showcases o1-preview's lack of understanding of the connections between key concepts when the potential training set provides limited information.

In sum, the o1-preview excelled in tasks requiring statistical computation and reasoning, demonstrating a solid grasp of standard psychometric concepts that are equivalent to or beyond a first-year master's or doctoral student's level in educational measurement and psychometrics. However, our tests revealed some weaknesses in its ability to distinguish closely related concepts even under simple questions.

### **Prompt**

A test with 10 questions has a reliability estimate of .70. According to the Spearman Brown Prophecy formula, what would the reliability be if 10 more questions (of equal quality) were added to the test?

### **O1-preview**

To determine the new reliability of the test after adding 10 more questions of equal quality, we'll use the **Spearman-Brown Prophecy Formula**. This formula predicts how changing the length of a test affects its reliability.

$$r_{new} = \frac{n \times r_{old}}{1 + (n - 1) \times r_{old}}$$

- $r_{new}$  = New reliability
- $r_{old}$  = Original reliability (0.70)
- $n$  = Factor by which the test length is increased

#### **Step-by-Step Calculation:**

##### **1. Determine the Value of $n$ :**

- Original number of questions: 10
- Additional questions added: 10
- Factor of increase ( $n$ ):

$$n = \frac{\text{New Length}}{\text{Original Length}} = \frac{20}{10} = 2$$

##### **2. Apply the Spearman-Brown Formula:**

$$r_{new} = \frac{2 \times 0.70}{1 + (2 - 1) \times 0.70}$$

##### **3. Simplify the Denominator:**

$$1 + (2 - 1) \times 0.70 = 1 + 1 \times 0.70 = 1 + 0.70 = 1.70$$

##### **4. Compute the New Reliability:**

$$r_{new} = \frac{1.40}{1.70}$$

##### **5. Calculate the Final Value:**

$$r_{new} = 0.8235 \approx 0.82$$

#### **Conclusion:**

- The new reliability estimate is approximately **0.82**
- By adding 10 more questions of equal quality to the original 10-question test, the reliability increases from 0.70 to approximately 0.82

#### **Reference Answer**

**0.82**

Figure 61: **Educational Measurement and Psychometrics: Case 1.** An example illustrating the capability of o1-preview to respond to specialized questions of test reliability in the domain of educational measurement and psychometrics. Green indicates complete consistency.

**Prompt**

The standard error of measurement of a set of test scores is 2. Kelly obtained a score of 28 on the test. In which of the following ranges of scores can we have 95% confidence that Kelly's true score falls?

- a. 22–34
- b. 27–39
- c. 26–30
- d. 24–32

**Q1-preview**

Answer:

**d. 24–32**

**Explanation:**

To determine the range within which we can have 95% confidence that Kelly's true score falls, we'll use the Standard Error of Measurement (SEM) and calculate the confidence interval around Kelly's observed score.

Understanding the **Standard Error of Measurement (SEM)**:

- SEM quantifies the amount of error inherent in an individual's observed test score due to imperfect reliability of the test.
- It provides an estimate of how much an individual's observed score might fluctuate upon repeated testing under identical conditions. Calculating the 95% Confidence Interval:

The formula for the confidence interval around an observed score is:

Confidence Interval = Observed Score  $\pm$  (Z  $\times$  SEM)

- **Observed Score:** 28
- **SEM:** 2
- **Z-score for 95% confidence:** Approximately 1.96

**Step-by-Step Calculation:****1. Calculate the Margin of Error:**

$$\text{Margin of Error} = Z \times \text{SEM} = 1.96 \times 2 = 3.92$$

**2. Determine the Lower and Upper Limits of the Confidence Interval:**

- Lower Limit: Lower Limit = Observed Score – Margin of Error = 28 – 3.92 = 24.08
- Upper Limit: Upper Limit = Observed Score + Margin of Error = 28 + 3.92 = 31.92

**3. Construct the Confidence Interval:** Confidence Interval = [24.08, 31.92]**Matching with the Given Options:**

- Option a. 22–34: Lower limit is less than calculated; upper limit is higher.
- Option b. 27–39: Upper limit exceeds the calculated interval significantly.
- Option c. 26–30: Narrower interval than calculated.
- Option d. 24–32: Matches most closely with the calculated interval.

**Conclusion:**

- Option d (24–32) aligns closely with the calculated 95% confidence interval.
- Therefore, we can be 95% confident that Kelly's true score falls between 24 and 32.

**Summary:**

Using the SEM and the Z-score for 95% confidence, we calculated the confidence interval to be approximately 24 to 32. Thus, option d is the correct choice.

**Reference Answer**

**Option d**

Figure 62: **Educational Measurement and Psychometrics: Case 2.** An example illustrating the capability of o1-preview to respond to specialized questions of standard error of measurement in the domain of educational measurement and psychometrics. Green indicates complete consistency.

**Prompt**

Which of the following is a measure of interrater reliability: (choose all that apply)

- a. Nominal agreement
- b. Intraclass correlation (ICC)
- c. Cohen's kappa
- d. Coefficient alpha

**O1-preview****Correct Answers:**

- a. Nominal agreement
- b. Intraclass correlation (ICC)
- c. Cohen's kappa

**Explanation:****a. Nominal Agreement:**

- **Definition:** Nominal agreement measures the percentage of times different raters agree when assigning items to nominal (categorical) categories.
- **Usage:** It's a straightforward method for assessing interrater reliability when data are categorical.
- **Limitation:** It doesn't account for agreement occurring by chance.

**b. Intraclass Correlation (ICC):**

- **Definition:** The Intraclass Correlation Coefficient (ICC) assesses the consistency or conformity of measurements made by multiple observers measuring the same quantity.
- **Usage:** Commonly used for continuous data, ICC evaluates both the correlation and agreement between raters.
- **Types:** There are different forms of ICC depending on the study design and the assumptions about the raters.

**c. Cohen's Kappa:**

- **Definition:** Cohen's kappa measures the level of agreement between two raters beyond what would be expected by chance.
- **Usage:** Suitable for categorical (nominal or ordinal) data, it provides a more robust measure than simple percent agreement.
- **Interpretation:** A kappa value of 1 indicates perfect agreement, while 0 indicates no agreement better than chance.

**d. Coefficient Alpha (Cronbach's Alpha):**

- Not a measure of interrater reliability.
- **Definition:** Cronbach's alpha assesses the internal consistency reliability of a set of test items, evaluating how closely related the items are as a group.
- **Usage:** It is used to estimate the reliability of a composite score (e.g., a scale or test), not the agreement between different raters.

**Summary:**

- **Interrater Reliability Measures:** Nominal agreement, Intraclass Correlation (ICC), and Cohen's kappa are all measures used to assess the degree of agreement or consistency between different raters or observers.
- **Internal Consistency Measure:** Coefficient alpha (Cronbach's alpha) measures internal consistency within a test or scale and is not used for assessing interrater reliability.

**Reference Answer**

Options b) and d)

Figure 63: **Educational Measurement and Psychometrics: Case 3.** An example illustrating the capacity of the o1-preview to respond to specialized questions of interrater reliability in the domain of educational measurement and psychometrics. Red represents inconsistency between the o1-preview response and the reference answer, yellow indicates partial consistency, while green signifies complete consistency.

---

## 4.24 Public Health Policy Analysis

Throughout the evaluation of Public Health Policy Analysis, we tested o1-preview's capacity to analyze health policy with respect to the Affordable Care Act (ACA). The test targeted measuring the model's accuracy in responding to questions about the ACA's effects on insurance coverage, healthcare disparities, preventive services, and Medicaid expansion.

Based on the test samples, o1-preview demonstrated strong alignment with reference answers, showing a clear grasp of the ACA's key impacts. For example, in 64 Case 1, the model successfully identified the overall trend but overlooked specifics, such as the increased listings for heart and lung transplants. In 64 Case 2, it provided precise insights into increased coverage and access to care, closely matching the reference answer.

However, the o1-preview model occasionally lacked depth in addressing specific topics, as shown in 65, where key disease-specific examples were omitted. This limitation reduced its ability to capture the nuances of the ACA's effects entirely on certain populations. This suggests that the model requires further refinement, especially in areas demanding advanced analysis and detailed specificity.

In this domain, o1-preview performed robustly in analyzing the Affordable Care Act and its broader public health impacts. The model effectively highlighted significant elements of health policy, such as Medicaid expansion, preventive services, and efforts to address healthcare disparities. Despite some limitations in providing detailed explanations, its reasoning aligned well with the ground truth across most cases.

For future evaluations, we could consider using larger datasets, such as the *Compilation of State Data on the Affordable Care Act*[44], which would allow for more thorough training and testing of o1-preview's abilities. Additionally, incorporating more evaluation metrics, along with broader perspectives on policy evaluation, would offer a more comprehensive assessment of the model's public health policy analysis capabilities.

With further improvements in providing more specific examples and advanced analysis, o1-preview will exhibit significant potential for excelling in public health policy analysis and policy-making tasks.

**Prompt**

How successful was the ACA at expanding insurance coverage and timely access to care among surgical populations? return brief summary

**O1-preview**

The Affordable Care Act (ACA) significantly expanded insurance coverage among surgical populations by increasing access through Medicaid expansion and health insurance marketplaces. Studies have shown a substantial decrease in the number of uninsured surgical patients post-ACA implementation. This increase in coverage improved timely access to surgical care, allowing more patients to receive necessary procedures without financial delays. However, while insurance coverage improved, some disparities in access to timely care persisted due to factors like provider availability and healthcare infrastructure.

**Reference Answer**

The ACA led to earlier presentation of acute surgical diseases, improved cancer screening, earlier stages of diagnosis , increased listings for heart and lung transplants, and increased rehabilitation after trauma. However, no access to significant inpatient mortality benefits were identified.

**Prompt**

In what ways did the ACA impact preventive services, particularly for chronic disease management in minority populations?

**O1-preview**

The Affordable Care Act (ACA) improved access to preventive services by requiring insurance plans to cover preventive care without cost-sharing. This mandate particularly benefited minority populations, who historically face higher rates of chronic diseases like diabetes and hypertension. By expanding Medicaid and offering subsidies for private insurance, the ACA increased coverage among minorities, enabling more individuals to access screenings, vaccinations, and routine check-ups. This led to earlier detection and better management of chronic conditions, helping to reduce health disparities and improve overall health outcomes in these communities.

**Reference Answer**

The ACA improved access to preventive services by ensuring insurance coverage for services like cancer screenings. Medicaid expansion led to earlier detection and treatment of diseases such as head and neck cancer particularly among low-income and minority populations.

Figure 64: **Public Health Policy Analysis: Case 1-2.** Two examples illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

### **Prompt**

How did the ACA address healthcare disparities, and what challenges remain in reducing these disparities?

### **O1-preview**

The Affordable Care Act (ACA) addressed healthcare disparities by expanding insurance coverage and improving access to care for underserved populations:

- Medicaid Expansion: The ACA allowed states to expand Medicaid eligibility to more low-income adults, disproportionately benefiting racial and ethnic minorities who were more likely to be uninsured.
- Health Insurance Marketplaces: Subsidies and tax credits made private insurance more affordable, increasing coverage among lower-income individuals and families.
- Preventive Services Coverage: Requiring insurers to cover preventive services without cost-sharing improved access to screenings and vaccinations, aiding early detection and management of diseases prevalent in minority communities.
- Funding for Community Health Centers: Increased support enhanced healthcare services in medically underserved areas, where many minority populations reside.
- Data Collection and Reporting: Improved data on healthcare disparities enabled better-targeted interventions and policies to address specific needs of different populations.

#### **Challenges Remaining:**

- Medicaid Expansion Gaps: Not all states chose to expand Medicaid, leaving many low-income individuals without coverage, particularly in states with larger minority populations.
- Affordability Issues: High premiums and out-of-pocket costs continue to be barriers for some, despite subsidies.
- Access to Care: Insurance coverage doesn't always equal access due to provider shortages, especially in rural and underserved urban areas.
- Cultural and Language Barriers: Lack of culturally competent care and language services can hinder effective healthcare delivery to minority groups.
- Social Determinants of Health: Factors like poverty, education disparities, and inadequate housing continue to contribute to health inequalities.
- Undocumented Immigrants: Exclusion from ACA benefits leaves this population without access to affordable healthcare.

Addressing these challenges requires continued efforts to expand coverage, reduce costs, enhance culturally competent care, and tackle the underlying social determinants that contribute to health disparities.

### **Reference Answer**

The ACA reduced the racial/ethnic insurance coverage gap, especially for Black and Hispanic populations in Medicaid expansion states. However, a significant portion of these populations still reside in states that have not expanded Medicaid, which continues to pose a challenge in fully addressing disparities.

Figure 65: **Public Health Policy Analysis: Case 3.** An example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview's response and the reference answer, and red indicates full consistency, while green signifies complete consistency.

---

## 4.25 Medical Genetics and Genomics Reasoning

To evaluate the model’s reasoning ability in genomics and medical genetics, we designed a series of experiments centred around genomics-related questions, requiring the model to generate appropriate responses to predefined questions. These questions focused on relationship predictions and their corresponding explanations. The tasks aimed to assess different facets of genomics reasoning and utilized selected QA tasks from the GenTurning benchmarks [52]. In all the tested 20 QA instances, o1-preview yields correct answers. Considering the related information we provided in the prompt, the o1-preview has the ability to generate reliable conclusions by following a correct chain-of-thought which is derived from the truth. On observing the details of the reasoning steps, the o1-preview provided explanations for all aspects and showed clearly how it came to the conclusions. We show a true-positive case in (Figure 66) and a false-negative case in (Figure 67), which in both cases, o1-preview shows satisfactory results. In the view of inference time consumption, it is observed that for the negative terms, the o1-preview takes a significantly longer time to draw the conclusion. In our test, the average running time of a true-positive conclusion is 12.67s, while the average running time of a false-negative conclusion is 20.09s. Based on this observation, it is reasonable to speculate that the o1-preview takes a longer chain-of-thought in attempting to reject a genomics hypothesis. Additional examples can be found in (Appendix A.18).

**Prompt**

You are the genetic medical researcher, based on the disease description and gene description. **reasoning if the gene is related to the disease, give your inference and argumentations.**

**Gene description:**

ELMO proteins, such as ELMO2, are homologs of the *C. elegans* Ced12 protein, which is required for apoptotic cell engulfment and cell migration. ELMO proteins have no apparent enzymatic activity and exert their biological functions through interactions with various plasma membrane-associated and cytosolic proteins. In general, ELMO proteins are recruited to the plasma membrane via interacting proteins for efficient activation of downstream effectors.

**Disease description:**

Primary intraosseous vascular malformation (VMPI), previously called intraosseous hemangioma, is a rare malformation that usually involves the vertebral column and the skull. The most commonly affected bones in the skull are the mandible and the maxilla, and life-threatening bleeding after a simple tooth extraction is frequent.

**O1-preview**

Based on the provided descriptions, there is a **plausible connection** between the **ELMO2 gene** and **primary intraosseous vascular malformation (VMPI)**. Here's the **reasoning**:

**\*\*Role of ELMO2 in Cellular Functions:\*\***

- **\*\*Cell Migration:\*\*** ELMO2 is involved in cell migration through its interactions with plasma membrane-associated and cytosolic proteins. Cell migration is a critical process in angiogenesis—the formation of new blood vessels from pre-existing ones.
- **\*\*Apoptotic Cell Engulfment:\*\*** ELMO2 **plays a role** in the clearance of apoptotic cells. Efficient removal of apoptotic cells is essential for normal tissue remodeling and development.

**\*\*Pathogenesis of VMPI:\*\***

- **\*\*Vascular Malformations:\*\*** VMPI involves abnormal vascular development within bones, leading to malformations. These malformations can cause life-threatening bleeding, especially in bones like the mandible and maxilla.
- **\*\*Endothelial Cell Dysfunction:\*\*** Vascular malformations are often linked to abnormalities in endothelial cell function and migration during vascular development.
- \*\*Potential Connection Between ELMO2 and VMPI:\*\***
- **\*\*Impact on Angiogenesis:\*\*** Given ELMO2's role in cell migration, it may influence endothelial cell behavior during angiogenesis. Dysregulation of ELMO2 could lead to abnormal vessel formation within bone tissue.
- **\*\*Tissue Remodeling:\*\*** ELMO2's involvement in apoptotic cell clearance may affect tissue remodeling processes essential for normal vascular development in bones.

**\*\*Inference:\*\***

- **\*\*Genetic Association:\*\*** Mutations or altered expression of ELMO2 might contribute to the development of VMPI by affecting endothelial cell migration and apoptotic cell clearance during bone vascularization.
- **\*\*Research Implications:\*\*** Further studies investigating ELMO2 expression levels and mutations in patients with VMPI could provide insights into its role in the disease.

**\*\*Conclusion:\*\***

Considering ELMO2's **functions in cell migration and apoptotic cell clearance**—both critical in vascular development and tissue remodeling—it is reasonable to infer that ELMO2 may be related to the pathogenesis of primary intraosseous vascular malformation. Further research is warranted to establish a definitive link.

**Running Time: (6s)**

**Reference Answer: Gene: ELMO2; Disease: Vascular malformation; Ground Truth: Related**

Figure 66: **Medical Genetic and Genomics Reasoning: Case 1.** Positive example illustrating the capability of the o1-preview to respond to gene-disease-related questions. Grey indicates the prompt instructions. Blue indicates the gene description and disease description. Green indicates the correct inferences. Yellow indicates the detailed deduction steps of the o1-preview.

**Prompt**

You are the genetic medical researcher, based on the disease description and gene description. **reasoning if the gene is related to the disease**, give your inference and argumentations.

**Gene description:**

LAMTOR2 is part of the LAMTOR complex, which is involved in recruitment and spatiotemporal activation of ERKs (see 601795) and MTOR complex-1 (MTORC1; see 601231). LAMTOR2 also participates in regulation of endosomal trafficking, growth factor signaling, cell proliferation, and Langerhans cell homeostasis.

**Disease description:**

Multiple familial trichoepithelioma (MFT), also called epithelioma adenoides cysticum (EAC), is an autosomal dominant dermatosis characterized by the presence of many skin tumors predominantly on the face. Since histologic examination shows dermal aggregates of basaloid cells with connection to or differentiation toward hair follicles, this disorder has been thought to represent a benign hamartoma of the pilosebaceous apparatus. Trichoepitheliomas can degenerate into basal cell carcinoma.

**O1-preview**

Based on the provided descriptions, there is **no direct evidence** to suggest that the **LAMTOR2** gene is related to **Multiple Familial Trichoepithelioma (MFT)**. However, given LAMTOR2's role in cell proliferation and growth factor signaling, it's worth exploring potential indirect connections. Here is the detailed reasoning:

### **Gene Description: LAMTOR2**

**- Function:**

- **LAMTOR2** is a component of the **Late Endosomal/Lysosomal Adaptor, MAPK and MTOR Activator 2 (LAMTOR) complex**.
- Involved in the **recruitment and spatiotemporal activation of ERKs** (Extracellular signal-regulated kinases) and **MTOR complex-1 (MTORC1)**.

**- Plays a crucial role in:**

- **Endosomal trafficking**. - **Growth factor signaling**. - **Cell proliferation**. - **Langerhans cell homeostasis**.

### **Disease Description: Multiple Familial Trichoepithelioma (MFT) Characteristics**

- An autosomal dominant skin disorder. Characterized by multiple skin tumors predominantly on the face.
- Histologically shows **dermal aggregates of basaloid cells** connected to or differentiating toward **hair follicles**.

**Considered a benign hamartoma of the pilosebaceous unit**

- **Trichoepitheliomas can degenerate into basal cell carcinoma**.

### **Inference and Argumentation**

**- Potential Connection via Cell Proliferation Pathways:**

- **LAMTOR2's Role in Cell Proliferation:**
  - Activates ERK and MTORC1 pathways, which are essential for cell growth and proliferation.
  - Dysregulation of these pathways is often implicated in tumorigenesis.

**- MFT and Cell Proliferation:**

- Abnormal signaling in pathways regulating cell proliferation could contribute to MFT development.

**- Conclusion:**

- While LAMTOR2 is involved in pathways that, when dysregulated, can lead to tumor growth, there is no specific evidence linking LAMTOR2 mutations to MFT.

**- LAMTOR2 Mutations:**

- Mutations in LAMTOR2 are associated with immunodeficiency disorders, not skin tumors.
- No documented cases link LAMTOR2 mutations to MFT or related skin conditions.

### **Conclusion**

- **No Direct Link Established**.

**Running Time: (21s)**

**Reference Answer: Gene: LAMTOR2; Disease: Trichoepithelioma; Ground Truth: False**

Figure 67: **Medical Genetics and Genomics Reasoning: Case 2.** Positive example illustrating the capability of o1-preview to respond to gene-disease-related questions. Grey indicates the prompt instructions. Blue indicates the gene description and disease description. Green indicates the correct inferences. Yellow indicates the detailed deduction steps of the o1-preview.

---

## 4.26 Medical Knowledge Question Answer

During the evaluation of o1-preview’s performance in Medical Knowledge Question Answer, we found that the model provided accurate and reasoned responses. The test aims to evaluate whether the model can select the correct options for medical knowledge questions and give reasonable explanations.

o1-preview demonstrated alignment with the reference answer. Upon selecting the correct option, it provided a reasonable explanation of the underlying knowledge, suggesting that the model comprehends the question’s content and identifies the key elements required to solve the problem. For instance, in Figure 68, when confronted with a relatively simple knowledge-based question, the model chose the correct option and provided a coherent explanation of the functions and effects of several fatty acids. Likewise, in Figure 69, when addressing a more complex scenario-based inference question, o1-preview selected the correct answer and offered common symptoms and explanations for four diseases, expanding upon the question’s scope. The explanation for the correct answer, Von Gierke disease, closely matched the standard answer explanation.

However, the model occasionally overlooks key information when dealing with specific topics, resulting in a lack of necessary logical judgment. Even when choosing the correct option, the model may miss key conditions when evaluating the wrong option. For example, in Figure 163, it ignores the fact that “autosomal recessive polycystic kidney disease” cannot occur in a “28-year-old male,” indicating an inability to use logical thinking to process multiple pieces of information in parallel. Despite these minor flaws, o1-preview demonstrates strong reasoning capabilities, especially when answering questions using large medical knowledge datasets.

In summary, o1-preview performs well in medical knowledge question answering. The model effectively captures the key information in the questions and can provide accurate options and effective explanations through retrieval. Although some explanations exhibit illogical reasoning and information gaps, in most cases, the overall reasoning is consistent with the facts. In the current test, the model achieved 100% accuracy. With further refinements, such as offering more specific examples and conducting deeper analyses, o1-preview shows great potential for medical knowledge question answering tasks.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "Which of the following is an essential fatty acid?"

"opa": "Linoleic acid", "opb": "Alpha linolenic acid", "opc": "Both of the above", "opd": "None of the above"

**O1-preview**

The correct answer is:

**C) Both of the above**

**Explanation:**

Essential fatty acids (EFAs) are fatty acids that the human body cannot synthesize on its own and must obtain through the diet. There are two primary families of essential fatty acids: omega-6 (n-6) and omega-3 (n-3) fatty acids.

1. **Linoleic Acid (LA):** This is an omega-6 fatty acid, specifically 18:2 (delta-9,12). **The human body lacks the enzymes necessary to introduce a double bond at the omega-6 position** (the sixth carbon from the methyl end). Therefore, linoleic acid must be obtained from dietary sources. It plays a crucial role in cell membrane structure and function, and it serves as a precursor for the synthesis of other bioactive lipids like arachidonic acid.
2. **Alpha-Linolenic Acid (ALA):** This is an omega-3 fatty acid, specifically 18:3 (delta-9,12,15). Similar to omega-6 fatty acids, **humans cannot introduce a double bond at the omega-3 position** (the third carbon from the methyl end). Alpha-linolenic acid is essential for normal growth and development and serves as a precursor for longer-chain omega-3 fatty acids like eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), which are important for cardiovascular and neurological health.

Since both linoleic acid and alpha-linolenic acid are essential fatty acids required in the human diet, the correct choice is:

**C) Both of the above**

**Reference Answer**

"exp": "The fatty acids that are required by humans, but are not synthesized in the body, hence need to be supplied in the diet are known as essential fatty acid (EFA). **Humans lack the enzymes that can introduce double bond beyond 9th Carbon**. They are Polyunsaturated Fatty Acid namely: Linoleic acid. Alpha linolenic acid. Arachidonic acid is considered as semi-essential fatty acid as, it can be synthesized from linoleic acid.",

"cop":3

Figure 68: **Medical Knowledge Question Answer: Case 1.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "A child presented with abdominal distension, hepatomegaly, doll like facies and recurrent episodes of hypoglycemia. Which of the following is the most likely diagnosis?"  
"opa": "Menkes disease", "opb": "Down syndrome", "opc": "Von gierke disease", "opd": "Lesch nyhan syndrome"

**O1-preview**

To determine the most likely diagnosis for the child described, let's analyze the key clinical features provided:

1. **Abdominal distension and hepatomegaly:** These symptoms indicate an enlarged abdomen and liver, suggesting storage of substances in the liver.
2. **Doll-like facies:** This refers to a characteristic facial appearance with fat cheeks, often seen in some metabolic disorders.
3. **Recurrent episodes of hypoglycemia:** Frequent low blood sugar levels indicate a problem with glucose metabolism or regulation.

Now, let's examine each option:

**Option A: Menkes Disease**

- **Menkes disease** is an X-linked recessive disorder caused by defective copper absorption and transport due to mutations in the ATP7A gene.
- **Clinical features** include:
  - **Brittle, kinky hair**
  - **Hypotonia (decreased muscle tone)**
  - **Developmental delay**
  - **Seizures**
- **Notable Absence:** Menkes disease does not typically cause hepatomegaly, abdominal distension, doll-like facies, or hypoglycemia.

**Option B: Down Syndrome**

- **Down syndrome** is a genetic disorder caused by trisomy 21.
- **Clinical features** include:
  - **Characteristic facial features** (flat facial profile, upward slanting eyes)
  - **Intellectual disability**
  - **Congenital heart defects**
  - **Increased risk of certain leukemias**
- **Notable Absence:** Down syndrome does not typically involve hepatomegaly or recurrent hypoglycemia, nor is "doll-like facies" a characteristic description.

**Option C: Von Gierke Disease**

- **Von Gierke disease** (Type I Glycogen Storage Disease) is caused by a deficiency of the enzyme **glucose-6-phosphatase**.

- Clinical features include:

- Severe fasting hypoglycemia due to impaired gluconeogenesis and glycogenolysis.
- Hepatomegaly and renomegaly (enlarged liver and kidneys) due to accumulation of glycogen and fat.
- Doll-like facies: Chubby cheeks and relatively thin extremities give a doll-like appearance.
- Protuberant abdomen due to enlarged organs.
- Hyperlipidemia and hyperuricemia.

- This option aligns closely with all the symptoms described.

**Option D: Lesch-Nyhan Syndrome**

- Lesch-Nyhan syndrome is an X-linked recessive disorder caused by deficiency of the enzyme hypoxanthine-guanine phosphoribosyltransferase (HGPRT).
- Clinical features include:
  - Self-mutilating behaviors (e.g., biting fingers or lips)
  - Neurological symptoms (spasticity, dystonia)
  - Hyperuricemia leading to gout and kidney stones
- Notable Absence: It does not typically present with hepatomegaly, abdominal distension, or hypoglycemia.

**Conclusion**

Considering the clinical presentation of abdominal distension, hepatomegaly, doll-like facies, and recurrent hypoglycemia, Von Gierke disease is the most consistent diagnosis. The features are classic for this glycogen storage disorder, making Option C the correct choice.

**Answer: Option C: Von Gierke disease**

**Reference Answer**

"exp": "Menkes disease- aka kinky hair disease - Defect in copper binding ATPase - Involves CNS-intellectual disability, seizures - Characterized by sparse hypopigmented kinky hairs. - Trichorrhexis nodosa and pili torti seen Von Gierke disease is glycogen storage disease type I, in which the child has doll like facies, abdominal distension, hepatomegaly and recurrent episodes of hypoglycemia. On investigation, there is hyperlipidemia, hyperuricemia and lactic acidosis. Lesch nyhan syndrome-mental retardation, microcephaly, seizures and gout.",  
"cop":3

Figure 69: **Medical Knowledge Question Answer: Case 2.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

---

## 4.27 Art Education

This study evaluates the performance of o1-preview in art education by comparing its responses to those of human experts. Although six diverse tasks were tested, detailed explanations are presented for these two key example tasks—explaining the concept of *currere* in education [134] and designing a cardboard assemblage art activity for children [109]—to assess the model’s ability to engage with both theoretical educational frameworks and practical curriculum design. The analysis concentrated on creativity, practical application, and depth of theoretical understanding.

In the task requiring an explanation of *currere* in education, o1-preview provided a well-organized and thoughtful response. It accurately described the four phases of *currere*—regressive, progressive, analytical, and synthetic—and elaborated on how these stages enable individuals to reflect on their educational experiences. The model emphasized the importance of personal reflection, memory, and synthesizing past and future learning experiences, closely aligning with William Pinar’s original work on *currere* [134]. This explanation showcased o1-preview’s capacity to engage with theoretical concepts and communicate their educational relevance clearly. As illustrated in Figure 70.

When tasked with designing a cardboard assemblage art activity, o1-preview demonstrated strong capabilities in creating a structured, step-by-step curriculum. It suggested a variety of materials, offered instructions that encouraged creativity and collaboration, and outlined clear learning objectives. However, when compared to Dr. Penfold’s approach—Harvard University’s Art Education Coordinator—o1-preview’s lesson appeared more rigid and structured, lacking the flexibility and adaptive scaffolding essential for fostering children’s open-ended exploration. Dr. Penfold emphasizes allowing children to freely explore materials, introducing techniques when appropriate to encourage spontaneity and imaginative problem-solving, both of which are essential to child-centered art education [109]. The more standardized curriculum model proposed by o1-preview highlights a gap between its procedural creativity and the responsive, individualized approach that human educators like Dr. Penfold implement. As illustrated in Figure 71.

In conclusion, while o1-preview excels in both practical curriculum planning and theoretical explanation, it lacks the depth of emotional insight and contextual sensitivity that human educators bring to art education. The model’s versatility in handling both practical and theoretical aspects is evident, yet it falls short in adapting to children’s varying needs in real-time and fostering the kind of open-ended exploration advocated by educators like Dr. Penfold. This highlights a key difference between AI models and human educators—o1-preview can be a valuable tool for generating ideas and structuring content, but human educators are essential for creating flexible, responsive learning environments that nurture children’s creativity in a more personal and adaptive manner.

**Prompt**

Can you explain the concept of *currere* in education?

**O1-preview**

The concept of *currere* in education is a transformative approach to understanding curriculum and personal educational experience. At its core, *currere* is about the "lived experiences" of individuals within educational contexts. Unlike traditional views of curriculum as a static set of courses or content to be delivered, *currere* focuses on the dynamic and evolving relationship between the individual and the curriculum. It views education as a personal and ongoing process rather than a predetermined path.

**Four-step method for engaging with *currere*:**

Pinar outlined a four-step method for engaging with *currere*, designed to facilitate deep reflection and personal growth:

**1. "Regressive Phase":**

- \*\*Description\*\*: Reflecting on past educational experiences.
- \*\*Purpose\*\*: To revisit and explore memories that have shaped one's learning and identity.
- \*\*Activity\*\*: Individuals recall and write about significant moments, emotions, and influences from their educational history.

**2. "Progressive Phase":**

- \*\*Description\*\*: Envisioning future possibilities in education.
- \*\*Purpose\*\*: To imagine potential futures and aspirations, considering how current actions impact these possibilities.
- \*\*Activity\*\*: Individuals project themselves into the future, outlining goals, hopes, and anticipated challenges.

**3. "Analytical Phase":**

- \*\*Description\*\*: Analyzing the interplay between past, present, and future.
- \*\*Purpose\*\*: To understand how past experiences and future aspirations influence current perceptions and actions.
- \*\*Activity\*\*: Critical examination of patterns, themes, and contradictions in one's educational narrative.

**4. "Synthetic Phase":**

- \*\*Description\*\*: Integrating insights to live more consciously in the present.
- \*\*Purpose\*\*: To achieve a cohesive understanding of oneself within the educational landscape.
- \*\*Activity\*\*: Synthesizing reflections to inform present decisions and actions, fostering self-awareness and agency.

**Reference Answer**

In William Pinar's conceptualization, *currere* transforms the curriculum into a dynamic, multifaceted conversation between individuals, society, history, and the self. It highlights education as an ongoing process of self-formation, where academic knowledge intersects with personal life history. Through the four steps—regressive, progressive, analytic, and synthetical—*currere* invites students to engage deeply with their past, envision future possibilities, critically analyze their present, and synthesize these elements into a lived, socially engaged understanding.

Rather than seeing education as the mastery of standardized curricula or tests, *currere* emphasizes subjective engagement, intellectual judgment, and the interplay between one's personal development and social participation. The aim is not just to accumulate knowledge but to understand oneself and the world more profoundly, using academic study as a vehicle for personal and social transformation.

This approach critiques the rigid, outcomes-based model of education, arguing that true learning must involve subjective, intellectual engagement with the curriculum, allowing individuals to explore their identities and their roles in society. In this way, *currere* is both a method of self-study and a form of cultural criticism, providing a path for educators and students alike to become more active, reflective participants in the world around them.

Figure 70: **Art Education: Case 1.** A example illustrating the capability of o1-preview to explain the concept of *currere* [134].

### **Prompt**

For cultivating children's creativity through art, can you design an art class/activity for children by guiding them to make a cardboard assemblage?

### **O1-preview**

\*\*Title:\*\* Cardboard Creations: A Journey into Imagination

\*\*Objective:\*\*

To cultivate children's creativity by guiding them through the process of designing and constructing a cardboard assemblage, encouraging them to think imaginatively, work collaboratively, and explore various artistic techniques.

\*\*Materials Needed:\*\*

- \*\*Cardboard Boxes:\*\* Various sizes and thicknesses (collect from donations, recycling centers, or local businesses).
- \*\*Cardboard Pieces:\*\* Sheets, tubes, egg cartons, and other cardboard scraps.
- \*\*Cutting Tools:\*\* Safety scissors appropriate for children, box cutters (for teacher use or older children under supervision).
- \*\*Adhesives:\*\* Non-toxic glue sticks, white glue, masking tape, duct tape.
- \*\*Decorative Materials:\*\*
  - Paints: Tempera or acrylic paints in assorted colors.
  - Brushes and sponges.
  - Markers and crayons.
  - Colored paper, tissue paper, and magazines for collage.
  - Fabric scraps, yarn, buttons, beads.
- \*\*Other Supplies:\*\* Rulers, pencils, erasers, stencils.
- \*\*Safety Equipment:\*\* Cutting mats, gloves, aprons, protective eyewear (if necessary).

\*\*Class Structure:\*\*

1. \*\*Introduction (15 minutes):\*\*

- \*\*Welcome the Students:\*\*
- Briefly discuss the concept of assemblage art — combining various elements to create a new whole.
  - Show examples of famous assemblage artists (e.g., Louise Nevelson, Joseph Cornell) to inspire students.
  - \*\*Set the Theme:\*\*
  - Offer a broad theme like "My Dream World," "Fantastic Creatures," or "Future Cities," or allow them to choose their own themes.

2. \*\*Brainstorming Session (20 minutes):\*\*

- \*\*Idea Generation:\*\*
- Encourage students to sketch their ideas or write down words and concepts related to their chosen theme.
  - Prompt them with open-ended questions:
  - "What shapes do you imagine?"
  - "How can you represent your favorite place or creature using cardboard?"
  - "What textures and layers can you create?"
- \*\*Group Discussion:\*\*
- Allow students to share their ideas with a partner or small group to foster collaboration.

3. \*\*Planning the Assemblage (15 minutes):\*\*

- \*\*Design Blueprint:\*\*
- Have students create a simple plan or outline of their assemblage, deciding which cardboard pieces they'll use and how they'll assemble them.
- Encourage experimentation — the plan doesn't have to be final or perfect.

4. \*\*Construction Phase (60 minutes):\*\*

- \*\*Building the Structure:\*\*
- Guide students in cutting and shaping the cardboard pieces safely.
- Demonstrate how to attach pieces using different adhesives and tapes.
- Encourage layering, stacking, and combining various shapes and sizes.
- \*\*Emphasize Creativity:\*\*
- Remind students there is no right or wrong way to create their assemblage.
- Suggest they think about dimensions — their work can be flat, relief, or fully three-dimensional.

5. **Break (10 minutes):**  
   - Provide a short break for students to rest, clean up scraps, and prepare for decorating.
6. **Decoration and Detailing (45 minutes):**  
   - **Adding Color and Texture:**  
     - Invite students to paint their assemblage, mix colors, and create patterns.  
     - Use decorative materials to add texture — collage with paper, attach fabric, or incorporate found objects.  
   - **Expression and Personalization:**  
     - Encourage students to include elements that reflect their personality or interests.  
     - Offer techniques like stamping, stenciling, or sponging for varied effects.
7. **Reflection and Sharing (15 minutes):**  
   - **Art Walk:**  
     - Display all the artworks in a gallery-style setup.  
   - **Presentation:**  
     - Give each student an opportunity to present their assemblage.  
     - Prompt them with questions:  
       - "What inspired your creation?"  
       - "What challenges did you overcome?"  
       - "What is your favorite part of your artwork?"
8. **Cleanup (10 minutes):**  
   - Engage students in cleaning up the workspace, teaching responsibility and care for materials.

**Learning Outcomes:**

- **Artistic Skills:**  
 - Understanding of assemblage as an art form.  
 - Development of fine motor skills through cutting, assembling, and decorating.
- **Creative Thinking:**  
 - Ability to generate original ideas and bring them to fruition.  
 - Exploration of personal expression and storytelling through art.
- **Problem-Solving:**  
 - Overcoming construction challenges.  
 - Making decisions on materials and techniques.
- **Collaboration and Communication:**  
 - Sharing ideas and learning from peers.  
 - Articulating thoughts and choices during presentations.

**Conclusion:**

This cardboard assemblage activity provides a tactile and immersive experience for children to explore their creativity. By transforming everyday materials into works of art, students learn to see possibilities in the mundane and gain confidence in their artistic abilities. The process emphasizes imagination, innovation, and the joy of creation, laying a foundation for a lifelong appreciation of art.

**Reference Answer**

Dr. Louisa Penfold, a Lecturer on Education at Harvard University, provides an insightful example of how cardboard can be used as a medium for art education, as discussed on her website: <https://louisapenfold.com/cardboard-art/>

Figure 71: **Art Education: Case 2.** A example illustrating the capability of o1-preview to design creative in-class activities for children.

[109]

---

## 4.28 Content Summarization

For the content summarization task, we designed the same prompt for all the news articles to evaluate the performance of the o1-preview. The prompt, “Please summarize the content in one sentence: . . .”, highlights the length of the summary and limits the results to a sentence. This test aims to evaluate whether the o1-preview model can summarize the textual content with the ability of context relevance, text content understanding and text generation. We compare the summary of the o1-preview with the manual summary for the corresponding article from three perspectives: accuracy, degree of abstraction and length.

Summaries generated by o1-preview show high similarity with a reference summary and original content in some samples, although they have different expressions with the same meaning. For instance, in Figure 72, the summary of both human beings and o1-preview has the same theme, which is “a type of fasting diet can regenerate pancreas cells”.

However, the o1-preview model tends to generate more detail, compared with manually summarizing. Human beings seem to be better at and more accustomed to a higher level of abstraction of text. For instance, in Figure 73, although both of the two summaries describe “Tiger Woods missed the cut at the Farmers Insurance Open”. Human beings tend to show the reason in a processed expression, which does not directly show the scores of the competition but shows a calculated and more abstractive result, while the o1-preview model lists several details about the competition. This kind of difference is shown not only in a single sample, this is common sense. In Figure 169, o1-preview model generated too much details so that the summary is not as easy to read as manual summary.

As for the length, summaries of o1-preview are longer than those of human beings in most samples. This is because summaries of o1-preview contain more details than manual summaries.

In summary, the o1-preview model is well performed in content summarization tasks. It can meet the basic needs of most summarization tasks, but compared with human summaries, it usually contains more details and is slightly inferior to humans in the ability to abstract information.

### **Prompt**

Please summarize the content in one sentence: Restoring the function of the organ - which helps control blood sugar levels - reversed symptoms of diabetes in animal experiments. The study, published in the journal Cell, says the diet reboots the body. Experts said the findings were "potentially very exciting" as they could become a new treatment for the disease. The experiments were on mice put on a modified form of the "fasting-mimicking diet". When people go on it they spend five days on a low calorie, low protein, low carbohydrate but high unsaturated-fat diet. It resembles a vegan diet with nuts and soups, but with around 800 to 1,100 calories a day. Then they have 25 days eating what they want - so overall it mimics periods of feast and famine. Previous research has suggested it can slow the pace of ageing. But animal experiments showed the diet regenerated a special type of cell in the pancreas called a beta cell. These are the cells that detect sugar in the blood and release the hormone insulin if it gets too high. Dr Valter Longo, from the University of Southern California, said: "Our conclusion is that by pushing the mice into an extreme state and then bringing them back - by starving them and then feeding them again - the cells in the pancreas are triggered to use some kind of developmental reprogramming that rebuilds the part of the organ that's no longer functioning." There were benefits in both type 1 and type 2 diabetes in the mouse experiments. Type 1 is caused by the immune system destroying beta cells and type 2 is largely caused by lifestyle and the body no longer responding to insulin. Further tests on tissue samples from people with type 1 diabetes produced similar effects. Dr Longo said: "Medically, these findings have the potential to be very important because we've shown - at least in mouse models - that you can use diet to reverse the symptoms of diabetes. "Scientifically, the findings are perhaps even more important because we've shown that you can use diet to reprogram cells without having to make any genetic alterations." BBC reporter Peter Bowes took part in a separate trial with Dr Valter Longo. He said: "During each five-day fasting cycle, when I ate about a quarter of the average person's diet, I lost between 2kg and 4kg (4.4-8.8lbs). "But before the next cycle came round, 25 days of eating normally had returned me almost to my original weight. "But not all consequences of the diet faded so quickly." His blood pressure was lower as was a hormone called IGF-1, which is linked to some cancers. He said: "The very small meals I was given during the five-day fast were far from gourmet cooking, but I was glad to have something to eat" Peter Bowes: Fasting for science Peter Bowes: Intermittent fasting and the good things it did to my body Separate trials of the diet in people have been shown to improve blood sugar levels. The latest findings help to explain why. However, Dr Longo said people should not rush off and crash diet. He told the BBC: "It boils down to do not try this at home, this is so much more sophisticated than people realise." He said people could "get into trouble" with their health if it was done without medical guidance. Dr Emily Burns, research communications manager at Diabetes UK, said: "This is potentially very exciting news, but we need to see if the results hold true in humans before we'll know more about what it means for people with diabetes. "People with type-1 and type-2 diabetes would benefit immensely from treatments that can repair or regenerate insulin-producing cells in the pancreas." Follow James on Twitter.

### **O1-preview**

A fasting-mimicking diet that regenerates insulin-producing pancreatic beta cells has reversed diabetes symptoms in mice, suggesting potential new treatments for diabetes in humans.

### **Reference Answer**

The pancreas can be triggered to regenerate itself through a type of fasting diet, say US researchers.

Figure 72: **Content Summarization: Case 1.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

### **Prompt**

Please summarize the content in one sentence: The 14-time major champion, playing in his first full PGA Tour event for almost 18 months, carded a level-par second round of 72, but missed the cut by four shots after his first-round 76. World number one Jason Day and US Open champion Dustin Johnson also missed the cut at Torrey Pines in San Diego. Overnight leader Rose carded a one-under 71 to put him on eight under. Canada's Adam Hadwin and USA's Brandt Snedeker are tied in second on seven under, while US PGA champion Jimmy Walker missed the cut as he finished on three over. Woods is playing in just his second tournament since 15 months out with a back injury. "It's frustrating not being able to have a chance to win the tournament," said the 41-year-old, who won his last major, the US Open, at the same course in 2008. "Overall today was a lot better than yesterday. I hit it better, I putted well again. I hit a lot of beautiful putts that didn't go in, but I hit it much better today, which was nice." Scotland's Martin Laird and England's Paul Casey are both on two under, while Ireland's Shane Lowry is on level par.

### **O1-preview**

In his first full PGA Tour event in nearly 18 months, Tiger Woods missed the cut at Torrey Pines by four shots after rounds of 76 and 72, while Justin Rose leads the tournament at eight under par.

### **Reference Answer**

Tiger Woods missed the cut at the Farmers Insurance Open, as England's Justin Rose maintained a one-shot lead.

Figure 73: **Content Summarization: Case 2.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

---

## 5 Discussion

Our comprehensive evaluation of o1-preview reveals both impressive capabilities and notable limitations across various domains. In coding tasks, the model demonstrated strong performance in problem-solving and algorithm implementation, while struggling with optimizing solutions for complex problems under time constraints. This suggests a need for improved efficiency in the model’s reasoning process for time-sensitive applications.

For mathematical reasoning, o1-preview excelled in solving high school-level problems but showed inconsistent performance in more advanced college-level mathematics. This indicates a potential gap in the model’s abilities to understand logical principles and handle long reasoning processes, highlighting an area for future development.

The model’s performance in scientific reasoning tasks was generally strong, particularly in integrating knowledge across different scientific disciplines. At times, it provided overly verbose explanations or failed to prioritize the most relevant information, suggesting a need for improved conciseness and information filtering mechanisms.

In language-related tasks, o1-preview demonstrated impressive multilingual capabilities and strong performance in natural language inference. Its performance in low-resource language translation revealed limitations in handling languages with limited training data. Future work should focus on improving the model’s ability to generalize linguistic patterns across diverse language families.

For medical and healthcare-related tasks, the model showed promise in generating coherent medical reports and analyzing electronic health records. Occasional production of inaccurate or inconsistent diagnoses underscores the critical need for further refinement and extensive validation before any potential clinical application.

In creative and analytical tasks such as geology reasoning and financial analysis, o1-preview demonstrated a remarkable ability to understand complex requirements and generate detailed outputs. The quality and practicality of these outputs varied, suggesting a need for better grounding in real-world constraints and domain-specific knowledge.

Future research directions should focus on:

- Enhancing the model’s efficiency in complex problem-solving under time constraints.
- Improving performance in advanced mathematical reasoning and abstract concept manipulation.
- Developing better mechanisms for concise and relevant information extraction and presentation.
- Expanding capabilities in low-resource languages and improving cross-lingual generalization.
- Refining domain-specific knowledge integration, particularly in specialized fields such as medicine, engineering and other sciences that require domain expertise [91, 70, 172, 105, 49, 171, 38, 151, 206, 37, 106, 99, 196, 162, 214, 192, 150, 50, 161, 28, 104, 16].
- Improving the model’s ability to handle multi-step reasoning tasks consistently across various domains, including sciences, engineering, medicine, public health, pharmacy, arts, humanities, education, agriculture, business, law, politics, and social studies. These domains can benefit from models that can reason well [91, 215, 112, 21, 104, 96, 207, 20, 51, 88, 87, 183, 101, 149, 102, 103, 181, 108, 98, 82].

---

## 5.1 LLM Agents and o1: Advancing Problem-Solving in Science and Engineering

Large language model agents are designed to handle complex tasks requiring sequential reasoning, planning, and memory. These agents typically consist of four key components: a core language model (the "brain"), a planning mechanism, memory systems, and the ability to use external tools [184]. By integrating these elements, LLM agents can break down complex problems, formulate strategies, and adapt their approaches based on feedback and new information.

o1 potentially changes the application potential of LLM agents. The combination of LLM agents with o1-preview's advanced capabilities presents exciting possibilities for tackling intricate science and engineering challenges. o1-preview has already demonstrated remarkable proficiency across various domains, from coding and mathematical reasoning to scientific analysis and creative problem-solving. When enhanced with the structured approach of LLM agents, o1 could potentially address even more complex, multi-step problems in scientific research and engineering design.

For instance, in scientific research, an o1-powered LLM agent could systematically approach experimental design by breaking down the process into subtasks such as literature review, hypothesis formation, methodology planning, and data analysis strategy. The agent's planning capabilities could help in outlining long-term research projects, while its memory systems could maintain coherence across extended investigations, recalling relevant past experiments or related findings. In engineering, such an agent could assist in complex system design by managing interconnected components, predicting potential issues, and suggesting optimizations. The tool use capability of LLM agents could allow o1 to interface with specialized engineering software, running simulations or accessing specific databases as needed during the design process.

Moreover, the self-reflection and improvement capabilities of LLM agents could enhance o1's already impressive reasoning skills, allowing for continuous refinement of solutions in iterative design processes or long-term research projects. This could be particularly valuable in fields like materials science or drug discovery, where vast parameter spaces need to be explored systematically.

However, it's important to note that challenges remain. The current limitations of LLM agents, such as context constraints and difficulties with long-term planning, would need to be addressed. Additionally, ensuring the reliability and consistency of outputs in critical scientific and engineering applications would be paramount.

---

## 6 Conclusion

This study provides a comprehensive evaluation of OpenAI’s o1-preview model across a diverse range of complex reasoning tasks. Our findings demonstrate that o1-preview represents a significant advancement in artificial intelligence, showing impressive capabilities in areas such as coding, scientific reasoning, language processing, and creative problem-solving. The model’s ability to integrate knowledge across different domains and apply it to novel problems is particularly noteworthy, suggesting progress towards more general artificial intelligence.

Despite these advancements, our study also uncovers important limitations and areas for improvement, particularly in advanced mathematical reasoning, handling of specialized domain knowledge, and consistent performance across all task types.

In conclusion, o1-preview is an unprecedented model that surpasses past landmarks in machine intelligence. This comprehensive study not only confirms its capabilities but also identifies critical pathways for future improvements toward AGI.

## A Appendix

### A.1 Code Generation

| Prompt  | O1-preview   |
|---|--|
| <p>You are given an array of integers start and an integer d, representing n intervals <math>[start[i], start[i] + d]</math>. You are asked to choose n integers where the ith integer must belong to the ith interval. The score of the chosen integers is defined as the minimum absolute difference between any two integers that have been chosen. Return the maximum possible score of the chosen integers.</p> <p>Example 1:<br/>Input: start = [6,0,3], d = 2<br/>Output: 4<br/>Explanation:<br/>The maximum possible score can be obtained by choosing integers: 8, 0, and 4. The score of these chosen integers is <math>\min( 8 - 0 ,  8 - 4 ,  0 - 4 )</math> which equals 4.</p> <p>Example 2:<br/>Input: start = [2,6,13,13], d = 5<br/>Output: 5<br/>Explanation:<br/>The maximum possible score can be obtained by choosing integers: 2, 7, 13, and 18. The score of these chosen integers is <math>\min( 2 - 7 ,  2 - 13 ,  2 - 18 ,  7 - 13 ,  7 - 18 ,  13 - 18 )</math> which equals 5.</p> <p>Constraints:<br/><math>2 \leq start.length \leq 105</math><br/><math>0 \leq start[i] \leq 109</math><br/><math>0 \leq d \leq 109</math></p> | <p><b>O1-preview</b></p> <pre>class Solution(object):     def maximumScore(self, start, d):         """         :type start: List[int]         :type d: int         :rtype: int         """          n = len(start)         intervals = [(s, s + d) for s in start]         intervals.sort()          def is_possible(candidate_min_diff):             pos = -1 &lt;&lt; 60 # Initialize to a very small number             for interval in intervals:                 start_i, end_i = interval                 # Assign the earliest possible position within the interval                 position = max(pos + candidate_min_diff, start_i)                 if position &gt; end_i:                     return False                 pos = position             return True          # Binary search for the maximum minimum difference         low, high = 0, 1 &lt;&lt; 60 # Set high to a large number         while low &lt; high:             mid = (low + high + 1) // 2             if is_possible(mid):                 low = mid             else:                 high = mid - 1         return low</pre> |

Figure 74: **Coding: Case 3.** A demonstration case from coding testing of a medium level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## Prompt

You are given an integer array nums of length n.

Your goal is to start at index 0 and reach index n - 1. You can only jump to indices greater than your current index.

The score for a jump from index i to index j is calculated as  $(j - i) * \text{nums}[i]$ .

Return the maximum possible total score by the time you reach the last index.

Example 1:

Input: nums = [1,3,1,5]

Output: 7

Explanation:

First, jump to index 1 and then jump to the last index. The final score is  $1 * 1 + 2 * 3 = 7$ .

Example 2:

Input: nums = [4,3,1,3,2]

Output: 16

Explanation:

Jump directly to the last index. The final score is  $4 * 4 = 16$ .

Constraints:

$1 \leq \text{nums.length} \leq 105$

$1 \leq \text{nums}[i] \leq 105$

## O1-preview

```
class Solution:
    def findMaximumScore(self, nums):
        n = len(nums)
        dp = [0] * n
        stack = []

        for i in range(n - 1, -1, -1):
            # Remove indices with nums smaller or equal to nums[i]
            while stack and nums[i] >= nums[stack[-1]]:
                stack.pop()
            if not stack:
                # Jump directly to the end
                dp[i] = (n - 1 - i) * nums[i]
            else:
                # Jump to the next greater element
                next_idx = stack[-1]
                dp[i] = (next_idx - i) * nums[i] + dp[next_idx]
            # Push the current index onto the stack
            stack.append(i)
        # The answer is dp[0]
        return dp[0]
```

Figure 75: **Coding: Case 4.** An additional case from coding testing of a medium level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## Prompt

There is a  $50 \times 50$  chessboard with one knight and some pawns on it. You are given two integers  $kx$  and  $ky$  where  $(kx, ky)$  denotes the position of the knight, and a 2D array  $\text{positions}$  where  $\text{positions}[i] = [xi, yi]$  denotes the position of the pawns on the chessboard. Alice and Bob play a turn-based game, where Alice goes first. In each player's turn:

The player selects a pawn that still exists on the board and captures it with the knight in the fewest possible moves. Note that the player can select any pawn, it might not be one that can be captured in the least number of moves.

In the process of capturing the selected pawn, the knight may pass other pawns without capturing them. Only the selected pawn can be captured in this turn. Alice is trying to maximize the sum of the number of moves made by both players until there are no more pawns on the board, whereas Bob tries to minimize them.

Return the maximum total number of moves made during the game that Alice can achieve, assuming both players play optimally.

Note that in one move, a chess knight has eight possible positions it can move to, as illustrated below. Each move is two cells in a cardinal direction, then one cell in an orthogonal direction.

**Example 1:**

Input:  $kx = 1$ ,  $ky = 1$ ,  $\text{positions} = [[0, 0]]$

Output: 4

Explanation:

The knight takes 4 moves to reach the pawn at  $(0, 0)$ .

**Example 2:**

Input:  $kx = 0$ ,  $ky = 2$ ,  $\text{positions} = [[1, 1], [2, 2], [3, 3]]$

Output: 8

Explanation:

Alice picks the pawn at  $(2, 2)$  and captures it in two moves:  $(0, 2) \rightarrow (1, 4) \rightarrow (2, 2)$ .

Bob picks the pawn at  $(3, 3)$  and captures it in two moves:  $(2, 2) \rightarrow (4, 1) \rightarrow (3, 3)$ .

Alice picks the pawn at  $(1, 1)$  and captures it in four moves:  $(3, 3) \rightarrow (4, 1) \rightarrow (2, 2) \rightarrow (0, 3) \rightarrow (1, 1)$ .

**Constraints:**

$0 \leq kx, ky \leq 49$

$1 \leq \text{positions.length} \leq 15$

$\text{positions}[i].length == 2$

$0 \leq \text{positions}[i][0], \text{positions}[i][1] \leq 49$

All  $\text{positions}[i]$  are unique.

The input is generated such that  $\text{positions}[i] \neq [kx, ky]$  for all  $0 \leq i < \text{positions.length}$ .

## O1-preview

```
class Solution:
    from collections import deque
    def maxMoves(self, kx, ky, positions):
        # Precompute distances between all positions
        positions = [[(kx, ky)]] + positions # Include knight's starting position
        n = len(positions) - 1 # Number of pawns
        dist = [[0] * (n + 1) for _ in range(n + 1)]

        board_size = 50
        knight_moves = [(-2, -1), (-2, 1), (-1, -2), (-1, 2),
                        (1, -2), (1, 2), (2, -1), (2, 1)]

        # Compute distances between all positions using BFS
        for idx in range(n + 1):
            sx, sy = positions[idx]
            visited = [[-1] * board_size for _ in range(board_size)]
            queue = deque()
            queue.append((sx, sy))
            visited[sx][sy] = 0
            while queue:
                x, y = queue.popleft()
                for dx, dy in knight_moves:
                    nx, ny = x + dx, y + dy
                    if 0 <= nx < board_size and 0 <= ny < board_size and
                       visited[nx][ny] == -1:
                        visited[nx][ny] = visited[x][y] + 1
                        queue.append((nx, ny))

        # Record distances to other positions
        for jdx in range(n + 1):
            tx, ty = positions[jdx]
            dist[idx][jdx] = visited[tx][ty]

        memo = {}
        def dp(mask, pos, turn):
            key = (mask, pos, turn)
            if key in memo:
                return memo[key]
            if mask == 0:
                return 0
            if turn == 0: # Alice's turn (maximize)
                max_moves = -float('inf')
                for i in range(1, n + 1):
                    if mask & (1 << (i - 1)):
                        new_mask = mask & ~(1 << (i - 1))
                        moves = dist[pos][i] + dp(new_mask, i, 1)
                        max_moves = max(max_moves, moves)
                memo[key] = max_moves
                return max_moves
            else: # Bob's turn (minimize)
                min_moves = float('inf')
                for i in range(1, n + 1):
                    if mask & (1 << (i - 1)):
                        new_mask = mask & ~(1 << (i - 1))
                        moves = dist[pos][i] + dp(new_mask, i, 0)
                        min_moves = min(min_moves, moves)
                memo[key] = min_moves
                return min_moves
        # Initial call: mask with all pawns, starting position index 0, Alice's turn (0)
        full_mask = (1 << n) - 1
        return dp(full_mask, 0, 0)
```

Figure 76: **Coding: Case 5.** A demonstration case from coding testing of a hard level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## Prompt

You are given two strings, coordinate1 and coordinate2, representing the coordinates of a square on an 8 x 8 chessboard.

Return true if these two squares have the same color and false otherwise.

The coordinate will always represent a valid chessboard square. The coordinate will always have the letter first (indicating its column), and the number second (indicating its row).

Example 1:

Input: coordinate1 = "a1", coordinate2 = "c3"

Output: true

Explanation:

Both squares are black.

Example 2:

Input: coordinate1 = "a1", coordinate2 = "h3"

Output: false

Explanation:

Square "a1" is black and "h3" is white.

Constraints:

coordinate1.length == coordinate2.length == 2

'a' <= coordinate1[0], coordinate2[0] <= 'h'

'1' <= coordinate1[1], coordinate2[1] <= '8'

## O1-preview

class Solution:

```
def checkTwoChessboards(self, coordinate1: str, coordinate2: str) -> bool:  
    """  
    :type coordinate1: str  
    :type coordinate2: str  
    :rtype: bool  
    """  
  
    # Helper function to determine the color parity of a square  
    def get_color_parity(coordinate):  
        # Map the column letter to a number (a-h to 1-8)  
        col = ord(coordinate[0]) - ord('a') + 1  
        # Convert the row number from string to integer  
        row = int(coordinate[1])  
        # Calculate the parity  
        return (col + row) % 2  
  
    # Get the parity for both coordinates  
    parity1 = get_color_parity(coordinate1)  
    parity2 = get_color_parity(coordinate2)  
  
    # Squares have the same color if their parities are equal  
    return parity1 == parity2
```

Figure 77: **Coding: Case 6.** A demonstration case from coding testing of a easy level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

| Prompt   | O1-preview  |
|--|---|
| <p>There is an infinite 2D plane.</p> <p>You are given a positive integer <math>k</math>. You are also given a 2D array <math>\text{queries}</math>, which contains the following queries:</p> <p><math>\text{queries}[i] = [x, y]</math>: Build an obstacle at coordinate <math>(x, y)</math> in the plane. It is guaranteed that there is no obstacle at this coordinate when this query is made.</p> <p>After each query, you need to find the distance of the <math>k</math>th nearest obstacle from the origin.</p> <p>Return an integer array <math>\text{results}</math> where <math>\text{results}[i]</math> denotes the <math>k</math>th nearest obstacle after query <math>i</math>, or <math>\text{results}[i] == -1</math> if there are less than <math>k</math> obstacles.</p> <p>Note that initially there are no obstacles anywhere. The distance of an obstacle at coordinate <math>(x, y)</math> from the origin is given by <math> x  +  y </math>.</p> <p><b>Example 1:</b><br/> Input: <math>\text{queries} = [[1,2],[3,4],[2,3],[-3,0]]</math>, <math>k = 2</math><br/> Output: <math>[-1,7,5,3]</math><br/> Explanation:<br/> Initially, there are 0 obstacles.<br/> After <math>\text{queries}[0]</math>, there are less than 2 obstacles.<br/> After <math>\text{queries}[1]</math>, there are obstacles at distances 3 and 7.<br/> After <math>\text{queries}[2]</math>, there are obstacles at distances 3, 5, and 7.<br/> After <math>\text{queries}[3]</math>, there are obstacles at distances 3, 5, and 7.</p> <p><b>Example 2:</b><br/> Input: <math>\text{queries} = [[5,5],[4,4],[3,3]]</math>, <math>k = 1</math><br/> Output: <math>[10,8,6]</math><br/> Explanation:<br/> After <math>\text{queries}[0]</math>, there is an obstacle at distance 10.<br/> After <math>\text{queries}[1]</math>, there are obstacles at distances 8 and 10.<br/> After <math>\text{queries}[2]</math>, there are obstacles at distances 6, 8, and 10.</p> <p><b>Constraints:</b><br/> <math>1 \leq \text{queries.length} \leq 2 * 10^5</math><br/> All <math>\text{queries}[i]</math> are unique.<br/> <math>-10^9 \leq \text{queries}[i][0], \text{queries}[i][1] \leq 10^9</math><br/> <math>1 \leq k \leq 105</math></p> | <pre>import heapq  class Solution:     def resultsArray(self, queries: List[List[int]], k: int) -&gt; List[int]:         """         :type queries: List[List[int]]         :type k: int         :type: List[int]         """          # Initialize a max-heap using a min-heap with negated distances         max_heap = []         results = []          for x, y in queries:             # Calculate the Manhattan distance of the obstacle from the origin             distance = abs(x) + abs(y)              if len(max_heap) &lt; k:                 # If the heap has less than k elements, push the negative distance                 heapq.heappush(max_heap, -distance)                 if len(max_heap) &lt; k:                     results.append(-1)                 else:                     # The kth nearest obstacle distance is the max element in the heap                     results.append(-max_heap[0])             else:                 # If the new distance is smaller than the largest distance in the heap                 if -max_heap[0] &gt; -distance:                     # Replace the largest distance with the new smaller distance                     heapq.heappushpop(max_heap, -distance)                     # Append the kth nearest obstacle distance                     results.append(-max_heap[0])          return results</pre> |

Figure 78: **Coding: Case 7.** A demonstration case from coding testing of a medium level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## Prompt

You are given an array `nums` of  $n$  integers, and a 2D integer array `queries` of size  $q$ , where  $\text{queries}[i] = [l_i, r_i]$ .  
For each query, you must find the maximum XOR score of any subarray of  $\text{nums}[l_i..r_i]$ .  
The XOR score of an array  $a$  is found by repeatedly applying the following operations on  $a$  so that only one element remains, that is the score:  
Simultaneously replace  $a[i]$  with  $a[i] \text{ XOR } a[i + 1]$  for all indices  $i$  except the last one.  
Remove the last element of  $a$ .  
Return an array answer of size  $q$  where  $\text{answer}[i]$  is the answer to query  $i$ .

Example 1:

Input: `nums` = [2,8,4,32,16,1], `queries` = [[0,2],[1,4],[0,5]]  
Output: [12,60,60]

Explanation:

In the first query, `nums[0..2]` has 6 subarrays [2], [8], [4], [2, 8], [8, 4], and [2, 8, 4] each with a respective XOR score of 2, 8, 4, 10, 12, and 6. The answer for the query is 12, the largest of all XOR scores.

In the second query, the subarray of `nums[1..4]` with the largest XOR score is `nums[1..4]` with a score of 60.

In the third query, the subarray of `nums[0..5]` with the largest XOR score is `nums[1..4]` with a score of 60.

Example 2:

Input: `nums` = [0,7,3,2,8,5,1], `queries` = [[0,3],[1,5],[2,4],[2,6],[5,6]]  
Output: [7,14,11,14,5]

Constraints:

```
1 <= n == nums.length <= 2000
0 <= nums[i] <= 231 - 1
1 <= q == queries.length <= 105
queries[i].length == 2
queries[i] = [l_i, r_i]
0 <= l_i <= r_i <= n - 1
```

## O1-preview

```
class Solution:
    def maximumSubarrayXor(self, nums, queries):
        n = len(nums)
        prefix_xor = [0] * (n + 1)
        for i in range(n):
            prefix_xor[i + 1] = prefix_xor[i] ^ nums[i]

        max_xor = [[0] * n for _ in range(n)]
        for i in range(n):
            current_xor = 0
            max_value = 0
            for j in range(i, n):
                current_xor ^= nums[j]
                if i == j:
                    max_value = current_xor
                else:
                    max_value = max(max_value, current_xor)
                max_xor[i][j] = max_value

        results = []
        for l, r in queries:
            results.append(max_xor[l][r])
        return results
```

Figure 79: **Coding: Case 8.** A demonstration case from coding testing of a hard level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

| Prompt  | O1-preview   |
|---|--|
| <p>You are given three positive integers num1, num2, and num3.</p> <p>The key of num1, num2, and num3 is defined as a four-digit number such that:</p> <p>Initially, if any number has less than four digits, it is padded with leading zeros.</p> <p>The ith digit (<math>1 \leq i \leq 4</math>) of the key is generated by taking the smallest digit among the ith digits of num1, num2, and num3.</p> <p>Return the key of the three numbers without leading zeros (if any).</p> <p><b>Example 1:</b><br/> Input: num1 = 1, num2 = 10, num3 = 1000<br/> Output: 0<br/> <b>Explanation:</b><br/> On padding, num1 becomes "0001", num2 becomes "0010", and num3 remains "1000".</p> <p>The 1st digit of the key is <math>\min(0, 0, 1)</math>.<br/> The 2nd digit of the key is <math>\min(0, 0, 0)</math>.<br/> The 3rd digit of the key is <math>\min(0, 1, 0)</math>.<br/> The 4th digit of the key is <math>\min(1, 0, 0)</math>.<br/> Hence, the key is "0000", i.e. 0.</p> <p><b>Example 2:</b><br/> Input: num1 = 987, num2 = 879, num3 = 798<br/> Output: 777</p> <p><b>Example 3:</b><br/> Input: num1 = 1, num2 = 2, num3 = 3<br/> Output: 1</p> | <pre>class Solution:     def generateKey(self, num1: int, num2: int, num3: int) -&gt; int:         # Pad each number with leading zeros to make them four digits         str_num1 = str(num1).zfill(4)         str_num2 = str(num2).zfill(4)         str_num3 = str(num3).zfill(4)          key_digits = []         # Iterate over each digit position         for i in range(4):             digit1 = int(str_num1[i])             digit2 = int(str_num2[i])             digit3 = int(str_num3[i])             # Find the minimum digit among the three numbers at this position             min_digit = min(digit1, digit2, digit3)             key_digits.append(str(min_digit))          # Build the key and remove leading zeros         key_str = ''.join(key_digits).lstrip('0')         # If the key is empty after stripping zeros, return 0         if not key_str:             return 0         else:             return int(key_str)</pre> |

Figure 80: **Coding: Case 9.** A demonstration case from coding testing of a easy level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## Prompt

You are given a string  $s$  of length  $n$  and an integer  $k$ , where  $n$  is a multiple of  $k$ . Your task is to hash the string  $s$  into a new string called result, which has a length of  $n / k$ .

First, divide  $s$  into  $n / k$  substrings, each with a length of  $k$ . Then, initialize result as an empty string.

For each substring in order from the beginning:

- The hash value of a character is the index of that character in the English alphabet (e.g., 'a' → 0, 'b' → 1, ..., 'z' → 25).
- Calculate the sum of all the hash values of the characters in the substring.
- Find the remainder of this sum when divided by 26, which is called hashedChar.
- Identify the character in the English lowercase alphabet that corresponds to hashedChar.
- Append that character to the end of result.
- Return result.

Example 1:

Input:  $s = "abcd"$ ,  $k = 2$

Output: "bf"

Explanation:

First substring: "ab",  $0 + 1 = 1$ ,  $1 \% 26 = 1$ ,  
 $result[0] = 'b'$ .

Second substring: "cd",  $2 + 3 = 5$ ,  $5 \% 26 = 5$ ,  
 $result[1] = 'f'$ .

Example 2:

Input:  $s = "mxz"$ ,  $k = 3$

Output: "i"

Explanation:

The only substring: "mxz",  $12 + 23 + 25 = 60$ ,  
 $60 \% 26 = 8$ ,  $result[0] = 'i'$ .

Constraints:

$1 \leq k \leq 100$

$k \leq s.length \leq 1000$

$s.length$  is divisible by  $k$ .

$s$  consists only of lowercase English letters.

## O1-preview

class Solution:

```
def stringHash(self, s: str, k: int) -> str:  
    result = ""  
    n = len(s)  
    for i in range(0, n, k):  
        substring = s[i:i+k]  
        total = sum(ord(c) - ord('a') for c in  
        substring)  
        hashedChar = total % 26  
        result += chr(ord('a') + hashedChar)  
    return result
```

Figure 81: **Coding: Case 10.** A demonstration case from coding testing of a medium level problem on o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

| Prompt  | O1-preview   |
|---|--|
| <p>You are given two positive integers n and k.</p> <p>An integer x is called k-palindromic if: x is a palindrome. x is divisible by k.</p> <p>An integer is called good if its digits can be rearranged to form a k-palindromic integer. For example, for k = 2, 2020 can be rearranged to form the k-palindromic integer 2002, whereas 1010 cannot be rearranged to form a k-palindromic integer.</p> <p>Return the count of good integers containing n digits.</p> <p>Note that any integer must not have leading zeros, neither before nor after rearrangement. For example, 1010 cannot be rearranged to form 101.</p> <p><b>Example 1:</b><br/>Input: n = 3, k = 5<br/>Output: 27<br/><b>Explanation:</b><br/>Some of the good integers are:<br/>551 because it can be rearranged to form 515.<br/>525 because it is already k-palindromic.</p> <p><b>Example 2:</b><br/>Input: n = 1, k = 4<br/>Output: 2<br/><b>Explanation:</b><br/>The two good integers are 4 and 8.</p> <p><b>Example 3:</b><br/>Input: n = 5, k = 6<br/>Output: 2468</p> <p><b>Constraints:</b><br/>1 &lt;= n &lt;= 10<br/>1 &lt;= k &lt;= 9</p> | <pre> class Solution:     def countGoodIntegers(self, n: int, k: int) -&gt; int:         from math import factorial         from collections import Counter          # Precompute factorials up to n         fact = [1] * (n + 1)         for i in range(1, n + 1):             fact[i] = fact[i - 1] * i          total_count = 0         half_length = (n + 1) // 2          # Calculate start and end for generating palindromic halves         start = 10 ** (half_length - 1) if half_length &gt; 1 else 1         end = 10 ** half_length          for i in range(start, end):             str_i = str(i)             # Generate palindrome             if n % 2 == 0:                 palindrome_str = str_i + str_i[::-1]             else:                 palindrome_str = str_i + str_i[-2::-1]              palindrome = int(palindrome_str)              # Check if palindrome is divisible by k             if palindrome % k != 0:                 continue              # Count digits             digit_counts = Counter(int(d) for d in palindrome_str)              # Total permutations             total_permutations = fact[n]             for count in digit_counts.values():                 total_permutations //= fact[count]              # Calculate permutations starting with zero             if 0 in digit_counts:                 counts_with_zero_fixed = digit_counts.copy()                 counts_with_zero_fixed[0] -= 1                 if counts_with_zero_fixed[0] == 0:                     del counts_with_zero_fixed[0]                  permutations_with_zero = fact[n - 1]                 for count in counts_with_zero_fixed.values():                     permutations_with_zero //= fact[count]                 else:                     permutations_with_zero = 0                  # Permutations without leading zeros                 permutations_without_leading_zero = total_permutations - permutations_with_zero                  total_count += permutations_without_leading_zero          return total_count </pre> |

Figure 82: **Coding: Case 11.** A demonstration case from coding testing of a hard level problem on o1-preview solution. o1-preview fails to pass this test problem. All of three attempts generate wrong answer.

## Prompt

You are given an integer power and two integer arrays damage and health, both having length n. Bob has n enemies, where enemy i will deal Bob damage[i] points of damage per second while they are alive (i.e. health[i] > 0). Every second, after the enemies deal damage to Bob, he chooses one of the enemies that is still alive and deals power points of damage to them. Determine the minimum total amount of damage points that will be dealt to Bob before all n enemies are dead.

Example 1:

Input: power = 4, damage = [1,2,3,4], health = [4,5,6,8]  
Output: 39

Explanation:

Attack enemy 3 in the first two seconds, after which enemy 3 will go down, the number of damage points dealt to Bob is  $10 + 10 = 20$  points.  
Attack enemy 2 in the next two seconds, after which enemy 2 will go down, the number of damage points dealt to Bob is  $6 + 6 = 12$  points.  
Attack enemy 0 in the next second, after which enemy 0 will go down, the number of damage points dealt to Bob is 3 points.  
Attack enemy 1 in the next two seconds, after which enemy 1 will go down, the number of damage points dealt to Bob is  $2 + 2 = 4$  points.

Example 2:

Input: power = 1, damage = [1,1,1,1], health = [1,2,3,4]  
Output: 20

Explanation:

Attack enemy 0 in the first second, after which enemy 0 will go down, the number of damage points dealt to Bob is 4 points.  
Attack enemy 1 in the next two seconds, after which enemy 1 will go down, the number of damage points dealt to Bob is  $3 + 3 = 6$  points.  
Attack enemy 2 in the next three seconds, after which enemy 2 will go down, the number of damage points dealt to Bob is  $2 + 2 + 2 = 6$  points.  
Attack enemy 3 in the next four seconds, after which enemy 3 will go down, the number of damage points dealt to Bob is  $1 + 1 + 1 + 1 = 4$  points.

Example 3:

Input: power = 8, damage = [40], health = [59]  
Output: 320

Constraints:

$1 \leqslant \text{power} \leqslant 104$   
 $1 \leqslant n \leqslant \text{damage.length} \leqslant \text{health.length} \leqslant 105$   
 $1 \leqslant \text{damage}[i], \text{health}[i] \leqslant 104$

## O1-preview

class Solution:

```
def minDamage(self, power: int, damage: List[int], health: List[int]) -> int:  
    n = len(damage)  
    enemies = []  
    for i in range(n):  
        work_i = (health[i] + power - 1) // power  
        enemies.append((work_i, damage[i], i))
```

# Define custom comparator

```
def comparator(a, b):
```

```
    work_a, damage_a, idx_a = a  
    work_b, damage_b, idx_b = b  
    num1 = work_a * damage_b  
    num2 = work_b * damage_a  
    if num1 < num2:
```

```
        return -1
```

```
    elif num1 > num2:
```

```
        return 1
```

```
    else:
```

```
        # Tie breaker, compare indices
```

```
        if idx_a < idx_b:
```

```
            return -1
```

```
        elif idx_a > idx_b:
```

```
            return 1
```

```
        else:
```

```
            return 0
```

```
from functools import cmp_to_key
```

```
enemies.sort(key=cmp_to_key(comparator))
```

# Now process the enemies in sorted order

```
total_time = 0
```

```
total_damage = 0
```

```
for work_i, damage_i, idx in enemies:
```

```
    total_time += work_i
```

```
    total_damage += damage_i * total_time
```

```
return total_damage
```

Figure 83: **Coding: Case 12.** A demonstration case from coding testing of a hard level problem and o1-preview solution. The generated code passes the Leetcode automated judgement system for the first attempt.

## A.2 Radiology Report Generation

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. \*\* Anterior basal segment of lower lobe of right lung \*\*: A pure ground glass nodule \*\* with a size of approximately \*\* 6 x 5 millimeters \*\* is observed, with a CT value of \*\* -703 HU \*\*, located approximately \*\* 2 millimeters \*\* from the pleura.
2. \*\* Subpleural segment of the outer basal segment of the lower lobe of the right lung \*\*: A solid nodule \*\* approximately 5 millimeters \*\* is observed.
3. \*\* Bilateral lungs \*\*: Increased bronchial vascular bundles.
4. Mediastinum: No enlarged lymph nodes were observed.
5. \*\* Bronchus \*\*: All lobes and segments of the bronchi are unobstructed.
6. \*\* Chest cavity \*\*: No pleural effusion was found on both sides of the chest cavity.

### **Reference Answer**

1. Pure ground glass nodule in the anterior basal segment of the lower lobe of the right lung, classified as Lu rads type 3. It is recommended to have a follow-up examination six months later.
2. There are scattered small nodules in both lungs, classified as Lu rads type 2. It is recommended to have an annual follow-up examination.

Figure 84: **Radiology: Case 3.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. A pure ground glass nodule (CT value -627HU) approximately 4mm in the posterior basal segment of the lower lobe of the left lung is considered as a small ground glass nodule, and regular follow-up observation is recommended.
2. A cord like high-density lesion in the posterior segment of the upper lobe of the right lung, with clear edges, is considered a fibrous cord shadow.

### **Reference Answer**

1. Fibrotic changes in the posterior segment of the upper lobe of the right lung, please combine clinical and medical history.
2. A ground glass nodule in the posterior basal segment of the lower lobe of the left lung, classified as LU-RADS class 2. Please combine clinical and annual follow-up examinations.

Figure 85: **Radiology: Case 4.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. Postoperative changes in the lungs: Increased texture in both lungs, new streaks appearing in the lower right and upper left lungs with unclear boundaries, consistent with the characteristics of postoperative changes.
2. Pure ground glass nodule in the dorsal segment of the lower lobe of the left lung: The nodule size is about 8.9 × 5.9 millimeters, with little change compared to before, and the density has slightly changed (-804~-226 HU), and further follow-up observation is needed.
3. \*\* Other pulmonary nodules \*\*: all less than 5 millimeters, no significant changes observed.
4. Mediastinal and thoracic conditions: No mediastinal lymph node enlargement was observed, no obvious pleural effusion was found on both sides, and the airway was unobstructed.

### **Reference Answer**

1. Changes after resection of ground glass nodules in the anterior segment of the left upper lobe, basal segment of the right lower lung, and dorsal segment of the right lower lung, with slight inflammation in the surgical area. Please consider the clinical situation.
2. Pure ground glass nodule in the dorsal segment of the left lower lung, LU-RADS class 3, recommended for a six-month follow-up.
3. Multiple nodules in both lungs, LU-RADS class 2, suggest follow-up examination.

Figure 86: **Radiology: Case 5.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

A mixed ground glass density nodule with a size of approximately  $12 \times 11\text{mm}$  was observed near the pleura in the outer basal segment of the lower lobe of the left lung, with a solid component diameter of approximately 5mm and a CT value of approximately -599HU. Considering the possibility of early lung adenocarcinoma, further examination and evaluation are recommended.

### **Reference Answer**

1. Mixed ground glass nodules in the outer basal segment of the lower lobe of the left lung, classified as Lung rad 3. It is recommended to focus on observation and undergo a 3-month follow-up examination.
2. Lung rads2-3 nodules in both lungs, it is recommended to have a follow-up examination.
3. Aortic and coronary artery sclerosis.

Figure 87: **Radiology: Case 6.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. \*\* Mixed ground glass nodule in the outer basal segment of the lower lobe of the right lung (approximately 5.7mm × 5.4mm) \*\*: Considering the possibility of early lung adenocarcinoma, further examination or regular follow-up is recommended.
2. Thickening of bronchial vascular bundles in both lungs and patchy increased density in both lower lungs: \*\* Consider the possibility of chronic inflammation or interstitial lung disease.
3. \*\* Enlargement of the heart and thickening of the inferior pulmonary veins \*\*: Indicating abnormal heart function or pulmonary congestion, it is recommended to undergo a cardiac function assessment.

### **Reference Answer**

1. Thickening of both inferior pulmonary veins, pulmonary congestion awaiting discharge, please evaluate left heart function.
2. Chronic inflammation in the middle lobe of the right lung and the lower tongue of the left lung.
3. Mixed ground glass nodules in the outer basal segment of the lower lobe of the right lung, inflammatory? Suggest re examination. Bilateral pulmonary miliary nodules, LU-RADS class 2, benign nodules.
4. Liver S2 and 3 cysts, please combine with ultrasound examination.

Figure 88: **Radiology: Case 7.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. \*\* Mixed ground glass nodule in the posterior segment of the lower lobe of the right lung, stable in size \*\*:  
The nodule is approximately 15 × 12 millimeters in size, irregular in shape, with clear boundaries, visible lobulation, and a solid part connected to the pleura. The CT value is about 60 HU, with no significant changes compared to before. Considering sub solid pulmonary nodules, it is recommended to continue close follow-up in combination with clinical practice.
2. Multiple small nodules in the basal segment of the lower right lung, stable: Several solid small nodules, with no significant changes in size and density compared to before, may be inflammation or benign lesions. Regular follow-up is recommended.
3. \*\* No new abnormalities found \*\*: Clear bronchial vascular bundles in both lungs, no thickening of bilateral pleura, and no pleural effusion on both sides.

### **Reference Answer**

Mixed ground glass nodules and several solid small nodules in the subpleural region of the lower right lung, roughly the same as before.

Figure 89: **Radiology: Case 8.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. Multiple small ground glass nodules in both lungs, with a maximum size of approximately 5mm. Regular follow-up is recommended to rule out the possibility of early lung tumors.
2. Fibrous cord shadow in the middle lobe of the right lung, consistent with fibrotic changes.
3. Multiple cysts in the liver.
4. A 1.5cm nodule in the inner quadrant of the left breast, further examination is recommended to differentiate between benign and malignant.

### **Reference Answer**

1. Multiple ground glass nodules in both lungs, considered inflammatory nodules, suggest follow-up observation and re examination.
2. There is a small fibrotic lesion in the right middle lung.
3. Low density cystic lesions in the liver, please refer to the abdominal CT report.
4. Nodule in the inner quadrant of the left breast, ultrasound is recommended.

Figure 90: **Radiology: Case 9.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### **Prompt**

Imagine you are a chest radiologist. Your daily responsibilities include thoroughly reviewing the report section of chest CT examinations and providing concise, professional diagnoses for patients. Based on the chest CT scan report provided below, please summarize the main findings or diagnosis.

### **O1-preview**

1. Mixed ground glass nodule (15mm × 9mm, CT value -371HU) in the posterior segment of the upper lobe of the right lung, suggesting a high possibility of lung adenocarcinoma. Further examination is recommended to clarify the nature.
2. Pure ground glass nodule (13mm × 13mm, CT value -529HU) in the upper lobe of the right lung, suggesting the possibility of early lung adenocarcinoma. Follow up observation or further examination is recommended.
3. Multiple patchy ground glass opacities around both lungs need to be considered in conjunction with clinical factors such as inflammation.
4. Multiple small nodular shadows (diameter<5mm) in both lungs, with clear boundaries, indicating a high possibility of benign lesions. Regular follow-up is recommended.
5. Aortic wall calcification indicates atherosclerosis.
6. Round low-density shadow in the left kidney, considering renal cyst.

### **Reference Answer**

1. Grinding glass nodules in the posterior and apical segments of the upper lobe of the right lung, LU-RADS 3L class, suggest a follow-up examination at 4-6 weeks.
2. Multiple ground glass exudate in the upper right lung, with a high possibility of infection. It is recommended to undergo anti-inflammatory treatment and follow-up examination.
3. Multiple nodules in both lungs, LU-RADS class 2.Aortic sclerosis.
4. There is a high possibility of left kidney cyst.

Figure 91: **Radiology: Case 10.** We observed that the o1-preview closely aligns with human writing patterns, exhibiting clear organization and concise language. To enhance reader accessibility, the radiological report generated in the zero-shot o1-preview task has been translated from Chinese to English.

### A.3 Robot Command Planning

#### **Prompt**

Can you help me write the code of ROS for SLAM navigation?

#### **O1-preview**

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import rospy
from sensor_msgs.msg import LaserScan
from nav_msgs.msg import Odometry, OccupancyGrid
from geometry_msgs.msg import TransformStamped
import tf
class SimpleSLAM:
    def __init__(self):
        rospy.init_node('simple_slam', anonymous=True)
        self.laser_sub = rospy.Subscriber('/scan', LaserScan, self.laser_callback)
        self.odom_sub = rospy.Subscriber('/odom', Odometry, self.odom_callback)
        self.map_pub = rospy.Publisher('/map', OccupancyGrid, queue_size=10)
        self.tf_broadcaster = tf.TransformBroadcaster()
        self.map = OccupancyGrid()

    def laser_callback(self, data):
        pass
    def odom_callback(self, data):
        pass
    def update(self):
        self.map_pub.publish(self.map)
        self.tf_broadcaster.sendTransform(
            (0, 0, 0),
            (0, 0, 0, 1),
            rospy.Time.now(),
            "base_link",
            "map" )
if __name__ == '__main__':
    slam = SimpleSLAM()
    rate = rospy.Rate(10)
    while not rospy.is_shutdown():
        slam.update()
        rate.sleep()
```

Figure 92: **Robot Command Planning: Case 2.** o1-preview's response demonstrates a ROS implementation for Simultaneous Localization and Mapping (SLAM), integrating laser scan and odometry data to build a dynamic map.

### Prompt

Can you help me write the code of ROS for given point navigation?

### O1-preview

```
import rospy
import actionlib
from move_base_msgs.msg import MoveBaseAction, MoveBaseGoal
from tf.transformations import quaternion_from_euler
def move_to_goal(x_goal, y_goal, yaw_goal):
    rospy.init_node('navigate_to_pose', anonymous=False)
    client = actionlib.SimpleActionClient('move_base', MoveBaseAction)
    client.wait_for_server()
    goal = MoveBaseGoal()
    goal.target_pose.header.frame_id = 'map'
    goal.target_pose.header.stamp = rospy.Time.now()
    goal.target_pose.pose.position.x = x_goal
    goal.target_pose.pose.position.y = y_goal
    goal.target_pose.pose.position.z = 0.0
    quaternion = quaternion_from_euler(0.0, 0.0, yaw_goal)
    goal.target_pose.pose.orientation.x = quaternion[0]
    goal.target_pose.pose.orientation.y = quaternion[1]
    goal.target_pose.pose.orientation.z = quaternion[2]
    goal.target_pose.pose.orientation.w = quaternion[3]
    client.send_goal(goal)
    client.wait_for_result()
    if client.get_state() == actionlib.GoalStatus.SUCCEEDED:
        rospy.loginfo("Reached the destination!")
        return True
    else:
        rospy.loginfo("Failed to reach the destination.")
        return False
if __name__ == '__main__':
    try:
        x_goal = 2.0
        y_goal = 3.0
        yaw_goal = 1.57 # 朝向90度
        move_to_goal(x_goal, y_goal, yaw_goal)
        rospy.spin()
    except rospy.ROSInterruptException:
        rospy.loginfo("Navigation terminated.")
```

Figure 93: **Robot Command Planning: Case 3.** The code generated by o1-preview implements point-to-point navigation using the ROS MoveBase action server, leveraging goal coordinates and quaternion-based orientation control.

**Prompt**

Can you help me write the code of ROS for robot RL?

**O1-preview**

```
import gym
import numpy as np
import tensorflow as tf
from tensorflow import keras
from collections import deque
import random
from robot_env import RobotEnv
env = RobotEnv()
state_size = env.observation_space.shape[0]
action_size = env.action_space.n
batch_size = 64
n_episodes = 1000
output_dir = 'model_output/dqn'
if not os.path.exists(output_dir):
    os.makedirs(output_dir)
class DQNAgent:
    def __init__(self, state_size, action_size):
        self.state_size = state_size
        self.action_size = action_size
        self.memory = deque(maxlen=2000)
        self.gamma = 0.95 # 折扣率
        self.epsilon = 1.0 # 探索率
        self.epsilon_min = 0.01
        self.epsilon_decay = 0.995
        self.learning_rate = 0.001
        self.model = self._build_model()
    def _build_model(self):
        model = keras.Sequential()
        model.add(keras.layers.Dense(24, input_dim=self.state_size, activation='relu'))
        model.add(keras.layers.Dense(24, activation='relu'))
        model.add(keras.layers.Dense(self.action_size, activation='linear'))
        model.compile(loss='mse', optimizer=keras.optimizers.Adam(lr=self.learning_rate))
        return model
    def remember(self, state, action, reward, next_state, done):
        self.memory.append((state, action, reward, next_state, done))
```

Figure 94: **Robot Command Planning: Case 4.** A deep Q-learning agent for robot control is generated by o1-preview using ROS, with a neural network built in TensorFlow to optimize decision-making in a simulated environment.

### Prompt

Can you help me write the code of ROS for autonomous driving?

### O1-preview

```
class VehicleControl:  
    def __init__(self):  
        rospy.init_node('vehicle_control', anonymous=True)  
        rospy.Subscriber('/planned_path', Path, self.path_callback)  
        rospy.Subscriber('/obstacle_detected', Bool, self.obstacle_callback)  
        self.control_pub = rospy.Publisher('/ackermann_cmd', AckermannDriveStamped, queue_size=10)  
        self.current_path = None  
        self.obstacle_detected = False  
    def path_callback(self, data):  
        self.current_path = data.poses  
    def obstacle_callback(self, data):  
        self.obstacle_detected = data.data  
    def control_loop(self):  
        rate = rospy.Rate(10) # 10Hz  
        while not rospy.is_shutdown():  
            if self.current_path and not self.obstacle_detected:  
                target_pose = self.current_path[0].pose  
                drive_msg = AckermannDriveStamped()  
                drive_msg.drive.speed = 2.0  
                drive_msg.drive.steering_angle = self.compute_steering_angle(target_pose)  
                self.control_pub.publish(drive_msg)  
            else:  
                drive_msg = AckermannDriveStamped()  
                drive_msg.drive.speed = 0.0  
                drive_msg.drive.steering_angle = 0.0  
                self.control_pub.publish(drive_msg)  
            rate.sleep()  
    def compute_steering_angle(self, target_pose):  
        return 0.0  
if __name__ == '__main__':  
    try:  
        controller = VehicleControl()  
        controller.control_loop()  
    except rospy.ROSInterruptException:  
        pass
```

Figure 95: **Robot Command Planning: Case 5.** o1-preview generates a ROS-based control loop for autonomous driving, processing planned paths and obstacle detection to adjust the vehicle's speed and steering dynamically.

---

#### A.4 Nature Language Inference

Additional examples are shown in Table 8.

Table 8: Additional examples of o1-preview responses on NLI task.

| Dataset | Premise / Context  | Hypothesis / Question   | o1-preview Response  | Label          |
|---------|--|---|--|----------------|
| MNLI    | The tourist industry continued to expand, and though it became one of the top two income earners in Spain, a realization that unrestricted mass tourism was leading to damaging long-term consequences also began to grow.   | Tourist's caused damage to Spain.   | The premise indicates that mass tourism was leading to "damaging long-term consequences" in Spain. This suggests that the influx of tourists was causing harm to the country over time. Therefore, the hypothesis that "Tourists caused damage to Spain" is directly supported by the premise. This means the premise entails the hypothesis. Answer: <b>Entailment</b>  | Entailment     |
| ANLI    | Delhi Public School Megacity was established in 2004 at Kolkata belonging to the Delhi Public School Society. It is the only DPS other than Delhi Public School New-town Kolkata to be affiliated to the Council for the Indian School Certificate Education(CISCE). | Delhi Public School Megacity is the only DPS affiliated with CISCE.   | The premise clearly states that there are two DPS schools affiliated with CISCE: Delhi Public School Megacity and Delhi Public School Newtown Kolkata. The hypothesis, however, claims that Delhi Public School Megacity is the only DPS affiliated with CISCE. This directly contradicts the information in the premise, which acknowledges the existence of another affiliated school. Therefore, the relationship between the premise and the hypothesis is a contradiction. Answer: <b>Contradiction</b> | Contradiction  |
| QNLI    | In addition, "Super Bowl City" opened on January 30 at Justin Herman Plaza on The Embarcadero, featuring games and activities that will highlight the Bay Area's technology, culinary creations, and cultural diversity.   | What was the name of the event at the The Embarcadero that was held prior to Super Bowl 50 to help show off some of the things that San Francisco has to offer? | Yes, the context sentence contains the answer. The event was called "Super Bowl City."   | Entailment     |
| MedNLI  | Mother developed separation of symphysis pubis and was put in traction.  | She has orthopedic injuries.  | The premise states that the mother experienced a separation of the symphysis pubis—a joint located in the pelvis—and required traction. This condition is an injury to the musculoskeletal system, specifically involving bones and joints, which falls under the field of orthopedics. Therefore, the hypothesis that "She has orthopedic injuries" is directly supported by the information in the premise. Thus, the premise entails the hypothesis. Answer: <b>Entailment</b>                            | Entailment     |
| RadQNLI | LEFT UPPER EXTREMITY ULTRASOUND: [**Doctor Last Name **] scale and Doppler son[**Name (N) 76**] of the left internal jugular, subclavian, axillary, basilic, cephalic, and brachial veins were performed.  | Are there any clots in the deep veins of the left upper extremity?  | No, the context sentence does not contain the answer to the question. It mentions that an ultrasound was performed on specific veins of the left upper extremity but does not indicate whether any clots were found.   | Not entailment |

## A.5 Quantitative Investing

### Prompt

How does the concept of mean reversion relate to quantitative momentum trading strategies?

### Q1-preview

Result:

Mean reversion and momentum are two fundamental concepts in quantitative finance that, at first glance, may seem contradictory. Mean reversion suggests that asset prices and returns tend to move back toward their historical averages over time. In contrast, momentum trading strategies capitalize on the continuation of existing trends, assuming that assets performing well will continue to do so in the near future. Understanding the relationship between these two concepts is crucial for developing effective quantitative momentum trading strategies that can adapt to changing market conditions.

Mean Reversion in Finance

• Definition:

- Mean reversion is the financial theory suggesting that asset prices and returns eventually revert to their long-term mean or average level.
- This mean can be the historical average price, average return, or another statistical benchmark.

• Mechanism:

- Overreaction and Correction: Markets may overreact to news, causing prices to deviate from intrinsic values. Over time, prices correct themselves.
- Statistical Properties: In a mean-reverting process, extreme values are followed by movements back toward the mean.

• Applications:

- Identifying Overbought/Oversold Conditions: Traders use indicators like the Relative Strength Index (RSI) to spot potential reversals.
- Pairs Trading: Involves taking positions on two correlated assets that have diverged from their typical price relationship.

### Reference Answer

Mean reversion suggests that **overvalued or undervalued assets will eventually return to their average prices**, influencing trading decisions in **quantitative momentum strategies to exploit short-term price deviations from their long-term trends**.

Figure 96: **Quantitative Investing: Case 3.** o1-preview successfully explained the concept of mean reversion, detailing its working mechanism, and introduced its downstream tasks, demonstrating an accurate understanding of the underlying principles.

### **Prompt**

How can time series decomposition techniques such as trend, seasonality, and residual analysis aid in forecasting stock market trends?

### **O1-preview**

#### Introduction

Time series decomposition is a fundamental technique in time series analysis that breaks down a complex time series into its constituent components: trend, seasonality, and residual (or irregular) components. In the context of the stock market, decomposing stock price data or market indices helps analysts and traders understand underlying patterns, isolate different effects, and improve the accuracy of forecasting models. By separating these components, one can model and predict future market trends more effectively.

#### Components of Time Series Decomposition

##### 1.Trend Component:

1. Definition: The long-term movement or direction in the data over a period of time.
2. Significance: Reflects the underlying growth or decline in stock prices or market indices due to economic, political, or technological factors.
3. Application: Identifying the trend helps in understanding the general direction of the market and in making long-term investment decisions.

##### 2.Seasonality Component:

1. Definition: Regular, periodic fluctuations that occur at fixed periods (daily, weekly, monthly, quarterly, or annually).
2. Significance: Captures patterns such as increased trading activity at certain times of the year or recurring economic cycles.
3. Application: Recognizing seasonal effects enables traders to anticipate and capitalize on predictable market movements.

##### 3.Residual (Irregular) Component:

1. Definition: The random variation or noise remaining after removing the trend and seasonal components.
2. Significance: Represents unpredictable factors affecting stock prices, such as sudden news events or market shocks.
3. Application: Analyzing the residuals helps in assessing the effectiveness of the decomposition and in modeling the stochastic (random) part of the series using statistical models.

### **Reference Answer**

Time series decomposition techniques help break down the time series data into its fundamental components like trend, seasonality, and residual variations. By analyzing these components separately, analysts can better understand past patterns and make more accurate forecasts of stock market trends.

Figure 97: **Quantitative Investing: Case 4.** o1-preview successfully analyzed the differences among trend, seasonality, and residual analysis in stock market forecasting, and explained how they work together, demonstrating a comprehensive understanding of their combined role in predicting market movements.

---

**Prompt**

How is the stochastic oscillator used as a trading signal?

**O1-preview**

The stochastic oscillator is a momentum indicator that compares a security's closing price to its price range over a specific period, oscillating between 0 and 100:

- Above 80 (Overbought):** Suggests the asset may be overbought and due for a price decline, signaling a potential sell opportunity.
  - Below 20 (Oversold):** Indicates the asset may be oversold and poised for a price increase, signaling a potential buy opportunity.
- Traders use these overbought and oversold levels, along with crossovers of the indicator's lines, to determine optimal entry and exit points in trading.

**Reference Answer**

The stochastic oscillator is a momentum indicator **that compares a security's closing price to its price range over a specified period of time.** Traders use it to identify **overbought and oversold conditions.**

Figure 98: **Quantitative Investing: Case 5.** o1-preview successfully explained the stochastic oscillator, provided examples, and demonstrated its application, showing an accurate understanding of the underlying concept.

## A.6 Low-Resource Language Translation

### Prompt

Help me translate: ḡW dEGG.I Uω TSR.I DL?

### O1-preview

The sentence "ḡW dEGG.I Uω TSR.I DL?" translates to:

"How old are you today?"

Breakdown of the Translation:

ḡW (hila): "How many" or "What"

dEGG.I (tsugvwalodi): "Your age"

d- (tsu-): Second-person possessive prefix meaning "your."

EGG.I (gvwalodi): "Age" or "years."

Uω (saquu): "Just", "only," or "exactly."

TSR.I (igasvti): "Today"

TS (iga): "Day."

R.I (svti): Suffix indicating "this" or "present."

Combined, igasvti means "this day" or "today."

DL (ada): A particle used for emphasis or to indicate a question, similar to "now" or "?"

Interpretation:

The sentence is asking someone how many years old they are as of today, likely in the context of a birthday. It emphasizes the age "just today", making it clear that the question pertains to the person's age on this specific day.

### Reference Answer

How much does one rick of wood cost?

Figure 99: **Low-Resource Language Translation: Case 3.** This test evaluates the o1-preview model's ability to translate the Cherokee sentence. The model translates the sentence as "How old are you today?" and provides a detailed breakdown of the translation components, explaining the meaning of each Cherokee word and particle. However, the translation is totally different from the reference answer due to mistakenly translating the words in the sentence.

**Prompt**

Help me translate: IrWS Dhł ḍō'žeō V.I Dłō'bōł.

**O1-preview**

The sentence "IrWS Dhł ḍō'žeō V.I Dłō'bōł." translates to:

"The chickens are climbing the trees because of them."

Breakdown of the Translation:

IrWS (tsitaga): "chickens"

IrWS is the Cherokee word for "chickens."

Dhł (anida): "trees"

Dh- (ani-): Plural prefix meaning "multiple" or "many."

ł (da): "tree."

Combined, anida means "trees."

ḍō'žeōV.I (tsudvhnisododi): "because of them"

đ- (tsu-): Third person plural prefix meaning "their" or "them."

žeō (dvh): Root verb meaning "they" or "them."

ōV.I (sdodi): A conjunction meaning "because" or "due to."

Combined, tsudvhnisododi translates to "because of them."

Dłō'bōł (adanvteha): "they are climbing" or "they are going up."

Dł- (ada-): Third person plural prefix meaning "they."

ō'bōł (nvteha): Verb meaning "are climbing" or "are ascending."

Combined, adanvteha means "they are climbing."

Interpretation:

The sentence describes the chickens climbing the trees because of someone or something. Given the context, "them" likely refers to the foxes mentioned in a previous sentence.

Contextual Connection:

In a prior sentence, foxes were causing concern:

"IrWS DhCēō O'YIrōđS.ñ DōBAT." ("The chickens are going to their coop because of the foxes.")

Continuing the narrative, the current sentence explains that the chickens are now climbing the trees to escape or avoid the foxes.

**Reference Answer**

He's thinking about raising chicks.

Figure 100: **Low-Resource Language Translation: Case 4.** The o1-preview model correctly translates the noun chickens and accurately captures the present progressive tense. The o1-preview model also attempts to establish a contextual connection, suggesting that the chickens are reacting to foxes mentioned in a previous sentence. However, this contextual interpretation is incorrect, as the reference answer, "He's thinking about raising chicks," is the ground truth, indicating a misalignment in the overall translation.

**Prompt**

Help me translate: DfiOCD DøðSø DløSE IrA6JøV.

**O1-preview**

The sentence "DfiOCD DøðSø DløSE IrA6JøV." translates to:

"I see the man helping the crying boy."

Breakdown of the Translation:

DfiOCD (ayo utlia): "crying boy"

Dfi (ayo): "boy"

O'CD (utlia): "is crying"

Combined, DfiOCD means "the boy who is crying" or "crying boy."

DøðSø (asgaya): "man"

DløSE (adadegv): "is helping"

D- (a-): Third person singular prefix meaning "he/she"

løSE (dadegv): "helping"

Together, DløSE means "he is helping."

IrA6JøV (tsigowhtiha): "I see"

Ir- (tsi-): First person singular prefix meaning "I"

A6JøV (gowhtiha): "see" or "am seeing"

Combined, IrA6JøV means "I see."

Interpretation:

The sentence describes you observing a man who is helping a boy that is crying. It emphasizes the action of the man assisting the boy, with you witnessing the event.

**Reference Answer**

I see a lame man limping.

Figure 101: **Low-Resource Language Translation: Case 5.** The o1-preview model successfully translates the sentence into "I see a man doing something". The provided detailed breakdown of each component reveals a mismatch in the rest interpretation. While the model succeeds in grammatical analysis, the translation result does not align with the intended meaning.

## A.7 Educational Q&A

**Prompt**

What type of organism is commonly used in preparation of foods such as cheese and yogurt?

Must Choose from below options:

- viruses
- protozoa
- gymnosperms
- mesophilic organisms

**O1-preview**

**Note** of the provided options correctly identify the organisms commonly used in the preparation of cheese and yogurt. The primary organisms involved in making these dairy products are **bacteria**, specifically **lactic acid bacteria** such as *Lactobacillus* and *Streptococcus* species. These bacteria are essential for fermenting milk, which leads to the production of yogurt and various types of cheese

**Reference Answer**

**mesophilic organisms**

Figure 102: **Educational Q&A: Case 3.** o1-preview identified that none of the provided options were correct for the question regarding organisms used in the preparation of cheese and yogurt. Instead, it overthought the problem by providing an explanation about lactic acid bacteria (such as *Lactobacillus* and *Streptococcus*) being responsible for the fermentation process, despite the question's simpler intent. This case demonstrates that o1-preview may sometimes overthink and fail to select from the given options. The dataset for this task is sourced from [193].

**Prompt**

What is the least dangerous radioactive decay?

Must Choose from below options:

- zeta decay
- beta decay
- gamma decay
- alpha decay

**O1-preview**

The least dangerous radioactive decay among the given options is alpha decay.

Correct Answer: **alpha decay**

**Reference Answer**

**alpha decay**

Figure 103: **Educational Q&A: Case 4.** o1-preview accurately determined that alpha decay is the least dangerous form of radioactive decay among the provided options, showcasing its ability to interpret and correctly answer safety-related questions about radioactive materials. The dataset for this analysis is sourced from [193].

**Prompt**

Kilauea in Hawaii is the world's most continuously active volcano. very active volcanoes characteristically eject red-hot rocks and lava rather than this?

Must Choose from below options:

- magma
- greenhouse gases
- carbon and smog
- smoke and ash

**O1-preview**

The phenomenon you're describing involves the type of material that very active volcanoes like Kilauea primarily eject. Kilauea is known for its effusive eruptions, which are characterized by the steady flow of red-hot lava and rocks rather than explosive emissions.

Correct Answer: **smoke and ash**

**Reference Answer**

**smoke and ash**

Figure 104: **Educational Q&A: Case 5.** o1-preview accurately determined that smoke and ash were the correct answer in a question about volcanic eruptions. The model identified the characteristics of Kilauea's effusive eruptions, showing that it understands the difference between material ejected in explosive versus effusive volcanic activities. The dataset for this task is sourced from [193].

## A.8 Logical Reasoning

### Prompt

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

Zhao Ming, Qian Hong and Sun Jie were admitted to Peking University, Tsinghua University and Beijing Normal University. Student A guessed: Zhao Ming was admitted to Tsinghua University, Sun Jie was admitted to Beijing Normal University. Guess B: Zhao Ming was admitted to Beijing Normal University and Qian Hong was admitted to Tsinghua University. Student C guess: Zhao Ming was admitted to Peking University and Sun Jie was admitted to Tsinghua University. The students' guesses turned out to be half right.

Based on the above experiment, which of the following must be true?

- A. Zhao Ming, Qian Hong and Sun Jie were admitted to Peking University, Tsinghua University and Beijing Normal University respectively.
- B. Zhao Ming, Qian Hong and Sun Jie were admitted to Tsinghua University, Beijing Normal University and Peking University respectively.
- C. Zhao Ming, Qian Hong and Sun Jie were admitted to Beijing Normal University, Tsinghua University and Peking University respectively.
- D. Zhao Ming, Qian Hong and Sun Jie were admitted to Peking University, Beijing Normal University and Tsinghua University respectively.

### O1-preview

From the information provided, we can deduce the following:

First, let's list all possible permutations of student-university assignments:

1. Zhao Ming (Z): Peking University (P), Qian Hong (Q): Tsinghua University (T), Sun Jie (S): Beijing Normal University (B)
2. Z: P, Q: B, S: T
3. Z: T, Q: P, S: B
4. Z: T, Q: B, S: P
5. Z: B, Q: P, S: T
6. Z: B, Q: T, S: P

Next, we analyze each student's guesses:

- Student A guessed:
  - A1: Z was admitted to T - A2: S was admitted to B
- Student B guessed:
  - B1: Z was admitted to B - B2: Q was admitted to T
- Student C guessed:
  - C1: Z was admitted to P - C2: S was admitted to T

We evaluate each permutation against these guesses to see which one results in each student having exactly one correct and one incorrect guess. Permutation 1:

- Evaluations:
  - A1: False (Z is not at T)
  - A2: True (S is at B)
  - B1: False (Z is not at B)
  - B2: True (Q is at T)
  - C1: True (Z is at P)
  - C2: False (S is not at T)

Each student has one correct and one incorrect guess, satisfying the condition "the students' guesses turned out to be half right." Thus, the correct assignment is:

- Zhao Ming was admitted to \*\*Peking University\*\*
- Qian Hong was admitted to \*\*Tsinghua University\*\*
- Sun Jie was admitted to \*\*Beijing Normal University\*\*

This matches \*\*Option A\*\*. \*\*Answer:\*\* A

Figure 105: **Logical Reasoning: Case 3.** Three students, Zhao Ming, Qian Hong, and Sun Jie, were admitted to Peking University, Tsinghua University, and Beijing Normal University. Student A, B, and C made guesses, but only half of each guess was correct. After analysis, it was determined that Zhao Ming was admitted to Peking University, Qian Hong to Tsinghua University.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

A suggestion was made to the manager of a garment factory to replace large sewing balls on the machine, so that there would be no need to stop frequently to change the thread balls, which would help reduce labor costs. Which of the following four options best fits the description recommendations?

- A. Large balls of sewing thread are not as strong as small ones.
- B. The factory operates on a hourly wage system, not a piecework wage system.
- C. sewing machines do not need to be shut down regularly for maintenance and repair.
- D. The operator is not allowed to leave the machine during the working period.

### **O1-preview**

The recommendation suggests replacing small sewing balls with larger ones to reduce the frequency of stops for changing thread balls, aiming to reduce labor costs. The key to understanding which option best fits this recommendation lies in examining how changing thread balls impacts labor costs under different wage systems.

Option B states: "The factory operates on an hourly wage system, not a piecework wage system."

In an hourly wage system, workers are paid based on the time they spend working, regardless of how much they produce. Frequent stops to change thread balls mean workers spend part of their paid time not producing garments, which increases labor costs per unit of output. By using larger thread balls, the factory reduces non-productive downtime, making workers more efficient during their paid hours, and thus reducing labor costs.

In contrast, under a piecework wage system, workers are paid based on the number of items they produce. Downtime caused by changing thread balls would primarily affect workers' earnings, not the factory's labor costs per unit. Therefore, the recommendation to reduce stops would not have the same impact on labor costs in a piecework system.

Options A, C, and D are less directly related to the labor cost implications of changing thread balls. Option A introduces a potential quality issue with large thread balls but doesn't address labor costs. Option C is about machine maintenance, which is separate from the frequency of changing thread balls. Option D discusses operator movement policies but doesn't directly relate to labor cost savings from reduced downtime.

**\*\*Answer:\*\* B. The factory operates on a hourly wage system, not a piecework wage system.**

Figure 106: **Logical Reasoning: Case 4.** The analysis performance of the o1-preview model is remarkable. It accurately grasps the core purpose of the suggestion in the question stem, which is to reduce labor costs. By in-depth analysis of the relationship between each option and this purpose, it clearly points out the close connection between option B and reducing labor costs. In the analysis process, a detailed and reasonable explanation is provided for the impact of replacing sewing balls under different wage systems on labor costs.

**Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

Schools that are 300 meters away from home cannot be attended, but are arranged to go to schools 2 kilometers away. A school-age child in a city encountered such an arrangement in the district education bureau when he was in elementary school, and this arrangement is district education. The bureau made according to the school district where the children's household registration is located. According to the Principle of Enrolling Near Schools stipulated by the city education bureau, the parents of the children took the district education bureau to the court and demanded that the original arrangement be revoked to allow their children to enter school nearby.

Which of the following is most likely a reasonable basis for the court?

A."Entry nearest" is not "recently enrolled", and the linear distance between the registered place of the enrolled child and the school cannot be used as the only basis for dividing the school district.

B.According to the specific geographical factors, each elementary school in the teaching area may not be in the center of the teaching area.

C.Which school to study for children's enrollment is not to let school-age children or their parents choose independently, but to obey the administrative arrangements of the competent government departments.

D.The administrative action of the district education bureau to divide the school districts complies with the law, and the school district where the plaintiff child household registration is located really needs to go to a school 2 kilometers away from home.

**O1-preview**

Based on the situation presented, the parents are arguing that their child should attend the school that is only 300 meters away from their home, in line with the city's "Principle of Enrolling Near Schools." However, the district education bureau has assigned the child to a school 2 kilometers away based on the school district where the child's household registration (hukou) is located.

The court must determine whether the district education bureau's action aligns with the "Principle of Enrolling Near Schools" and whether the parents' claim has merit.

Option A states: \*\*"Enrolling near" is not "enrolled at the nearest school," and the linear distance between the registered place of the enrolled child and the school cannot be used as the only basis for dividing the school district.\*\*

This option directly addresses the key issue: the interpretation of the "Principle of Enrolling Near Schools." It clarifies that "enrolling near" doesn't necessarily mean the "nearest" school in terms of physical distance. The policy may consider other factors, such as school capacity, resource allocation, and administrative boundaries, rather than just the shortest distance from the child's home to the school.

Therefore, the court is likely to base its decision on the interpretation that "enrolling near" allows for administrative discretion in assigning schools and that physical proximity alone isn't the sole determinant. This means the district education bureau's assignment is reasonable and lawful under the given policy.

**Answer: A\***

Figure 107: **Logical Reasoning: Case 5.** It can closely focus on the core of the problem and deeply analyze the relationship between each option and the key issue of the case. The interpretation of option A is accurate and in place. It not only considers the specific meaning of the policy but also makes reasonable inferences based on the actual situation, providing a clear logical thinking for reaching the correct answer.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

Mobile payment is now gaining popularity in large and medium-sized cities such as Beijing and Shanghai. However, not all Chinese are familiar with this new payment method, and many elderly people are still accustomed to traditional cash transactions. Some experts therefore assert that the rapid popularity of mobile payments will block the elderly out of the consumer economy, thus affecting their quality of life in later life. Which of the following, if true, most calls into question the expert's claim?

- A. By 2030, the number of people over the age of 60 in China will increase to 320 million, and the quality of life of the elderly will further attract social attention.
- B. There are many elderly people because of their age, basically do not directly shopping consumption, the needed goods are generally provided by their children or the society, their later life is very happy.
- C. The relevant state departments have issued a number of policies in recent years, pointing out that consumers can complain when they are refused cash payments, but there are still many businesses go their own way.
- D. Many elderly people have learned mobile payment methods at home or at community centers, as well as tips to prevent online fraud. Use a language style that meets the application scenario.

### **O1-preview**

To determine which option most calls into question the expert's claim that the rapid popularity of mobile payments will block the elderly out of the consumer economy, thus affecting their quality of life, let's analyze each option:

\*\*A.\*\* \*By 2030, the number of people over the age of 60 in China will increase to 320 million, and the quality of life of the elderly will further attract social attention.\*

- This option mentions the growing elderly population and increased attention to their quality of life. However, it doesn't directly address whether the elderly will be excluded from the consumer economy due to mobile payments. It doesn't undermine the expert's claim but rather provides context.

\*\*B.\*\* \*There are many elderly people who, because of their age, basically do not directly shop; the needed goods are generally provided by their children or society; their later life is very happy.\*

- This suggests that the elderly are less involved in direct shopping and rely on others for their needs, implying that being blocked from the consumer economy might not affect their quality of life. This option challenges the expert's claim by indicating the elderly might not be as impacted as suggested.

\*\*C.\*\* \*The relevant state departments have issued a number of policies in recent years, pointing out that consumers can complain when they are refused cash payments, but there are still many businesses that go their own way.\*

- This supports the expert's concern by highlighting that despite policies, many businesses refuse cash, possibly excluding the elderly who use cash. It does not call into question the expert's claim.

\*\*D.\*\* \*Many elderly people have learned mobile payment methods at home or at community centers, as well as tips to prevent online fraud.\*

- This directly challenges the expert's claim by stating that many elderly people are adapting to mobile payments. If the elderly are learning to use mobile payment methods, they are less likely to be blocked from the consumer economy, and their quality of life may not be negatively affected.

\*\*Conclusion:\*\*

Option \*\*D\*\* most effectively calls into question the expert's claim by providing evidence that the elderly are adapting to mobile payments, thereby reducing the likelihood of being excluded from the consumer economy.

**\*\*Answer: D. Many elderly people have learned mobile payment methods at home or at community centers\***

Figure 108: **Logical Reasoning: Case 6.** The analysis performance of the o1-preview model is rather excellent. It conducts a meticulous analysis of each option, clearly expounds the relationship between each option and the expert's view. Through step-by-step reasoning, it accurately identifies the option that most effectively challenges the expert's view. In the analysis process, the logic is rigorous and comprehensive considerations are made, which can effectively help users understand the problem and reach the correct answer.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

A certain bank was robbed, and four people, namely A, B, C, and D, are suspected and detained for interrogation. The investigation result shows that the criminal is one of them. A said, "It was stolen by C." B said, "I didn't steal it." C said, "I didn't steal it either." D said, "If B didn't steal it, then it was stolen by me." It has now been verified that only one of them told a lie.

Based on the above conditions, which of the following can be determined to be established?

- A. A stole.
- B. B stole.
- C. C stole.
- D. D stole.

### **O1-preview**

To solve this logical reasoning problem, we'll analyze the statements made by each suspect and use the information that only one person is lying.

**Statements:**

- **A:** "It was stolen by C."
- **B:** "I didn't steal it."
- **C:** "I didn't steal it either."
- **D:** "If B didn't steal it, then it was stolen by me."

**Analysis:**

1. **Assume A is lying** (the only liar), and B, C, and D are telling the truth.

- Since A is lying about C stealing, **C did not steal it**.
- B says he didn't steal it and is telling the truth, so **B did not steal it**.
- C says he didn't steal it and is telling the truth, confirming **C did not steal it**.
- D's truthful statement: "If B didn't steal it, then it was stolen by me."
- Since B didn't steal it, according to D's statement, **D stole it**.

This scenario is consistent with only one person lying (A) and leads to D being the thief.

2. **Assuming any other person is the liar** leads to contradictions:

- **If B is lying**: He did steal it. But this contradicts D's statement if D is telling the truth.
- **If C is lying**: He did steal it. But this contradicts the truthfulness of A's statement, suggesting more than one person is lying.
- **If D is lying**: His statement is false, but this leads to contradictions with C's truthfulness.

**Conclusion:**

The only consistent scenario is when **A is lying**, and the thief is **D**.

**Answer: D. D stole.**

Figure 109: **Logical Reasoning: Case 7.** The o1-preview model performs extremely well in analyzing the bank robbery problem. It closely adheres to the key point that only one person is lying. It meticulously analyzes each person's statement and conducts step-by-step reasoning through assumptions. This method is both organized and efficient. When delving into each assumption in combination with the problem's conditions, for example, when assuming that A is lying, it accurately deduces the result. The logic is strict and convincing. Additionally, the model fully considers various possibilities to avoid one-sided conclusions, providing users with comprehensive and accurate analysis results, which helps users better understand the problem and reach the correct answer.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

Either Li has been admitted to Tsinghua University or Sun has not been admitted to Peking University. Which of the following conditions should be added so that it can be inferred that Li has been admitted to Tsinghua University?

- A. At least one of Zhang and Sun has not been admitted to Peking University.
- B. At least one of Zhang and Li has not been admitted to Tsinghua University.
- C. Both Zhang and Sun have been admitted to Peking University.
- D. Neither Zhang nor Sun has been admitted to Peking University.

### **O1-preview**

To solve this logical reasoning problem, let's begin by restating the premise and the options using logical symbols for clarity.

**\*\*Premise\*\*:**

- Either \*\*Li has been admitted to Tsinghua University ( $L$ )\*\* \*\*or\*\* \*\*Sun has not been admitted to Peking University\*\*.

This can be symbolically represented as:

$$1. (\ L \lor \neg S )$$

We are asked to determine which additional condition, when combined with the premise, allows us to conclude that **\*\*Li has been admitted to Tsinghua University ( $L$ )\*\***.

Let's analyze each option:

**\*\*Option A\*\*:** At least one of Zhang and Sun has not been admitted to Peking University.  
- Symbolically:  $( \neg Z \lor \neg S )$

**\*\*Option B\*\*:** At least one of Zhang and Li has not been admitted to Tsinghua University.  
- Symbolically:  $( \neg Z_T \lor \neg L )$  (Where  $( Z_T )$  represents "Zhang has been admitted to Tsinghua University")

**\*\*Option C\*\*:** Both Zhang and Sun have been admitted to Peking University.  
- Symbolically:  $( Z \land S )$

**\*\*Option D\*\*:** Neither Zhang nor Sun has been admitted to Peking University.  
- Symbolically:  $( \neg Z \land \neg S )$

Our goal is to find which option, when combined with the premise  $( L \lor \neg S )$ , allows us to deduce  $( L )$  is true. **Analyzing Option C\*\*:**

Given:

$$1. (\ L \lor \neg S )$$

$$2. ( Z \land S )$$

From (2),  $( S )$  is true.

Substituting  $( S )$  into the premise:

- Since  $( S )$  is true,  $( \neg S )$  is false.
- Thus,  $( L \lor \text{False} )$  simplifies to  $( L )$ .
- Therefore,  $( L )$  must be true.

**Conclusion\*\*:**

**Only "Option C" provides sufficient information to deduce that Li has been admitted to Tsinghua University.**

**Answer. C**

Figure 110: **Logical Reasoning: Case 8.** Red indicates that the model's output is completely inconsistent with the reference diagnosis. The o1-preview model is highly effective in analyzing this logical reasoning problem. It presents the premise "Either Li has been admitted to Tsinghua University or Sun has not been admitted to Peking University" using logical symbols. When analyzing options, it closely integrates the problem content. For option C, starting from the given condition "Both Zhang and Sun have been admitted to Peking University," it conducts accurate deduction, showing a tight logical process and helping users understand and find the correct answer.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

A:Do not do what you don't want to do.B: I am against it.I want to do what I want.  
Which of the following is most similar to the above dialogue?

- A.A: People are not plants, can they be ruthless? B: I object.The plants are ruthless, but people are affectionate.  
B.A: People do not offend me, I do not offend.B: I object.If a person commits a crime, I commit a crime.  
C.A: If people do not have far-sightedness, they must have near-worry.B: I object.People have long-term worries and near worries.  
D.A: It is not in its position, and it does not seek its politics.B: I object.In its place, it will do its thing.

### **O1-preview**

To solve this problem, we need to analyze the structure and content of the original dialogue and compare it with the options to find the one that is most similar.

**\*\*Original Dialogue:\*\***

- **\*\*A:\*\*** "Do not do what you don't want to do."
- **\*\*B:\*\*** "I object. I want to do what I want."

Here, A is advising against doing things one doesn't like (implying not to impose on others what you wouldn't want yourself), while B misunderstands or twists this advice to assert their desire to do what they want, effectively contradicting A.

Now, let's examine each option:

**\*\*Option A:\*\***

- **\*\*A:\*\*** "People are not plants, can they be ruthless?"
- **\*\*B:\*\*** "I object. Plants are emotionless, but people are affectionate."

Here, B is agreeing with A by reinforcing the idea that people have emotions, so this doesn't match the original dialogue where B contradicts A.

**\*\*Option B:\*\***

- **\*\*A:\*\*** "If people do not offend me, I do not offend."
- **\*\*B:\*\*** "I object. If a person commits an offense, I commit an offense."

B is stating a conditional retaliation, which is a different concept from the original dialogue's structure.

**\*\*Option C:\*\***

- **\*\*A:\*\*** "If people do not have far-sightedness, they must have near-worry."
- **\*\*B:\*\*** "I object. People have long-term worries and near worries."

B is adding information rather than contradicting or twisting A's statement, so this doesn't match the original dialogue.

**\*\*Option D:\*\***

- **\*\*A:\*\*** "It is not in its position, and it does not seek its politics."
- **\*\*B:\*\*** "I object. In its place, it will do its thing."

In this option, A advises not to involve oneself in matters outside one's position, while B counters by emphasizing action when in the appropriate position. B's response is a direct twist or contradiction of A's statement, similar to the original dialogue.

**\*\*Conclusion:\*\***

Option D mirrors the original dialogue's structure and the nature of B's objection. Both involve A giving an advice or principle, and B responding by asserting their own stance that twists or contradicts A's statement.

**\*\*Answer: D\*\***

Figure 111: **Logical Reasoning: Case 9.** The o1-preview model demonstrates excellent performance in analyzing this problem. When analyzing option D, it clearly points out that A's statement "If one is not in a particular office, one should not meddle in its affairs" and B's response "I object. If one is in a particular office, one will handle its affairs" are similar in structure to the original dialogue. In both cases, A presents a principle-based view, and B responds in a way that twists or rebuts.

### **Prompt**

Here are some logical reasoning problems and please deduce the most realistic one from the four options given according to the description.

As long as cancer patients receive chemotherapy, they will suffer from strong side effects, and the side effects of chemotherapy will lead to a decrease in the resistance of cancer patients. Therefore, people with decreased resistance are prone to getting cancer.

Which of the following is the closest to the reasoning in the above text?

- A. As long as students study hard, their grades will be greatly improved, and if their grades are improved, they can be awarded a scholarship. So if they study hard, they can be awarded a scholarship.
- B. Only physicists understand the theory of relativity, and only those who understand the theory of relativity can engage in related research. Mr. Wang is engaged in work related to the theory of relativity, so he must understand the theory of relativity.
- C. If Ballack plays, the German team will definitely win. If the German team wins, they can get a high bonus. So when the German team gets a high bonus, Ballack must have played.
- D. Lao Tian is Xiao Tian's father, and Da Tian is Lao Tian's younger brother. So, Da Tian is Xiao Tian's uncle.

### **O1-preview**

To solve this problem, let's analyze the reasoning presented in the passage and compare it with the reasoning in each option to find the one that is most similar.

**\*\*Passage Analysis:\*\***

- 1. **Premise 1:** Cancer patients who receive chemotherapy will suffer strong side effects.
- 2. **Premise 2:** The side effects of chemotherapy lead to decreased resistance (immunity).
- 3. **Conclusion:** People with decreased resistance are prone to getting cancer.

**\*\*Logical Flaw in the Passage:\*\***

- The reasoning is flawed because it suggests that decreased resistance leads to getting cancer, but the patients already have cancer. The conclusion doesn't logically follow from the premises. It implies a circular reasoning or a logical fallacy where the conclusion doesn't directly result from the premises.

**\*\*Option Analysis:\*\***

**\*\*Option A:\*\***

- 1. Students who study hard will greatly improve their grades.
- 2. Improved grades can lead to a scholarship.
- 3. Conclusion: If they study hard, they can be awarded a scholarship.

**\*\*Reasoning Type:\*\*** Logical and straightforward; the conclusion follows from the premises.

**\*\*Option B:\*\***

- 1. Only physicists understand the theory of relativity.
- 2. Only those who understand it can engage in related research.
- 3. Mr. Wang is engaged in related research, so he must understand it.

**\*\*Reasoning Type:\*\*** Logical deduction; the conclusion is valid.

**\*\*Option C:\*\***

- 1. If Ballack plays, the German team will definitely win.
- 2. If the German team wins, they can get a high bonus.
- 3. Conclusion: When the German team gets a high bonus, Ballack must have played.

**\*\*Reasoning Type:\*\*** Flawed reasoning; the conclusion assumes that because the team got a bonus (they won), Ballack must have played, which isn't necessarily true.

**\*\*Option D:\*\***

- Family relationships are outlined, leading to a logical and factual conclusion.

**\*\*Comparison:\*\***

- **Option C** mirrors the flawed reasoning in the passage. It makes an invalid inference by assuming that a result (the team getting a high bonus) confirms a specific cause (Ballack playing), ignoring other possibilities (the team could have won without Ballack).

**\*\*Conclusion:\*\***

Option C is the closest to the reasoning in the passage, as both contain flawed logic by incorrectly inferring causation from a sequence of events.

**\*\*Answer:\*\* C**

Figure 112: **Logical Reasoning: Case 10.** It meticulously analyzes and compares the reasoning of each option, elaborating that Option A is logically reasonable, Option B's logical deduction is effective, Option C has a similar defect to the question stem, and Option D is a clear conclusion about family relationships. Through rigorous comparison, it finds Option C, which is closest to the question stem. The entire analysis process is logically strict and well-structured, providing users with problem-solving ideas and accurate answers.

## A.9 High School Level Math Competition

### Prompt

How many vertical asymptotes does the graph of  $y=\frac{2}{x^2+x-6}$  have?

### O1-preview

To determine the number of vertical asymptotes for the function  $y=\frac{2}{x^2+x-6}$ , we need to find the values of  $x$  that make the denominator zero, as these are the points where the function is undefined and may have vertical asymptotes.

#### Step 1: Factor the Denominator

First, factor the quadratic expression in the denominator:

$$x^2 + x - 6 = (x + 3)(x - 2)$$

#### Step 2: Find the Zeros of the Denominator

Set each factor equal to zero to find the zeros:

- $x+3=0 \Rightarrow x=-3$
- $x-2=0 \Rightarrow x=2$

#### Step 3: Determine if They Are Vertical Asymptotes

Since the numerator 2 is a constant and does not cancel out any factors in the denominator, both  $x=-3$  and  $x=2$  are vertical asymptotes.

#### Conclusion:

The graph of  $y = \frac{2}{x^2+x-6}$  has two vertical asymptotes at  $x=-3$  and  $x=2$ .

**Answer: 2**

### Reference Answer

The denominator of the rational function factors into  $x^2+x-6=(x-2)(x+3)$ . Since the numerator is always nonzero, there is a vertical asymptote whenever the denominator is 0, which occurs for  $x = 2$  and  $x = -3$ . Therefore, the graph has 2 vertical asymptotes.

Figure 113: **High School Level Math Competition: Case 5: Algebra** This level 3 algebra problem involves determining the number of vertical asymptotes for the rational function. The yellow highlights indicate shared steps in both the o1-preview output and the reference answer, while the green highlights mark the shared final answer.

### **Prompt**

Find the distance between the vertex of the graph of the equation  $f(x) = x^2 - 8x + 15$  and the point  $(0, 2)$ .

### **O1-preview**

To find the distance between the vertex of the parabola defined by  $f(x)=x^2-8x+15$  and the point  $(0,2)$ , we'll follow these steps:

Step 1: Find the Vertex of the Parabola

A quadratic function in the form  $f(x)=ax^2+bx+c$  has its vertex at:

$$x_v=-b/2a$$

Given:

$$a=1$$

$$b=-8$$

$$c=15$$

Calculate the x-coordinate of the vertex:

$$x_v=-(-8)/2\times1=8/2=4$$

Now, find the y-coordinate by plugging  $x_v$  back into the function:

$$y_v=f(x_v)=f(4)=(4)^2-8\times4+15=16-32+15=-1$$

So, the vertex is at the point  $(4, -1)$ .

Step 2: Calculate the Distance Between the Vertex and the Point  $(0,2)$

Use the distance formula between two points  $(x_1,y_1)$  and  $(x_2,y_2)$ :

$$d=[(x_2-x_1)^2+(y_2-y_1)^2]^{1/2}$$

Plug in the coordinates:

$$d=[(0-4)^2+(2-(-1))^2]^{1/2}$$

$$=[(-4)^2+(3)^2]^{1/2}$$

$$=[16+9]^{1/2}$$

$$=[25]^{1/2}$$

$$=5$$

Conclusion:

The distance between the vertex  $(4, -1)$  and the point  $(0,2)$  is 5 units.

**Answer:** 5

### **Reference Answer**

Completing the square, we get  $f(x) = (x-4)^2 - 1$ . The vertex of the graph of this equation is thus  $(4, -1)$ . Using the Pythagorean Theorem, it follows that the distance between  $(0, 2)$  and  $(4, -1)$  is  $\boxed{5}$ .

Figure 114: **High School Level Math Competition: Case 6: Algebra** This level 4 algebra problem involves determining the distance between the vertex of a parabola and a given point. The yellow highlights indicate shared steps in both the o1-preview output and the reference answer, while the green highlight marks the shared final answer.

### **Prompt**

What is the simplified numerical value of  $\frac{a+11b}{a-b}$  if  $\frac{4a+3b}{a-2b}=5$ ?

### **O1-preview**

To find the simplified numerical value of  $(a+11b)/(a-b)$  given that  $(4a+3b)/(a-2b)=5$ , we'll follow these steps:

Step 1: Solve the Given Equation for aa in Terms of bb

We are given:

$$(4a+3b)/(a-2b)=5$$

Cross-multiply to eliminate the denominator:

$$4a+3b=5(a-2b)$$

Simplify the right-hand side:

$$4a+3b=5a-10b$$

Bring all terms to one side to solve for a:

$$4a+3b-5a+10b=0$$

Simplify:

$$(-a)+13b=0$$

Solve for aa:

$$-a+13b=0 \Rightarrow a=13b$$

Step 2: Substitute a=13b into the Expression

We need to find:

$$(a+11b)/(a-b)$$

Substitute a=13b:

$$(13b+11b)/(13b-b)$$

Simplify numerator and denominator:

Numerator:

$$13b+11b=24b$$

Denominator:

$$13b-b=12b$$

So the expression becomes:

$$(24b)/(12b)$$

Step 3: Simplify the Expression

Simplify by dividing numerator and denominator by bb (assuming b≠0b̸=0):

$$24b/12b=24/12=2$$

Conclusion

The simplified numerical value of  $(a+11b)/(a-b)$  is 2.

**Answer: 2**

### **Reference Answer**

Let's play with the given condition a little. Clearing out the denominator gives  $4a+3b=5(a-2b)=5a-10b$ .

Selectively combine like terms by adding  $9b-4a$  to both sides to get  $12b=a-b$ . This gives  $\frac{12b}{a-b}=1$ . Now, we want to find  $\frac{a+11b}{a-b}$ . Rewrite this as  $\frac{a-b+12b}{a-b}=\frac{a-b}{a-b}+\frac{12b}{a-b}=1+\frac{12b}{a-b}$ , and we are done.

Figure 115: **High School Level Math Competition: Case 7: Algebra** This level 4 algebra problem requires simplifying a given expression based on a relationship between two variables. The yellow highlights indicate the shared steps in both the o1-preview output and the reference answer.

**Prompt**

The probability of rain tomorrow is  $\frac{1}{11}$ . What is the probability that it will not rain tomorrow? Express your answer as a common fraction.

**O1-preview**

To find the probability that it will not rain tomorrow, subtract the probability of rain from 1 (since the total probability of all possible outcomes is 1).

Given:

Probability of rain =  $\frac{1}{11}$

Step 1: Calculate the Probability of No Rain

$$\text{Probability of no rain} = 1 - \text{Probability of rain} = 1 - \frac{1}{11} = \frac{10}{11}$$

Conclusion:

The probability that it will not rain tomorrow is  $\frac{10}{11}$ .

Answer:  $\frac{10}{11}$

**Reference Answer**

It must either rain tomorrow or not rain tomorrow, so the sum of the probability that it rains and the probability it doesn't rain is 1. Therefore, the probability it doesn't rain is  $1 - \frac{1}{11} = \boxed{\frac{10}{11}}$ .

Figure 116: **High School Level Math Competition: Case 8: Counting and Probability**  
This level 1 probability problem involves calculating the likelihood that it will not rain tomorrow, given the probability of rain. The yellow highlights indicate shared steps in both the o1-preview output and the reference answer, with the green highlight marking the correct final answer.

### **Prompt**

We flip a fair coin 10 times. What is the probability that we get heads in at least 6 of the 10 flips?

### **O1-preview**

To find the probability of getting at least 6 heads when flipping a fair coin 10 times, we'll use the binomial probability formula. This scenario follows a binomial distribution with parameters  $n=10$  (number of trials) and  $p=0.5$  (probability of success on each trial).

Step 1: Understand the Binomial Probability Formula

The probability of getting exactly  $k$  successes in  $n$  independent trials is given by:

$$P(X=k)=C(n,k)\times p^k\times(1-p)^{n-k}$$

Where:

- $C(n,k)$  is the number of combinations (ways to choose  $k$  successes out of  $n$  trials).
- $p$  is the probability of success on a single trial.
- $(1-p)$  is the probability of failure on a single trial.

Step 2: Calculate Probabilities for  $k=6$  to  $k=10$

We need to find the sum of probabilities from  $k=6$  to  $k=10$ :

$$P(X \geq 6) = \sum_{k=6}^{10} P(X = k)$$

Since  $p=0.5$ , the term  $p^k \times (1-p)^{n-k}$  simplifies to  $(0.5)^{10}$  for all  $k$ .

$$\text{So, } P(X=k)=C(10,k)\times(0.5)^{10}$$

Compute  $C(10,k)$  for each  $k$ :

- For  $k=6$ :  $C(10,6)=10! / 6!(10-6)!=210$
- For  $k=7$ :  $C(10,7)=10! / 7!(10-7)!=120$
- For  $k=8$ :  $C(10,8)=10! / 8!(10-8)!=45$
- For  $k=9$ :  $C(10,9)=10! / 9!(10-9)!=10$
- For  $k=10$ :  $C(10,10)=10! / 10!(10-10)!=1$

Step 3: Sum the Probabilities

Add up the probabilities for each  $k$ :

$$P(X \geq 6)=[C(10,6)+C(10,7)+C(10,8)+C(10,9)+C(10,10)]\times(0.5)^{10}=[210+120+45+10+1]\times(0.5)^{10}=386\times(0.5)^{10}$$

Since  $(0.5)^{10}=1/1024$  :

$$P(X \geq 6)=386/1024$$

Step 4: Simplify the Fraction

Simplify  $386/1024$ :

$$386/1024=193/512$$

(Note: 193 is a prime number, so the fraction is in lowest terms.)

Conclusion

The probability of getting at least 6 heads in 10 flips of a fair coin is  $193/512$ .

**Answer:**  $193/512$

### **Reference Answer**

There are  $2^{10} = 1024$  possible outcomes of the 10 coin flips. The probability that we flip at least 6 heads is equal to the probability that we flip at least 6 tails, by symmetry. Let's call this probability  $p$ . The only other possibility is that we flip exactly 5 heads and 5 tails, for which the probability is  $\frac{\binom{10}{5}}{2^{10}} = \frac{252}{1024} = \frac{63}{256}$ . Therefore,  $\frac{63}{256} + 2p = 1$ , giving  $p = \frac{1}{2} \left(1 - \frac{63}{256}\right) = \boxed{\frac{193}{512}}$

Figure 117: High School Level Math Competition: Case 9: Counting and Probability  
This level 4 probability problem involves finding the likelihood of getting heads in at least 6 out of 10 coin flips. The yellow highlights indicate the shared steps in both the o1-preview output and the reference answer, while the green highlights mark the shared final answer.

### **Prompt**

For a particular peculiar pair of dice, the probabilities of rolling 1, 2, 3, 4, 5, and 6 on each die are in the ratio \$1:2:3:4:5:6\$. What is the probability of rolling a total of 7 on the two dice?

### **O1-preview**

To find the probability of rolling a total of 7 with this peculiar pair of dice, we'll follow these steps:

Understanding the Dice Probabilities

Each die has outcomes (1 through 6) with probabilities proportional to the ratios 1:2:3:4:5:6. Let's calculate the actual probabilities for each face on a single die.

Step 1: Calculate the Total Weight

Add up the weights for all possible outcomes:

$$\text{Total weight} = 1+2+3+4+5+6=21$$

Step 2: Calculate Individual Probabilities

The probability of each outcome is its weight divided by the total weight.

$$P(\text{rolling a } 1)=1/21$$

$$P(\text{rolling a } 2)=2/21$$

$$P(\text{rolling a } 3)=3/21$$

$$P(\text{rolling a } 4)=4/21$$

$$P(\text{rolling a } 5)=5/21$$

$$P(\text{rolling a } 6)=6/21$$

Calculating the Probability of a Total of 7

We need to find all pairs  $(x,y)$  such that  $x+y=7$ , where  $x$  is the outcome on the first die and  $y$  is the outcome on the second die.

Possible Pairs Summing to 7:

$$(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$$

Step 3: Calculate the Probability for Each Pair

The probability of each pair is the product of the probabilities of the individual outcomes:

$$\text{Pair (1,6): } P(1,6)=P(1)\times P(6)=1/21\times 6/21=6/441$$

$$\text{Pair (2,5)(2,5): } P(2,5)=2/21\times 5/21=10/441$$

$$\text{Pair (3,4)(3,4): } P(3,4)=3/21\times 4/21=12/441$$

$$\text{Pair (4,3)(4,3): } P(4,3)=4/21\times 3/21=12/441$$

$$\text{Pair (5,2)(5,2): } P(5,2)=5/21\times 2/21=10/441$$

$$\text{Pair (6,1)(6,1): } P(6,1)=6/21\times 1/21=6/441$$

Step 4: Sum the Probabilities

Add up the probabilities of all pairs that sum to 7:

$$P(\text{total of 7})=P(1,6)+P(2,5)+P(3,4)+P(4,3)+P(5,2)+P(6,1)=6/441+10/441+12/441+12/441+10/441+6/441=56/441$$

Step 5: Simplify the Fraction

Simplify  $56/441$ :

Find the Greatest Common Divisor (GCD) of 56 and 441:

$$\text{GCD}(56,441)=7$$

Divide numerator and denominator by 7:

$$56\div 7441\div 7=8/63$$

Conclusion

The probability of rolling a total of 7 with this peculiar pair of dice is  $8/63$ .

Answer:  $8/63$

### **Reference Answer**

On each die the probability of rolling  $k$ , for  $1 \leq k \leq 6$ , is  $\frac{k}{21}$ . There are six ways of rolling a total of 7 on the two dice, represented by the ordered pairs  $(1,6)$ ,  $(2,5)$ ,  $(3,4)$ ,  $(4,3)$ ,  $(5,2)$ , and  $(6,1)$ . Thus the probability of rolling a total of 7 is  $\frac{6}{21^2}=\frac{6}{441}=\boxed{\frac{8}{63}}$ .

Figure 118: High School Level Math Competition: Case 10: Counting and Probability  
This is a level 4 probability problem.

## A.10 College-level Math Problems

|  |  |
|--|--|
| <p><b>Prompt</b></p> <p>Consider the set of integers <math>\{1, 2, \dots, n\}</math>. How many subsets of this set do not contain consecutive integers? Please deduce an explicit formula step by step.</p> <p><b>O1-preview</b></p> <p>To find the number of subsets of the set <math>\{1, 2, \dots, n\}</math> that do not contain consecutive integers, we'll derive an explicit formula step by step.</p> <p><b>Step 1: Model the Problem with Binary Sequences</b></p> <p>Consider each integer in the set <math>\{1, 2, \dots, n\}</math> as a position in a binary sequence of length <math>(n)</math>. Each position can be either:</p> <ul style="list-style-type: none"> <li>- <b>0</b>: the integer is not included in the subset.</li> <li>- <b>1</b>: the integer is included in the subset.</li> </ul> <p>The constraint that no two consecutive integers are included translates to the condition that no two consecutive positions in the binary sequence can both be <b>1</b>.</p> <p><b>Step 2: Recognize the Fibonacci Relationship</b></p> <p>The number of binary sequences of length <math>(n)</math> with no two consecutive <b>1</b>'s is known to follow the Fibonacci sequence. Specifically, the count satisfies the recurrence relation:</p> $[ S(n) = S(n - 1) + S(n - 2) ]$ <p>with initial conditions:</p> $[ S(0) = 1 \quad (\text{the empty subset}), \quad S(1) = 2 \quad (\text{either include 1 or not}). ]$ <p>This is because:</p> <ul style="list-style-type: none"> <li>- <b>Case 1:</b> If the <math>(n)</math>th element is <b>0</b>, the problem reduces to counting sequences of length <math>(n - 1)</math>, which is <math>(S(n - 1))</math>.</li> </ul> | <ul style="list-style-type: none"> <li>- <b>Case 2:</b> If the <math>(n)</math>th element is <b>1</b>, the <math>((n - 1))</math>th element must be <b>0</b> (to avoid consecutive <b>1</b>'s), so the problem reduces to sequences of length <math>(n - 2)</math>, which is <math>(S(n - 2))</math>.</li> </ul> <p><b>Step 3: Express <math>(S(n))</math> Using Fibonacci Numbers</b></p> <p>By observing the initial conditions and the recurrence relation, <math>(S(n))</math> corresponds to the <math>((n + 2))</math>th Fibonacci number <math>(F_{n+2})</math>. The Fibonacci sequence is defined by:</p> $[ F_n = F_{n-1} + F_{n-2} ]$ <p>with initial conditions <math>(F_0 = 0)</math> and <math>(F_1 = 1)</math>.</p> <p><b>Step 4: Write the Explicit Formula Using Binet's Formula</b></p> <p>The explicit formula for the <math>(n)</math>th Fibonacci number is given by Binet's formula:</p> $[ F_n = \frac{\phi^n - \psi^n}{\sqrt{5}} ]$ <p>where:</p> $[ \phi = \frac{1 + \sqrt{5}}{2} \quad (\text{the golden ratio}), \quad \psi = \frac{1 - \sqrt{5}}{2} \quad (\text{the conjugate of the golden ratio}). ]$ <p><b>Step 5: Substitute into the Formula for <math>(S(n))</math></b></p> <p>Since <math>(S(n) = F_{n+2})</math>, we have:</p> $[ S(n) = \frac{\phi^{n+2} - \psi^{n+2}}{\sqrt{5}} ]$ <p><b>Final Answer:</b></p> <p>An explicit formula: <math>S(n) = [\varphi^{n+2} - \psi^{n+2}] / \sqrt{5}</math>, where <math>\varphi = (1 + \sqrt{5})/2</math> and <math>\psi = (1 - \sqrt{5})/2</math></p> |
|--|--|

Figure 119: **College-Level Math: Case 2.** A demonstration case of solving a basic discrete math problem. The generated answer is manually evaluated as correct.

|  |   |
|--|---|
| <p><b>Prompt</b></p> <p>Consider two sequences of integers <math>a_1 &lt; a_2 &lt; \dots &lt; a_n</math> and <math>b_1 &gt; b_2 &gt; \dots &gt; b_n</math>, where the set <math>\{a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n\} = \{1, 2, \dots, 2n\}</math>. Prove that <math>\sum_{i=1}^n  a_i - b_i  = n^2</math>. Be careful not assume any additional relationships between <math>a_i</math>'s and <math>b_i</math>'s.</p> <p><b>O1-preview</b></p> <p>To prove that <math>(\sum_{i=1}^n  a_i - b_i  = n^2)</math> given the conditions, we'll proceed in a systematic manner.</p> <p><b>Given:</b></p> <ul style="list-style-type: none"> <li>- Two sequences of integers (<math>a_1 &lt; a_2 &lt; \dots &lt; a_n</math>) and (<math>b_1 &gt; b_2 &gt; \dots &gt; b_n</math>).</li> <li>- The set <math>\{a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n\} = \{1, 2, \dots, 2n\}</math>.</li> </ul> <p><b>Objective:</b></p> <ul style="list-style-type: none"> <li>- Prove that <math>(\sum_{i=1}^n  a_i - b_i  = n^2)</math>.</li> </ul> <p><b>Proof:</b></p> <ol style="list-style-type: none"> <li>1. <b>Observation about the Sums of (<math>a_i</math>) and (<math>b_i</math>):</b><br/>The sum of all numbers from 1 to <math>(2n)</math> is: <math>\sum_{k=1}^{2n} k = n(2n + 1)</math>. Thus, the total sum of the (<math>a_i</math>) and (<math>b_i</math>) sequences is: <math>\sum_{i=1}^n a_i + \sum_{i=1}^n b_i = n(2n + 1)</math>.</li> <li>2. <b>Observation about Pairwise Sums:</b><br/>If we can show that <math>(a_i + b_i = 2n + 1)</math> for all <math>(i)</math>, then: <math> a_i - b_i  =  2a_i - (2n + 1) </math>. Let's check if this condition can be met.</li> <li>3. <b>Constructing Sequences with <math>(a_i + b_i = 2n + 1)</math>:</b><br/>Let's define: <math>a_i = i</math>, <math>b_i = 2n + 1 - i</math>, for <math>i = 1, 2, \dots, n</math>. This ensures that <math>(a_i &lt; a_{i+1})</math> and <math>(b_i &gt; b_{i+1})</math>, satisfying the conditions of the sequences.</li> </ol> | <p>4. <b>Verifying that the Numbers are Distinct and Cover (1) to <math>(2n)</math>:</b><br/>The (<math>a_i</math>) are <math>(1, 2, \dots, n)</math>, and the (<math>b_i</math>) are <math>(2n, 2n - 1, \dots, n + 1)</math>. Together, they cover all integers from <math>(1)</math> to <math>(2n)</math>.</p> <p>5. <b>Calculating (<math> a_i - b_i </math>):</b><br/><math> a_i - b_i  =  i - (2n + 1 - i)  =  2i - (2n + 1) </math>.</p> <p>6. <b>Summing Up (<math> a_i - b_i </math>):</b><br/>The sum is: <math>\sum_{i=1}^n  2i - (2n + 1)  = \sum_{i=1}^n (2n + 1 - 2i) = n(2n + 1) - 2\sum_{i=1}^n i</math>. Calculate <math>(\sum_{i=1}^n i = \frac{n(n + 1)}{2})</math>, so: <math>\sum_{i=1}^n  a_i - b_i  = n(2n + 1) - n(n + 1) = n(2n + 1 - n - 1) = n^2</math>.</p> <p>7. <b>General Case:</b><br/>Regardless of how the sequences (<math>a_i</math>) and (<math>b_i</math>) are chosen under the given conditions, the pairing will always satisfy <math>(a_i + b_i = 2n + 1)</math>, because the total sum must be <math>(n(2n + 1))</math>, and the sequences are strictly increasing and decreasing, respectively.</p> <p>8. <b>Conclusion:</b><br/>Thus, <math>(\sum_{i=1}^n  a_i - b_i  = n^2)</math>.</p> <p><b>Answer:</b><br/><b>Answer:</b> Proved: the sum of <math> a_i - b_i </math> to <math> a_n - b_1 </math> always equals <math>n^2</math>—that is, <math>\sum  a_i - b_i  = n^2</math>.</p> |
|--|---|

Figure 120: **College-Level Math: Case 3.** A demonstration case of solving an advanced discrete math problem. The generated answer is manually evaluated as incorrect.

|   |   |
|---|---|
| <p><b>Prompt</b></p> <p>Prove that for any integer <math>n \geq 2</math>, there exists a permutation of the sequence <math>(1, 2, \dots, n)</math> such that when the integers are concatenated into a single integer, the resulting number is divisible by 7. For example, when <math>n = 2</math>, the permutation <math>(2, 1)</math> forms the integer 21, which is divisible by 7.</p> <p><b>O1-preview</b></p> <p>To prove that for any integer (<math>n \geq 2</math>), there exists a permutation of the sequence <math>((1, 2, \dots, n))</math> such that the concatenated number is divisible by 7, we'll use modular arithmetic and the pigeonhole principle.</p> <p><b>Understanding the Problem:</b></p> <p>When we concatenate the numbers <math>(a_1, a_2, \dots, a_n)</math> into a single integer (<math>N</math>), it can be represented as: <math>[N = a_1 \times 10^{L_1} + a_2 \times 10^{L_2} + \dots + a_n]</math> where <math>(L_i)</math> is the total number of digits after <math>(a_i)</math>. Our goal is to find a permutation <math>((a_1, a_2, \dots, a_n))</math> such that <math>(N)</math> is divisible by 7, i.e., <math>[N \equiv 0 \pmod{7}]</math>.</p> <p><b>Key Observations:</b></p> <ol style="list-style-type: none"> <li><b>Powers of 10 Modulo 7:</b> The powers of 10 modulo 7 repeat every 6 steps because <math>(10^6 \equiv 1 \pmod{7})</math>. Specifically: <math>\begin{cases} 10^k \pmod{7} = 1 &amp; \text{if } k \equiv 0 \pmod{6} \\ 10^k \pmod{7} = 2 &amp; \text{if } k \equiv 1 \pmod{6} \\ 10^k \pmod{7} = 4 &amp; \text{if } k \equiv 2 \pmod{6} \\ 10^k \pmod{7} = 8 &amp; \text{if } k \equiv 3 \pmod{6} \\ 10^k \pmod{7} = 6 &amp; \text{if } k \equiv 4 \pmod{6} \\ 10^k \pmod{7} = 5 &amp; \text{if } k \equiv 5 \pmod{6} \end{cases}</math></li> <li><b>Coefficients Modulo 7:</b> For each position (<math>i</math>), define the coefficient <math>(c_i = 10^{n-i} \pmod{7})</math>. The sequence <math>(c_i)</math> repeats every 6 positions due to the periodicity of <math>(10^k \pmod{7})</math>.</li> <li><b>Linear Congruence:</b> We need to solve the congruence: <math>[\sum_{i=1}^n c_i a_i \equiv 0 \pmod{7}]</math> where <math>(a_i)</math> are the numbers <math>(1)</math> through <math>(n)</math> in some permutation.</li> </ol> | <p><b>Proof by Pigeonhole Principle:</b></p> <ul style="list-style-type: none"> <li>- The number of possible permutations of <math>(n)</math> distinct numbers is <math>(n!)</math>.</li> <li>- The number of possible remainders when divided by 7 is 7.</li> <li>- For <math>(n \geq 3)</math>, <math>(n! \geq 6)</math>. For <math>(n \geq 7)</math>, <math>(n! \geq 5040)</math>.</li> </ul> <p>By the pigeonhole principle, since there are more permutations than possible remainders, there must be at least one permutation where <math>(\sum_{i=1}^n c_i a_i \equiv 0 \pmod{7})</math>.</p> <p><b>Constructing an Example for Clarity (Case (<math>n = 6</math>)):</b></p> <p>Let's compute <math>(c_i)</math> for <math>(n = 6)</math>: <math>\begin{aligned} c_1 &amp;= 10^5 \pmod{7} = 5 \\ c_2 &amp;= 10^4 \pmod{7} = 4 \\ c_3 &amp;= 10^3 \pmod{7} = 6 \\ c_4 &amp;= 10^2 \pmod{7} = 2 \\ c_5 &amp;= 10^1 \pmod{7} = 3 \\ c_6 &amp;= 10^0 \pmod{7} = 1 \end{aligned}</math></p> <p>We need to find <math>(a_i)</math> such that: <math>[5a_1 + 4a_2 + 6a_3 + 2a_4 + 3a_5 + a_6 \equiv 0 \pmod{7}]</math> with <math>(a_i)</math> being a permutation of <math>(1, 2, 3, 4, 5, 6)</math>.</p> <p>After some trial (as detailed in the solution), one such permutation is <math>(a_1, a_2, a_3, a_4, a_5, a_6) = (1, 2, 3, 5, 6, 4)</math>, satisfying the congruence: <math>[5 \times 1 + 4 \times 2 + 6 \times 3 + 2 \times 5 + 3 \times 6 + 1 \times 4 \equiv 0 \pmod{7}]</math></p> <p><b>Conclusion:</b></p> <p>Since for <math>(n \geq 2)</math>, the number of permutations exceeds the number of possible remainders modulo 7, there must exist at least one permutation satisfying the condition. Therefore, for any integer <math>(n \geq 2)</math>, there exists a permutation of <math>((1, 2, \dots, n))</math> such that the concatenated number is divisible by 7.</p> <p><b>Answer:</b></p> <p>Yes; for any <math>(n \geq 2)</math>, such a permutation exists—so the concatenated number divides by 7 for all <math>n \geq 2</math>.</p> |
|---|---|

Figure 121: College-Level Math: Case 4. A demonstration case of solving an advanced discrete math problem. The generated answer is manually evaluated as incorrect.

|  |   |
|--|---|
| <p><b>Prompt</b></p> <p>Let <math>f: [0, 1] \rightarrow \mathbb{R}</math> be continuously differentiable with <math>f(0) = 0</math>. Prove that <math>\sup_{0 \leq x \leq 1}  f(x)  \leq (\int_0^1 (f'(x))^2 dx)^{1/2}</math>.</p> <p><b>O1-preview</b></p> <p>To prove the inequality <math>\sup_{0 \leq x \leq 1}  f(x)  \leq \left(\int_0^1 [f'(x)]^2 dx\right)^{1/2}</math>, we start by utilizing the Fundamental Theorem of Calculus and the Cauchy-Schwarz inequality.</p> <p><b>Step 1: Express (<math>f(x)</math>) as an integral of its derivative.</b></p> <p>Since (<math>f</math>) is continuously differentiable on <math>([0, 1])</math> and (<math>f(0) = 0</math>), we have: <math>[f(x) = f(0) + \int_0^x f'(t) dt = \int_0^x f'(t) dt]</math>.</p> <p><b>Step 2: Apply the Cauchy-Schwarz inequality.</b></p> <p>We can estimate (<math> f(x) </math>) using the Cauchy-Schwarz inequality: <math>[ f(x)  = \left  \int_0^x f'(t) dt \right  \leq \left( \int_0^x 1^2 dt \right)^{1/2} \left( \int_0^x [f'(t)]^2 dt \right)^{1/2}]</math>. Simplify the first integral: <math>[\int_0^x 1^2 dt = x]</math>. So, <math>[ f(x)  \leq \sqrt{x} \left( \int_0^x [f'(t)]^2 dt \right)^{1/2}]</math>.</p> | <p><b>Step 3: Use the fact that (<math>x \leq 1</math>) and compare integrals.</b></p> <p>Since (<math>x \leq 1</math>) on the interval <math>([0, 1])</math>, we have (<math>\sqrt{x} \leq 1</math>). Also, the integral from <math>(0)</math> to <math>(x)</math> is less than or equal to the integral from <math>(0)</math> to <math>(1)</math>: <math>[\int_0^x [f'(t)]^2 dt \leq \int_0^1 [f'(t)]^2 dt]</math>. Therefore, <math>[ f(x)  \leq \left( \int_0^1 [f'(t)]^2 dt \right)^{1/2}]</math>.</p> <p><b>Step 4: Conclude the proof by taking the supremum.</b></p> <p>Since the above inequality holds for all (<math>x</math>) in <math>([0, 1])</math>, we have: <math>[\sup_{0 \leq x \leq 1}  f(x)  \leq \left( \int_0^1 [f'(t)]^2 dt \right)^{1/2}]</math>. This completes the proof.</p> <p><b>Conclusion:</b></p> <p>We have shown that the supremum of (<math> f(x) </math>) on the interval <math>([0, 1])</math> is bounded above by the square root of the integral of (<math>[f'(x)]^2</math>) over the same interval, as required.</p> <p><b>Answer: Proof:</b> Since <math>f(0) = 0</math>, we have <math> f(x)  \leq [\int_0^x [f'(t)]^2 dt]^{1/2}</math> for all <math>x</math>; thus the inequality holds.</p> |
|--|---|

Figure 122: College-Level Math: Case 6. A demonstration case of solving a Calculus problem. The generated answer is manually evaluated as correct.

### **Prompt**

Consider the function space  $C[0,1]$ , which is the space of all continuous real valued functions defined on  $[0,1]$  equipped with supremum norm. Show that every compact subset of  $C[0,1]$  has empty interior.

### **O1-preview**

To show that every compact subset ( $K$ ) of ( $C[0,1]$ ) (the space of continuous real-valued functions on  $([0,1])$  with the supremum norm) has an empty interior, we proceed by contradiction.

**Assumption for Contradiction:** Suppose there exists a compact subset ( $K \subseteq C[0,1]$ ) that has a non-empty interior. This means there exists a function ( $f_0 \in K$ ) and a radius ( $r > 0$ ) such that the open ball  $B_r(f_0) = \{f \in C[0,1] \mid |f - f_0|_{\infty} < r\}$  is entirely contained within ( $K$ ).

#### **Properties of ( $B_r(f_0)$ ):**

1. **Infinite Dimensionality:** The space ( $C[0,1]$ ) is infinite-dimensional. Any open ball in ( $C[0,1]$ ) is an infinite-dimensional subset because you can find infinitely many linearly independent functions within any open ball.
2. **Non-Compactness of Open Balls:** In infinite-dimensional normed spaces, open balls are not compact. This is because compactness in normed spaces requires sequential compactness, which fails in infinite dimensions due to the Riesz lemma.

#### **Contradiction:**

- Since ( $B_r(f_0) \subseteq K$ ) and ( $K$ ) is compact, it would imply that ( $B_r(f_0)$ ) is compact.
- However, this contradicts the fact that open balls in infinite-dimensional normed spaces like ( $C[0,1]$ ) are not compact.

#### **Conclusion:**

- Our assumption that ( $K$ ) has a non-empty interior must be false.
- Therefore, every compact subset ( $K$ ) of ( $C[0,1]$ ) must have an empty interior.

#### **Key Point:**

In infinite-dimensional Banach spaces like ( $C[0,1]$ ), compact subsets cannot contain open sets because open subsets (like open balls) are not compact. Hence, compact subsets must have empty interiors.

#### **Final Answer:**

Because in  $C[0,1]$ , any compact set is too “small” to contain an open set, every compact subset has empty interior.

Figure 123: **College-Level Math: Case 8.** A demonstration case of o1-preview reproducing the proof of an advanced math theorem. The generated answer is manually evaluated as correct.

|  |  |
|--|--|
| <p><b>Prompt</b></p> <p>Can you prove the fact that open balls in infinite-dimensional normed spaces are not compact?</p> <p><b>O1-preview</b></p> <p>Yes, I can prove that open balls in infinite-dimensional normed spaces are not compact.</p> <p><b>Proof:</b></p> <p>Let (<math>X</math>) be an infinite-dimensional normed space, and consider an open ball (<math>B_r(0)</math>) of radius (<math>r &gt; 0</math>) centered at the origin: <math>\{x \in X \mid \ x\  &lt; r\}</math></p> <p><b>Objective:</b></p> <p>Show that (<math>B_r(0)</math>) is not compact.</p> <p><b>Strategy:</b></p> <p>We will construct a sequence within (<math>B_r(0)</math>) that has no convergent subsequence. This will demonstrate that (<math>B_r(0)</math>) is not sequentially compact, and hence not compact, because in metric spaces (which normed spaces are), compactness is equivalent to sequential compactness.</p> <p><b>Step 1: Understand the Infinite-Dimensionality</b></p> <p>In an infinite-dimensional normed space, there exists an infinite sequence of vectors that are nearly "orthogonal" in the sense that they are linearly independent and maintain a certain minimal distance from each other.</p> <p><b>Step 2: Utilize Riesz's Lemma</b></p> <p>Riesz's Lemma states:</p> <p><i>In an infinite-dimensional normed space (<math>X</math>), for any proper closed subspace (<math>Y \subsetneq X</math>) and any (<math>\epsilon &gt; 0</math>), there exists (<math>x \in X</math>) with (<math>\ x\  = 1</math>) such that <math>\inf_{y \in Y} \ x - y\  \geq 1 - \epsilon</math>.</i></p> <p><b>Step 3: Construct the Sequence</b></p> <p><b>Initial Vector:</b></p> <p>Choose (<math>x_1 \in X</math>) with (<math>\ x_1\  = 1</math>).</p> <p>Since (<math>X</math>) is infinite-dimensional, such a vector exists.</p> <p><b>Inductive Step:</b></p> <p>Assume we have constructed (<math>x_1, x_2, \dots, x_{n-1}</math>) in (<math>X</math>) such that:</p> $\ x_i\  = 1 \quad \forall i$ $\ x_i - x_j\  \geq \delta > 0 \quad \forall i \neq j$ | <ul style="list-style-type: none"> <li>- Let (<math>Y_{n-1} = \text{span}\{x_1, x_2, \dots, x_{n-1}\}</math>).</li> <li>- (<math>Y_{n-1}</math>) is a finite-dimensional (hence closed) subspace of (<math>X</math>).</li> </ul> <p><b>Applying Riesz's Lemma:</b></p> <ul style="list-style-type: none"> <li>- Set (<math>\epsilon = \frac{1}{2}</math>).</li> <li>- By Riesz's Lemma, there exists (<math>x_n \in X</math>) such that: <ul style="list-style-type: none"> <li>- (<math>\ x_n\  = 1</math>).</li> <li>- (<math>\inf_{y \in Y_{n-1}} \ x_n - y\  \geq \frac{1}{2}</math>).</li> </ul> </li> <li>- Scale (<math>x_n</math>) to have norm (<math>\frac{1}{2}</math>): <ul style="list-style-type: none"> <li>- Set (<math>x_n' = \frac{1}{2}x_n</math>), so (<math>\ x_n'\  = \frac{1}{2}</math>).</li> </ul> </li> <li>- For any (<math>y \in Y_{n-1}</math>): <ul style="list-style-type: none"> <li>- (<math>\ x_n' - y\  = \frac{1}{2}\ x_n - y\  \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}</math>), where (<math>y = \frac{1}{2}x_m</math> for some <math>m \in \{1, 2, \dots, n-1\}</math>).</li> </ul> </li> <li>- This implies (<math> x_n' - x_i  \geq \frac{1}{4}</math>) for all (<math>i &lt; n</math>).</li> </ul> <p><b>Resulting Sequence:</b></p> <ul style="list-style-type: none"> <li>- The sequence (<math>\{x_n\}</math>) (where (<math>x_n = x_n'</math>)) satisfies: <ul style="list-style-type: none"> <li>- (<math>\ x_n\  = \frac{1}{2} &lt; r</math>), so (<math>x_n \in B_r(0)</math>).</li> <li>- (<math> x_n - x_m  \geq \frac{1}{4}</math>) for all (<math>n \neq m</math>).</li> </ul> </li> </ul> <p><b>Step 4: Show the Sequence Has No Convergent Subsequence</b></p> <ul style="list-style-type: none"> <li>- <b>Contradiction Argument:</b> <ul style="list-style-type: none"> <li>- Suppose (<math>\{x_{n_k}\}</math>) is a convergent subsequence of (<math>\{x_n\}</math>), converging to some (<math>x \in X</math>).</li> <li>- Then (<math>x_{n_k} \rightarrow x</math>) implies (<math>\ x_{n_k} - x\  \rightarrow 0</math>) as (<math>k \rightarrow \infty</math>).</li> <li>- However, from our construction, (<math>\ x_{n_k} - x\  \geq \frac{1}{4}</math>) for all (<math>k \neq l</math>).</li> <li>- This is a contradiction, as the distance between distinct terms of the subsequence cannot both be arbitrarily small and at least (<math>\frac{1}{4}</math>).</li> </ul> </li> </ul> <p><b>Step 5: Conclusion</b></p> <ul style="list-style-type: none"> <li>- Since every subsequence of (<math>\{x_n\}</math>) does not have a convergent subsequence, (<math>B_r(0)</math>) is not sequentially compact.</li> <li>- In metric spaces, compactness is equivalent to sequential compactness.</li> <li>- Therefore, (<math>B_r(0)</math>) is not compact.</li> </ul> <p><b>Final Answer:</b></p> <p>Therefore, open balls in infinite-dimensional normed spaces are not compact.</p> |
|--|--|

Figure 124: College-Level Math: Case 8.5. A demonstration case of o1-preview reproducing the proof of an advanced math theorem. The generated answer is manually evaluated as correct.

|   |  |
|---|--|
| <p><b>Prompt</b></p> <p>Let <math>E</math> be an infinitely dimensional Banach space, and let <math>\{x_n\}</math> be a sequence of linearly independent elements of norm 1. Show that there exists an element in the closure of the space generated by all <math>x_n</math> which does not lie in any subspace generated by a finite number of <math>x_n</math>.</p> <p><b>O1-preview</b></p> <p>To solve this problem, we aim to find an element in the closure of the linear span of the sequence <math>(x_n)</math> that is not contained in any finite-dimensional subspace generated by finitely many <math>(x_n)</math>. Here's how we can construct such an element:</p> <p><b>1. Construct an Infinite Series in <math>(E)</math>:</b></p> <p>Since <math>(E)</math> is a Banach space (a complete normed vector space), any absolutely convergent series in <math>(E)</math> converges to an element in <math>(E)</math>. Consider the series:</p> $[ s = \sum_{n=1}^{\infty} a_n x_n, ]$ <p>where <math>\{a_n\}</math> is a sequence of scalars such that <math>(\sum_{n=1}^{\infty}  a_n  &lt; \infty)</math> (i.e., <math>\{a_n\} \in \ell^1</math>).</p> <p><b>2. Ensure the Series Converges in <math>(E)</math>:</b></p> <p>Because each <math>(x_n)</math> has norm 1, the partial sums satisfy:</p> | $[ \left  \sum_{k=1}^n a_k x_k \right  \leq \sum_{k=1}^n  a_k  ]$ <p>Since <math>(\sum_{n=1}^{\infty}  a_n  &lt; \infty)</math>, the series <math>(\sum_{n=1}^{\infty} a_n x_n)</math> converges in <math>(E)</math> due to the completeness of <math>(E)</math>.</p> <p><b>3. Show the Limit is Not in Any Finite-Dimensional Subspace:</b></p> <p>The limit <math>(s)</math> lies in the closure of the linear span of <math>(x_n)</math>, denoted by <math>(\overline{\text{span}\{x_n\}})</math>. However, <math>(s)</math> cannot be expressed as a finite linear combination of the <math>(x_n)</math> unless almost all <math>(a_n = 0)</math>, which is not the case since <math>(\{a_n\})</math> is an infinite sequence in <math>(\ell^1)</math>.</p> <p><b>4. Conclude the Existence of Such an Element:</b></p> <p>Thus, <math>(s)</math> is an element in <math>(\overline{\text{span}\{x_n\}})</math> that does not belong to any finite-dimensional subspace generated by finitely many <math>(x_n)</math>. This demonstrates that such an element exists.</p> <p><b>Answer:</b></p> <p>Yes; for example, any infinite sum <math>s = \sum_{n=1}^{\infty} a_n x_n</math> with <math>\sum_{n=1}^{\infty}  a_n  &lt; \infty</math> converges in the Banach space to an element in the closure of the span of the <math>x_n</math>, but <math>s</math> is not in any finite-dimensional subspace spanned by finitely many <math>x_n</math>.</p> |
|---|--|

Figure 125: **College-Level Math: Case 9.** A demonstration case of o1-preview reproducing the proof of an advanced math theorem. The generated answer is manually evaluated as incorrect.

|   |   |
|---|---|
| <p><b>Prompt</b></p> <p>Prove Hilbert's Nullstellensatz. You can assume the following form of weak Nullstellensatz: Let <math>k</math> be an algebraically closed field, then the maximal ideals of <math>k[x_1, \dots, x_n]</math> are exactly of the form <math>(x_1 - a_1, x_2 - a_2, \dots, x_n - a_n)</math> where <math>a_i \in k</math>.</p> <p><b>O1-preview</b></p> <p>To prove Hilbert's Nullstellensatz, we will show that for any ideal <math>(I)</math> in the polynomial ring <math>(k[x_1, \dots, x_n])</math> over an algebraically closed field <math>(k)</math>, the radical of <math>(I)</math> equals the ideal of all polynomials that vanish on the zero set of <math>(I)</math>, i.e., <math>\sqrt{I} = I(V(I))</math>, where <math>V(I) = \{(a_1, \dots, a_n) \in k^n \mid f(a_1, \dots, a_n) = 0 \text{ for all } f \in I\}</math> and <math>I(V(I)) = \{f \in k[x_1, \dots, x_n] \mid f(a) = 0 \text{ for all } a \in V(I)\}</math>.</p> <p><b>Assumption (Weak Nullstellensatz):</b> The maximal ideals of <math>(k[x_1, \dots, x_n])</math> are precisely the ideals of the form <math>(x_1 - a_1, \dots, x_n - a_n)</math> with <math>(a_i \in k)</math>.</p> <p><b>Proof:</b></p> <p><b>Step 1: Show that <math>(\sqrt{I}) \subseteq I(V(I))</math>.</b></p> <p>Since <math>(\sqrt{I} \subseteq I(V(I)))</math> (every polynomial in <math>(I)</math> vanishes on <math>(V(I))</math> by definition), and <math>(I(V(I)))</math> is a radical ideal (because the set of common zeros of <math>(I)</math> and its radical are the same), it follows that: <math>\sqrt{I} \subseteq I(V(I))</math>.</p> <p><b>Step 2: Show that <math>(I(V(I))) \subseteq \sqrt{I}</math>.</b></p> <p>Let <math>(f \in I(V(I)))</math>. We need to show that <math>(f \in \sqrt{I})</math>, i.e., there exists an integer <math>(r \geq 1)</math> such that <math>(f^r \in I)</math>.</p> | <p><b>Assume, for contradiction,</b> that <math>(f \notin \sqrt{I})</math>. Then <math>(f \notin \bigcap_{m \in \mathfrak{M}} \supseteq I_m)</math>, where the intersection is over all maximal ideals <math>(\mathfrak{m})</math> containing <math>(I)</math>.</p> <p><b>Using the assumption,</b> there exists a maximal ideal <math>(\mathfrak{m}) = (x_1 - a_1, \dots, x_n - a_n)</math> such that:</p> <ul style="list-style-type: none"> <li>- <math>(I \subsetneq \mathfrak{m})</math>,</li> <li>- <math>(f \notin \mathfrak{m})</math>.</li> </ul> <p>This implies:</p> <ul style="list-style-type: none"> <li>- <math>((a_1, \dots, a_n) \in V(I))</math> because all polynomials in <math>(I)</math> vanish at <math>(a)</math>.</li> <li>- <math>(f(a) \neq 0)</math> because <math>(f \notin \mathfrak{m})</math>.</li> </ul> <p><b>This contradicts</b> the assumption that <math>(f \in I(V(I)))</math> (since <math>(f)</math> should vanish at all points of <math>(V(I))</math>).</p> <p><b>Therefore,</b> our assumption is false, and <math>(\sqrt{I} = I(V(I)))</math>. Hence, <math>(I(V(I))) = \sqrt{I}</math>.</p> <p><b>Conclusion:</b></p> <p>Combining both inclusions, we have: <math>\sqrt{I} = I(V(I))</math>.</p> <p>This equality is precisely Hilbert's Nullstellensatz, showing that the radical of an ideal <math>(I)</math> equals the ideal of polynomials vanishing on the zero set of <math>(I)</math>.</p> <p><b>Therefore, Hilbert's Nullstellensatz is proved under the given assumption.</b></p> <p><b>Answer:</b> By showing that for any ideal <math>I</math>, its radical equals the polynomials vanishing on <math>V(I)</math>—that is, <math>\sqrt{I} = I(V(I))</math></p> |
|---|---|

Figure 126: College-Level Math: Case 10. A demonstration case of o1-preview reproducing the proof of an advanced math theorem. The generated answer is manually evaluated as correct.

|  |   |
|--|---|
| <p><b>Prompt</b></p> <p>Prove Hilbert's Nullstellensatz. You can assume the following form of weak Nullstellensatz: Let <math>\mathbb{k}</math> be an algebraically closed field. If <math>I</math> is an ideal of <math>\mathbb{k}[x_1, \dots, x_n]</math> such that <math>V(I)</math> is empty, then <math>I = (1)</math>.</p> <p><b>O1-preview</b></p> <p>To prove Hilbert's Nullstellensatz using the given version of the weak Nullstellensatz, we'll show that for any ideal <math>(I)</math> in the polynomial ring <math>(k[x_1, \dots, x_n])</math> over an algebraically closed field <math>(k)</math>, the radical of <math>(I)</math> equals the ideal of polynomials that vanish on the zero set of <math>(I)</math>. Specifically, we'll prove:</p> <pre>[ \sqrt{I} = I(V(I)), ]</pre> <p>Where:</p> <ul style="list-style-type: none"> <li>- <math>(V(I))</math> is the set of common zeros of <math>(I)</math> in <math>(k^n)</math>: <math>V(I) = \{ a \in k^n \mid f(a) = 0 \text{ for all } f \in I \}</math>.</li> <li>- <math>(I(V(I)))</math> is the ideal of all polynomials in <math>(k[x_1, \dots, x_n])</math> that vanish on <math>(V(I))</math>: <math>I(V(I)) = \{ f \in k[x_1, \dots, x_n] \mid f(a) = 0 \text{ for all } a \in V(I) \}</math>.</li> </ul> <p><b>Assumption (Weak Nullstellensatz):</b> If <math>(I)</math> is an ideal of <math>(k[x_1, \dots, x_n])</math> such that <math>(V(I)) = \emptyset</math>, then <math>(I) = (1)</math>.</p> <p><b>Proof:</b></p> <p><b>Step 1: Show that <math>(\sqrt{I}) \subseteq I(V(I))</math>.</b></p> <p>Every polynomial <math>(f \in \sqrt{I})</math> satisfies <math>(f^r \in I)</math> for some <math>(r \geq 1)</math>. Since <math>(f^r \in I)</math>, and every polynomial in <math>(I)</math> vanishes on <math>(V(I))</math>, it follows that <math>(f^r(a) = 0)</math> for all <math>(a \in V(I))</math>. Consequently, <math>(f(a) = 0)</math> for all <math>(a \in V(I))</math> (since <math>(k)</math> is algebraically closed and has no nilpotent elements). Therefore, <math>(f \in I(V(I)))</math>, and thus:</p> <pre>[ \sqrt{I} \subseteq I(V(I)). ]</pre> <p><b>Step 2: Show that <math>(I(V(I))) \subseteq \sqrt{I}</math>.</b></p> <p>Let <math>(f \in I(V(I)))</math>. We need to show that <math>(f \in \sqrt{I})</math>, meaning there exists <math>(r \geq 1)</math> such that <math>(f^r \in I)</math>.</p> <p><b>Assume, for contradiction,</b> that <math>(f \notin \sqrt{I})</math>. This means <math>(f^r \notin I)</math> for all <math>(r \geq 1)</math>.</p> <p><b>Construct a New Ideal:</b></p> <p>Consider the extended polynomial ring <math>(k[x_1, \dots, x_n, t])</math>, where <math>(t)</math> is a new variable. Define the ideal:</p> <pre>[ J = I \cdot k[x_1, \dots, x_n, t] + (1 - f t). ]</pre> <p><b>Claim:</b> The variety <math>(V(J))</math> in <math>(k^{n+1})</math> is empty.</p> <p><b>Proof of Claim:</b></p> <p>Suppose there exists a point <math>(a, b) \in k^{n+1}</math> such that:</p> <ul style="list-style-type: none"> <li>- <math>(f_i(a) = 0)</math> for all <math>(f_i \in I)</math>,</li> <li>- <math>(1 - f(a)b = 0)</math>.</li> </ul> | <p>Since <math>(f \in I(V(I)))</math>, <math>(f(a) = 0)</math> for all <math>(a \in V(I))</math>. Therefore, <math>(f(a) = 0)</math>, and the equation <math>(1 - f(a)b = 0)</math> simplifies to <math>(1 - 0 \cdot b = 1 = 0)</math>, which is a contradiction. Hence, <math>(V(J)) = \emptyset</math>.</p> <p><b>Apply the Weak Nullstellensatz:</b></p> <p>By the assumption, since <math>(V(J)) = \emptyset</math>, the ideal <math>(J)</math> must be the entire ring:</p> <pre>[ J = (1) \quad \text{in} \quad k[x_1, \dots, x_n, t]. ]</pre> <p><b>Express <math>(J)</math> in Terms of Generators of <math>(J)</math>:</b></p> <p>There exist polynomials <math>(g_i(x, t) \in k[x_1, \dots, x_n, t])</math> and <math>(h(x, t) \in k[x_1, \dots, x_n, t])</math> such that:</p> <pre>[ 1 = \sum_i g_i(x, t) f_i(x) + h(x, t)(1 - f(x)t). ]</pre> <p>Here, <math>(f_i(x) \in I)</math>.</p> <p><b>Specialize <math>(t)</math> to <math>(t = \frac{1}{f(x)})</math>:</b></p> <p>When <math>(f(x) \neq 0)</math>, set <math>(t = \frac{1}{f(x)})</math>. Substituting into the equation:</p> <pre>[ 1 = \sum_i g_i(x, \frac{1}{f(x)}) f_i(x) + h(x, \frac{1}{f(x)})(1 - f(x) \cdot \frac{1}{f(x)}). ]</pre> <p>Simplify the right-hand side:</p> <pre>[ 1 = \sum_i g_i(x, \frac{1}{f(x)}) f_i(x) + h(x, \frac{1}{f(x)})(1 - 1) = \sum_i g_i(x, \frac{1}{f(x)}) f_i(x). ]</pre> <p>Since <math>(f_i(x) \in I)</math> and <math>(f_i(x) = 0)</math> for all <math>(a \in V(I))</math>, the right-hand side evaluates to zero at any point <math>(a)</math> where <math>(f(a) \neq 0)</math>. Therefore, we have:</p> <pre>[ 1 = 0 \quad \text{at these points}. ]</pre> <p>which is a contradiction. Hence, <math>(f(x) = 0)</math> everywhere in <math>(V(I))</math>, which implies that <math>(f)</math> is nilpotent modulo <math>(I)</math>.</p> <p><b>Conclusion:</b></p> <p>There must exist an integer <math>(r \geq 1)</math> such that <math>(f^r \in I)</math>, meaning <math>(f \in \sqrt{I})</math>. Thus:</p> <pre>[ I(V(I)) \subseteq \sqrt{I}. ]</pre> <p><b>Final Equality:</b></p> <p>Combining both inclusions:</p> <pre>[ \sqrt{I} = I(V(I)). ]</pre> <p>This equality is Hilbert's Nullstellensatz, showing that the radical of an ideal <math>(I)</math> is precisely the set of polynomials that vanish on the common zeros of <math>(I)</math>.</p> <p><b>Answer:</b> By showing that the radical of any ideal equals the polynomials vanishing on its zero set—that is, <math>\sqrt{I} = I(V(I))</math>—we prove Hilbert's Nullstellensatz using the given weak Nullstellensatz.</p> |
|--|---|

Figure 127: College-Level Math: Case 10.5. A demonstration case of o1-preview reproducing the proof of an advanced math theorem. The generated answer is manually evaluated as correct.

## A.11 Electronic Health Record Diagnosis

### Prompt

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Cutaneous lesions of disseminated histoplasmosis in human immunodeficiency virus-infected patients. Disseminated histoplasmosis is being diagnosed more frequently in persons infected with the human immunodeficiency virus and is often the initial manifestation of the acquired immunodeficiency syndrome (AIDS). Disease-related cutaneous features of HIV-associated disseminated histoplasmosis are defined as mucocutaneous lesions from which fungal organisms were either cultured or demonstrated histopathologically. We report four HIV-seropositive patients with disseminated histoplasmosis who had culture-positive skin or oral lesions of histoplasmosis and review the specific cutaneous manifestations of HIV-associated disseminated histoplasmosis.

Including our patients, disease-related skin and/or mucosal lesions were present in 11% of patients (26% of 239) with HIV-associated disseminated histoplasmosis.

The possibility of disseminated histoplasmosis should be considered in all HIV-infected persons and in persons with AIDS risk factors who have fever, weight loss, hepatosplenomegaly, and new cutaneous lesions.

An early skin or mucosal biopsy specimen for crushed tissue preparation, histologic evaluation, and fungal culture is a simple, rapid diagnostic procedure.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### O1-preview

Based on the medical abstract provided, the patient's condition falls under the category of **Bacterial Infections and Mycoses**.

Explanation:

Disseminated Histoplasmosis:

The abstract focuses on disseminated histoplasmosis in HIV-infected patients.

Histoplasmosis is an infection caused by the fungus *Histoplasma capsulatum*.

It is a type of mycosis, which refers to diseases caused by fungi.

Cutaneous Lesions:

The patients exhibited cutaneous (skin) and mucosal lesions from which the fungal organisms were cultured or identified histopathologically.

While these are skin manifestations, they are a result of a systemic fungal infection rather than a primary skin disease.

Association with HIV/AIDS:

The patients are immunocompromised due to HIV infection, making them more susceptible to opportunistic infections like histoplasmosis.

Although HIV is a viral disease, the primary focus here is on the fungal infection resulting from immunosuppression.

Conclusion:

Given that histoplasmosis is a fungal infection and the primary subject of the abstract, the most appropriate category is **Bacterial Infections and Mycoses**.

### Reference Answer

**Virus Diseases**

Figure 128: **Electronic Health Record Diagnosis: Case 5.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Lack of effect of clonidine and pentoxifylline in short-term therapy of diabetic peripheral neuropathy. The goal of this study was to confirm or rule out anecdotal reports of beneficial effects of clonidine and pentoxifylline in the treatment of painful diabetic peripheral neuropathy. Clonidine was administered to 16 subjects at two dosage levels (0.1 and 0.2 mg/day) and was compared to placebo in a crossover design, with each phase lasting 4 wk. Either pentoxifylline (400 mg 3 times/day) or placebo was given to 21 subjects in a 12-wk trial. Discomfort was characterized and rated with a subjective pain score (range 0-20). There was a significant decrease in pain score from baseline with both active drugs ( $P$  less than 0.05), but this was no better than the response to placebo ( $P$  less than 0.30 for clonidine and  $P$  less than 0.95 for pentoxifylline). This study does not demonstrate a short-term benefit of either clonidine or pentoxifylline in the treatment of peripheral neuropathy.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Nervous System Diseases**.

Explanation:

Diabetic Peripheral Neuropathy:

The abstract focuses on painful diabetic peripheral neuropathy, which is a common complication of diabetes. Peripheral neuropathy involves damage to the peripheral nerves, leading to symptoms like pain, tingling, and numbness, primarily in the hands and feet.

This condition results from prolonged high blood sugar levels causing nerve damage.

Relevance to the Category:

Peripheral neuropathy is a disorder of the nervous system because it affects nerve function.

While diabetes is an Endocrine Disease, the primary focus here is on the neurological complications arising from it. Therefore, the condition is best classified under Nervous System Diseases.

Conclusion:

Given that the patient's primary issue is diabetic peripheral neuropathy—a neurological disorder—the most appropriate category is **Nervous System Diseases**.

### **Reference Answer**

**Endocrine Diseases**

Figure 129: **Electronic Health Record Diagnosis: Case 6.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Treatment of Fournier's gangrene with adjunctive hyperbaric oxygen therapy.  
Fournier's gangrene is a devastating infection and often is associated with a high morbidity and mortality.  
Surgical debridement and antibiotics are the cornerstones of therapy.  
This case describes the use of hyperbaric oxygen as an adjunct in the treatment of Fournier's gangrene.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Bacterial Infections and Mycoses**.

Explanation:

Fournier's Gangrene:

Fournier's gangrene is a severe, rapidly progressing bacterial infection of the perineal, genital, or perianal regions. It is a form of necrotizing fasciitis, involving the destruction of skin, subcutaneous tissue, and fascia. The condition is typically polymicrobial, involving a mix of aerobic and anaerobic bacteria.

Clinical Features:

The infection leads to high morbidity and mortality if not promptly treated. Symptoms include severe pain, swelling, and discoloration of the affected area. Rapid surgical debridement and broad-spectrum antibiotics are essential for management. Hyperbaric oxygen therapy is mentioned as an adjunctive treatment, which can inhibit anaerobic bacterial growth and enhance wound healing.

Relevance to the Category:

As a severe bacterial infection, Fournier's gangrene is best classified under Bacterial Infections and Mycoses. While it affects the skin and genital regions, its primary concern is the infectious process caused by bacteria. The involvement of multiple bacterial species and the need for antibiotic therapy emphasize its classification as a bacterial infection.

Conclusion:

Given that Fournier's gangrene is a life-threatening bacterial infection requiring urgent medical intervention, the most appropriate category for this patient's condition is **Bacterial Infections and Mycoses**.

### **Reference Answer**

**Pathological Conditions, Signs and Symptoms**

Figure 130: **Electronic Health Record Diagnosis: Case 7.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Factor XIII deficiency and intracranial hemorrhages in infancy.

We report an infant with Factor XIII deficiency who had 2 seemingly spontaneous intracranial hemorrhages. It is important to consider Factor XIII deficiency as a possible cause of unexplained intracranial hemorrhages in infancy.

Ongoing factor replacement therapy is recommended to prevent further bleeding episodes.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Hemic and Lymphatic Diseases**.

Explanation:

Factor XIII Deficiency:

Factor XIII deficiency is a rare bleeding disorder caused by the lack or malfunction of clotting factor XIII.

Factor XIII is crucial for the final step of the blood coagulation cascade, stabilizing the formation of a blood clot.

Deficiency leads to a tendency for severe bleeding, including spontaneous intracranial hemorrhages in infants.

Hemic and Lymphatic Diseases:

The term "hemic" refers to the blood, and "lymphatic" pertains to the lymphatic system.

This category includes disorders affecting the blood's components, such as red cells, white cells, platelets, and coagulation factors.

Examples include anemia, hemophilia, thrombocytopenia, and other coagulation factor deficiencies.

Relevance to the Patient's Condition:

The infant's spontaneous intracranial hemorrhages are a direct result of the blood clotting disorder caused by Factor XIII deficiency.

While intracranial hemorrhage affects the nervous system, in this case, it is a secondary manifestation of a primary blood disorder.

Conclusion:

Given that Factor XIII deficiency is a disorder of the blood's coagulation system, the most appropriate category for this patient's condition is **Hemic and Lymphatic Diseases**.

### **Reference Answer**

**Hemic and Lymphatic Diseases**

Figure 131: **Electronic Health Record Diagnosis: Case 8.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Effects of haem arginate on variegate porphyria.

Four patients with variegate porphyria (VP) were treated with repeated haem arginate infusions daily for 4 days and then weekly for 4 weeks.

After the initial four daily doses of haem arginate (haem 3 mg/kg), the excretion of faecal protoporphyrin (mean 579 nmol/g dry wt) fell to an almost normal level (mean 123 nmol/g dry wt), and that of coproporphyrin (mean 162 nmol/g dry wt) to the normal level (mean 21 nmol/g dry wt) in all patients.

However, during the period of the four weekly infusions of haem the excretion of porphyrins increased almost to the pretreatment level.

Phototesting showed no changes in the photoreactivity of the skin, and no improvement in skin lesions was seen during the treatment.

Except for one case of thrombophlebitis no side-effects occurred.

In a child with homozygous VP, four daily infusions of haem arginate (2 mg/kg) normalized the faecal protoporphyrin content, but had no effect on the increased erythrocyte protoporphyrin concentration.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Nutritional and Metabolic Diseases**.

Explanation:

Variegate Porphyria (VP):

Variegate porphyria is a type of porphyria, which is a group of disorders caused by abnormalities in the heme biosynthesis pathway. It results from a deficiency of the enzyme protoporphyrinogen oxidase, leading to the accumulation of porphyrin precursors. These accumulated substances can cause various symptoms, including skin sensitivity to light (photosensitivity) and neurological complications.

Metabolic Disorder:

Porphyrias are considered metabolic diseases because they involve disruptions in the body's normal metabolic processes—in this case, the synthesis of heme, an essential component of hemoglobin.

The treatment discussed involves administering haem arginate to reduce the accumulation of porphyrin precursors, directly addressing the metabolic defect.

Relevance to the Category:

While variegate porphyria can have skin manifestations and occasionally neurological symptoms, its primary cause is a metabolic enzyme deficiency.

Therefore, it is best classified under Nutritional and Metabolic Diseases rather than skin or nervous system diseases.

Conclusion:

Given that variegate porphyria is a metabolic disorder affecting heme synthesis, the most appropriate category for this patient's condition is **Nutritional and Metabolic Diseases**.

### **Reference Answer**

**Skin and Connective Tissue Diseases**

Figure 132: **Electronic Health Record Diagnosis: Case 9.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

### **Prompt**

You are a professional cardiovascular diseases doctor. Here is a patient's medical abstract:

Rapidly progressive outer retinal necrosis in the acquired immunodeficiency syndrome  
Two patients, both seropositive for the human immunodeficiency virus, developed rapidly progressive retinal necrosis associated with a systemic herpes zoster infection.  
The retinitis in these patients was characterized by primary involvement of the outer retina, with sparing of the inner retina and retinal vasculature until late in the disease process; a rapidly progressive course; poor response to intravenous acyclovir; and development of rhegmatogenous retinal detachment.  
In one of the patients, the retinitis was initially multifocal.  
Electron microscopy of a retinal biopsy specimen from one of the patients demonstrated virus particles consistent with a herpesvirus, and polymerase chain reaction disclosed herpesvirus in a retinal biopsy specimen of the other patient.  
This entity may represent a distinct form of acute retinal necrosis that is seen in immunocompromised individuals.

There are 23 cardiovascular diseases categories in total: Bacterial Infections and Mycoses; Virus Diseases; Parasitic Diseases; Neoplasms; Musculoskeletal Diseases; Digestive System Diseases; Stomatognathic Diseases; Respiratory Tract Diseases; Otorhinolaryngologic Diseases; Nervous System Diseases; Eye Diseases; Urologic and Male Genital Diseases; Female Genital Diseases and Pregnancy Complications; Cardiovascular Diseases; Hemic and Lymphatic Diseases; Neonatal Diseases and Abnormalities; Skin and Connective Tissue Diseases; Nutritional and Metabolic Diseases; Endocrine Diseases; Immunologic Diseases; Disorders of Environmental Origin; Animal Diseases; Pathological Conditions, Signs and Symptoms.

Could you please tell me what cardiovascular disease this patient category is?

### **O1-preview**

Based on the medical abstract provided, the patient's condition falls under the category of **Eye Diseases**.

Explanation:

Rapidly Progressive Outer Retinal Necrosis:

The abstract discusses two patients who developed rapidly progressive retinal necrosis.

Retinal necrosis refers to the death of retinal tissue, leading to severe vision loss and potential retinal detachment.

The disease primarily affects the outer retina, with sparing of the inner retina and retinal vasculature until late stages.

Association with Herpes Zoster Virus:

The retinal necrosis is associated with a systemic herpes zoster infection. Herpes zoster, caused by the varicella-zoster virus, can reactivate in immunocompromised individuals, leading to severe ocular complications.

While the infection is viral, the primary clinical manifestation is in the eye.

Immunocompromised State (HIV/AIDS):

Both patients are seropositive for the human immunodeficiency virus (HIV). Immunosuppression increases susceptibility to opportunistic infections affecting the eye.

Relevance to the Category:

The main issue is the ocular manifestation of a systemic viral infection. The condition leads to significant eye pathology, including retinitis and retinal detachment. While Virus Diseases is a relevant category due to the herpesvirus involvement, the most specific classification pertains to the eye.

Conclusion:

Given that the primary clinical concern is severe retinal damage leading to vision loss, the most appropriate category for this patient's condition is **Eye Diseases**.

### **Reference Answer**

**Eye Diseases**

Figure 133: **Electronic Health Record Diagnosis: Case 10.** One specific example of a demonstration cases for EHR diagnosis involves the use of the o1-preview model. Red indicates that the model's output is completely inconsistent with the reference diagnosis, while green signifies complete consistency.

## A.12 Stochastic Processes in Statistics

### Prompt

Here are some questions about stochastic process, please answer step by step.

I have 4 umbrellas, some at home, some in the office. I keep moving between home and office. I take an umbrella with me only if it rains. If it does not rain, I leave the umbrella behind (at home or in the office). It may happen that all umbrellas are in one place, I am at the other, it starts raining, and I must leave, so I get wet.

If the probability of rain is  $p$ , what is the probability that I get wet?

### O1-preview

Due to the complexity of analytically solving for the stationary distribution in general, we can make an important observation:

- **Symmetry:** Because the system is symmetric between home and office, over the long run, each state  $S_k$  has an equal probability of occurring.
- **Uniform Distribution:** Therefore, the stationary probability  $\pi_k$  of being in state  $S_k$  is  $\pi_k = \frac{1}{N+1}$ .

#### Probability of Getting Wet

You get wet only when: - You are in state  $S_0$  (no umbrellas at your location), **and**  
- It rains (probability  $p$ ).

Therefore, the probability  $P_{\text{wet}}$  that you get wet is:

$$P_{\text{wet}} = \pi_0 \times p = \frac{1}{N+1} \times p = \frac{p}{N+1}.$$

For  $N = 4$ :

$$P_{\text{wet}} = \frac{p}{4+1} = \frac{p}{5}.$$

### Reference Answer

$$P_{\text{wet}} = \frac{qp}{q+1}, q = 1 - p$$

Figure 134: **Stochastic Processes in Statistics: Case 3.** A demonstration case of Stochastic Process problem solving. The o1-preview successfully solved the problem.

## Prompt

Here are some questions about stochastic process, please answer step by step.

Smith is in jail and has 3 dollars; he can get out on bail if he has 8 dollars. A guard agrees to make a series of bets with him. If Smith bets  $A$  dollars, he wins  $A$  dollars with probability 0.4 and loses  $A$  dollars with probability 0.6.

1. **Find the probability** that he wins 8 dollars before losing all of his money if:

- He bets 1 dollar each time (**timid strategy**).
- He bets, each time, as much as possible but not more than necessary to bring his fortune up to 8 dollars (**bold strategy**).

2. **Which strategy** gives Smith the better chance of getting out of jail?

## O1-preview

### Problem 1(a):

The probability  $P_{\text{win}}$  that Smith reaches \$8 before losing all his money is:

$$P_{\text{win}} = \frac{r^x - 1}{r^N - 1} = \frac{\left(\frac{3}{2}\right)^3 - 1}{\left(\frac{3}{2}\right)^8 - 1} = \frac{\frac{27}{8} - 1}{\frac{6561}{256} - 1} = \frac{\frac{19}{8}}{\frac{6305}{256}} = \frac{19 \times 256}{8 \times 6305} = \frac{608}{6305}$$

So, the probability is:

$$P_{\text{win}} = \frac{608}{6305} \approx 0.0965$$

### Problem 1(b):

Using these transitions, we set up equations for the probability of reaching \$8 from each state: -  $P_4 = 0.4 \times 1 + 0.6 \times 0 = 0.4$  -  $P_6 = 0.4 \times 1 + 0.6 \times P_4 = 0.4 + 0.6 \times 0.4 = 0.64$  -  $P_3 = 0.4 \times P_6 = 0.4 \times 0.64 = 0.256$

So, the probability is:

$$P_{\text{win}} = P_3 = \frac{32}{125} \approx 0.256$$

### Problem 2:

Since  $0.256 > 0.0965$ , the bold strategy gives Smith a better chance of getting out of jail.

## Reference Answer

0.0964 0.256

The bold strategy gives Smith a better chance of getting out of jail.

Figure 135: Stochastic Processes in Statistics: Case 4. A demonstration case of Stochastic Process problem solving. The o1-preview successfully solved the problem.

## Prompt

Here are some questions about stochastic process, please answer step by step.

A Markov chain with state space {1,2,3} has transition probability matrix:

$$P = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{3} & \frac{3}{3} & \frac{3}{3} \\ 0 & 1 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}$$

**Starting from state 1, find the expected time until absorption occurs.**

## O1-preview

Let's denote  $m_i$  as the expected number of steps until absorption when starting from state  $i$ . Since state 3 is absorbing,  $m_3 = 0$ .

We can set up equations for  $m_1$  and  $m_2$  using the law of total expectation (first-step analysis):

1. **For state 1:**

$$\begin{aligned} m_1 &= 1 + \sum_{j=1}^3 P_{1,j} \cdot m_j \\ &= 1 + \left(\frac{1}{3}\right)m_1 + \left(\frac{1}{3}\right)m_2 \end{aligned}$$

2. **For state 2:**

$$\begin{aligned} m_2 &= 1 + \sum_{j=1}^3 P_{2,j} \cdot m_j \\ &= 1 + \left(\frac{1}{2}\right)m_2 \\ m_2 &= 2 \end{aligned}$$

Next, substitute  $m_2 = 2$  into the equation for  $m_1$ :

$$m_1 = 2.5$$

## Reference Answer

2.5

Figure 136: Stochastic Processes in Statistics: Case 5. A demonstration case of Stochastic Process problem solving. The o1-preview successfully solved the problem.

## Prompt

Which is greater, 9.11 or 9.8?

## O1-preview

9.8 is greater than 9.11.

**Specifically:**

- 9.8 can be represented as 9.80.
- When comparing 9.80 and 9.11, the integer parts are both 9, so we compare the decimal parts.
- Here, 0.80 is greater than 0.11.

**Therefore, 9.8 is greater than 9.11.**

## Reference Answer

9.8 is greater than 9.11

Figure 137: Stochastic Processes in Statistics: Case 6. A demonstration case of tricky problem. The o1-preview successfully solved the problem.

## A.13 Medical Text Anonymization

### Prompt

Please anonymize the following text. Replace all the private information with the term "[redacted]".:

### O1-preview

Record date: [redacted] Silver Ridge Internal Medicine [redacted] [redacted] [redacted] [redacted]  
[redacted]

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is 61 years old with previous history of diabetes, as well as bilateral fem-pop surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to L5-S1 epidural steroid injection once, about 3 months ago. He states that he is reasonably symptom free but the pain is coming back. Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.

PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A 17-gauge, Weiss needle used to access epidural space at L5-S1. Three cc of Depo-Medrol and 3 cc of 0.5% lidocaine were allowed to drip in via gravity. No complications. No heme, no CSF.

No paresthesia.

FOLLOW-UP PLAN: The patient to report for further injection in case needed in the next 2 or 3 months.

Hardy Ostrowski, M.D.

Dictated By: Hardy Ostrowski

eScription document: [redacted]

DD: [redacted] DT: [redacted] DV: [redacted]

### Reference Answer

Record date: 2086-12-08 Silver Ridge Internal Medicine 3 Eaton Place OROZCO, KYLE

Mansfield, NV 29473 56040785 ----- 12/08/86

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is 61 years old with previous history of diabetes, as well as bilateral fem-pop surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to L5-S1 epidural steroid injection once, about 3 months ago. He states that he is reasonably symptom free but the pain is coming back. Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.

PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A 17-gauge, Weiss needle used to access epidural space at L5-S1. Three cc of Depo-Medrol and 3 cc of 0.5% lidocaine were allowed to drip in via gravity. No complications. No heme, no CSF. No paresthesia.

FOLLOW-UP PLAN: The patient to report for further injection in case needed in the next 2 or 3 months.

Hardy Ostrowski, M.D.

Dictated By: Hardy Ostrowski

eScription document: 1-1277442 EMSSten Tel

DD: 12/08/86 DT: 12/09/86 DV: 12/08/86

Figure 138: **Medical Text Anonymization: Case 4.** An example showing the capability of o1-preview to anonymize medical texts with the implicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### **Prompt**

What are the names? Who are the professions? What are the locations? What are ages? What are dates? What are contacts? What are IDs? What are the phone numbers? Replace the matching characters with [redacted]:

### **O1-preview**

Record date: [redacted][redacted][redacted] [redacted][redacted]

[redacted]----- [redacted]

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is [redacted] years old with previous history of diabetes, as well as bilateral fem-pop surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to L5-S1 epidural steroid injection once, about [redacted] months ago. He states that he is reasonably symptom free but the pain is coming back.

Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.

PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A 17-gauge, Weiss needle used to access epidural space at L5-S1. Three cc of Depo-Medrol and 3 cc of 0.5% lidocaine were allowed to drip in via gravity. No complications. No heme, no CSF. No paresthesia.

FOLLOW-UP PLAN: The patient to report for further injection incase needed in the next 2 or 3 months.

[redacted]

Dictated By: [redacted]

[redacted]

DD: [redacted] DT: [redacted] DV: [redacted]

### **Reference Answer**

Record date: 2086-12-08 Silver Ridge Internal Medicine 3 Eaton Place OROZCO, KYLE

Mansfield, NV 29473 56040785 ----- 12/08/86

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is 61 years old with previous history of diabetes, as well as bilateral fem-pop surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to L5-S1 epidural steroid injection once, about 3 months ago. He states that he is reasonably symptom free but the pain is coming back. Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.

PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A 17-gauge, Weiss needle used to access epidural space at L5-S1. Three cc of Depo-Medrol and 3 cc of 0.5% lidocaine were allowed to drip in via gravity. No complications. No heme, no CSF. No paresthesia.

FOLLOW-UP PLAN: The patient to report for further injection in case needed in the next 2 or 3 months.

Hardy Ostrowski, M.D.

Dictated By: Hardy Ostrowski

eScription document: 1-1277442 EMSSten Tel

DD: 12/08/86 DT: 12/09/86 DV: 12/08/86

Figure 139: **Medical Text Anonymization: Case 5.** An example showing the capability of o1-preview to anonymize medical texts with the general prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### **Prompt**

Please anonymize the following clinical note. Specifically, replace all the following information with the term "[redacted]": redact any strings that might be a name or acronym or initial, redact any strings separated by the "V" symbol, redact patients' names, doctors' names and the strings in front of M.D. or after Dr., redact pager names and medical staff names, redact any strings that look like something years old or age 37, redact any dates and IDs and numbers and record dates, redact locations and addresses and clinic names, redact professions and ages and contacts, redact any acronyms and initials.:

### **O1-preview**

[redacted] [redacted]; [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

[redacted]

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is [redacted] years old with previous history of diabetes, as well as bilateral [redacted] surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to [redacted] epidural steroid injection once, about [redacted] months ago. He states that he is reasonably symptom-free but the pain is coming back. Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.  
PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A [redacted]-gauge, [redacted] needle used to access epidural space at [redacted]. [redacted] cc of Depo-Medrol and [redacted] cc of [redacted] % lidocaine were allowed to drip in via gravity. No complications. No heme, no [redacted]. No paresthesia. FOLLOW-UP PLAN: The patient to report for further injection in case needed in the next [redacted] or [redacted] months.

[redacted], [redacted]

Dictated By: [redacted][redacted] [redacted];[redacted] [redacted]

[redacted]; [redacted] [redacted]; [redacted] [redacted]; [redacted]

### **Reference Answer**

Record date: 2086-12-08 Silver Ridge Internal Medicine 3 Eaton Place OROZCO, KYLE

Mansfield, NV 29473 56040785 ----- 12/08/86

INTERVAL HISTORY: The patient is status post one lumbar epidural injection. He is 61 years old with previous history of diabetes, as well as bilateral fem-pop surgery. He has pain radiating to both legs up to calves and thighs. He has moderate-to-severe spinal stenosis. He had good response to L5-S1 epidural steroid injection once, about 3 months ago. He states that he is reasonably symptom free but the pain is coming back. Today, we are going to repeat the injection.

PHYSICAL EXAMINATION: Essentially unchanged from prior.

PROCEDURE NOTE: After obtaining informed consent, the patient placed in sitting position. Lower back prepped and draped in usual sterile fashion. A 17-gauge, Weiss needle used to access epidural space at L5-S1. Three cc of Depo-Medrol and 3 cc of 0.5% lidocaine were allowed to drip in via gravity. No complications. No heme, no CSF. No paresthesia.

FOLLOW-UP PLAN: The patient to report for further injection in case needed in the next 2 or 3 months.

Hardy Ostrowski, M.D.

Dictated By: Hardy Ostrowski eScription document:1-1277442 EMSSten Tel

DD: 12/08/86 DT: 12/09/86 DV: 12/08/86

Figure 140: **Medical Text Anonymization: Case 6.** An example showing the capability of o1-preview to anonymize medical texts with the explicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### **Prompt**

Please anonymize the following text. Replace all the private information with the term "[redacted]".:

### **O1-preview**

Record date: [redacted]

NAME: [redacted]

MRN: [redacted]

S: Here for follow-up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with ADLs. He says he uses Viagra occasionally and it works for him.

O: Well-appearing male. Blood pressure 142/74. Chest clear. COR: No murmur.

A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check FLP.(5) Worried about dementia. Will check B12, TSH, RPR for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted 8 Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him 8 Percocet.

P: Follow up with [redacted] in 3 months.

[redacted]

[redacted]

### **Reference Answer**

Record date: 2075-01-07

NAME: Villegas, Yosef

MRN: 8249813

S: Here for follow up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with ADLs. He says he uses Viagra occasionally and it works for him.

O: Well-appearing male. Blood pressure 142/74. Chest clear. COR: No murmur.

A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check FLP.(5) Worried about dementia. Will check B12, TSH, RPR for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted 8 Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him 8 Percocet.

P: Follow up with Dr. Hobbs in 3 months.

Gilbert P. Perez, M.D.

GPP/church/olinger

Figure 141: **Medical Text Anonymization: Case 7.** An example showing the capability of o1-preview to anonymize medical texts with the implicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### **Prompt**

What are the names? Who are the professions? What are the locations? What are ages? What are dates? What are contacts? What are IDs? What are the phone numbers? Replace the matching characters with [redacted]:

### **O1-preview**

Record date: [redacted]

NAME: [redacted]

MRN: [redacted]

S: Here for follow-up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with ADLs. He says he uses Viagra occasionally and it works for him.

O: Well-appearing male. Blood pressure 142/74. Chest clear. COR: No murmur.

A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check FLP.(5) Worried about dementia. Will check B12, TSH, RPR for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted 8 Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him 8 Percocet.

P: Follow up with [redacted] in 3 months.

[redacted]

[redacted]

### **Reference Answer**

Record date: 2075-01-07

NAME: Villegas, Yosef

MRN: 8249813

S: Here for follow up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with ADLs. He says he uses Viagra occasionally and it works for him.

O: Well-appearing male. Blood pressure 142/74. Chest clear. COR: No murmur.

A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check FLP.(5) Worried about dementia. Will check B12, TSH, RPR for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted 8 Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him 8 Percocet.

P: Follow up with Dr. Hobbs in 3 months.

Gilbert P. Perez, M.D.

GPP/church/olinger

Figure 142: **Medical Text Anonymization: Case 8.** An example showing the capability of o1-preview to anonymize medical texts with the general prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

### **Prompt**

Please anonymize the following clinical note. Specifically, replace all the following information with the term "[redacted]": redact any strings that might be a name or acronym or initial, redact any strings separated by the "V" symbol, redact patients' names, doctors' names and the strings in front of M.D. or after Dr., redact pager names and medical staff names, redact any strings that look like something years old or age 37, redact any dates and IDs and numbers and record dates, redact locations and addresses and clinic names, redact professions and ages and contacts, redact any acronyms and initials.:

### **O1-preview**

Record date: [redacted]

NAME: [redacted]

MRN: [redacted]

S: Here for follow-up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with [redacted]. He says he uses Viagra occasionally and it works for him. O: Well-appearing male. Blood pressure [redacted]. Chest clear.

[redacted]. No murmur. A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check [redacted]. (5) Worried about dementia. Will check [redacted] for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted [redacted] Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him [redacted] Percocet.

P: Follow up with [redacted] in [redacted].

[redacted][redacted]

### **Reference Answer**

Record date: 2075-01-07

NAME: Villegas, Yosef

MRN: 8249813

S: Here for follow up of his colonic polyps. He wants to know the results from them. He is not taking hydrochlorothiazide and is curious about his blood pressure. He said he has cut his alcohol back to a 6-pack once a week. He has cut back his cigarettes to one time per week. He does say he has had some problems with his memory recently, although no other manifestations of dementia and no difficulty with ADLs. He says he uses Viagra occasionally and it works for him.

O: Well-appearing male. Blood pressure 142/74. Chest clear. COR: No murmur.

A:(1) Alcohol excess discussed. I told him to cut it out entirely.(2) Cigarette smoking. Discussed.(3) Hypertension. Under good control.(4) Elevated lipids. Will check FLP.(5) Worried about dementia. Will check B12, TSH, RPR for now.(6) Status-post knee replacement. Gets occasional Vicodin from his orthopedist. Says he wanted 8 Percocet for the next month or two that he will take on rare occasions when his pain is more severe, and I did give him 8 Percocet.

P: Follow up with Dr. Hobbs in 3 months.

Gilbert P. Perez, M.D.

GPP/church/olinger

Figure 143: **Medical Text Anonymization: Case 9.** An example showing the capability of o1-preview to anonymize medical texts with the explicit prompt. Yellow highlights the annotations made by human experts. Green marks the sensitive privacy information detected by o1-preview. Red indicates modifications beyond the human-marked annotation.

## A.14 Social Media Analysis

### Prompt

Which of these 4 emotions– [anger, joy, optimism, sadness]–best represents the mental state of the person writing the following text?

Answer only with a number: 0 if anger, 1 if joy, 2 if optimism, 3 if sadness. Here is the text:

Of course I've got a horrible cold and am breaking out 2 days before grad 

### O1-preview

0 (anger)

### Reference Answer

3 (sadness)

### Prompt

Which of these 4 emotions– [anger, joy, optimism, sadness]–best represents the mental state of the person writing the following text?

Answer only with a number: 0 if anger, 1 if joy, 2 if optimism, 3 if sadness. Here is the text:

The stupidest and weirdest thing people do. And what's more stupid than that? They upload it online. Oh my god. But good for laughter ah. 

### O1-preview

1

### Reference Answer

1 (joy)

### Prompt

Which of these 4 emotions– [anger, joy, optimism, sadness]–best represents the mental state of the person writing the following text?

Answer only with a number: 0 if anger, 1 if joy, 2 if optimism, 3 if sadness. Here is the text:

Pages like AjPlus profit off our outrage... they serve no purpose other than to show us shit that pisses us off. Don't need that negativity

### O1-preview

0

### Reference Answer

0 (anger)

### Prompt

Which of these 4 emotions– [anger, joy, optimism, sadness]–best represents the mental state of the person writing the following text?

Answer only with a number: 0 if anger, 1 if joy, 2 if optimism, 3 if sadness. Here is the text:

#good to learning #wisdom &lt;&lt; reform (v): make in order to improve something &gt;&gt;

### O1-preview

2

### Reference Answer

2 (optimism)

Figure 144: **Social Media Analysis: Case 6-9.** Demonstration cases for emotion recognition. Red indicates that the output of the o1-preview model is completely inconsistent with the reference answer, while green signifies full consistency.

**Prompt**

*Is the following post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post:*

#JoeScarborough is a vile excuse for a journalist". Here he is criticizing #Trump yet refuses to come clean about the death of an intern in his office. Stop talking about everyone else and start talking about YOUR INVOLVEMENT WITH #LORIKAUSUTIS. @user

**O1-preview**

1

**Reference Answer**

1 (offensive)

**Prompt**

*Is the following post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post:*

#AIATravelTips Always monitor your luggage at the airport. Do not leave your luggage unattended and do not accept unopened packages. ---- #Tips #Travel #Explore #AIA #Aircraft #Airport #Caribbean #Destination #Travelling #Connect #TravelTheWorld #Adventure #Adventures #Luggage

**O1-preview**

0

**Reference Answer**

0 (non-offensive)

Figure 145: **Social Media Analysis: Case 10-11.** Demonstration cases for irony detection. Red indicates that the output of the o1-preview model is completely inconsistent with the reference answer, while green signifies full consistency.

## A.15 Analogical Reasoning

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

```
{ } ^ {  
^  
} * {
```

### **O1-preview**



### **Reference Answer**



### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

```
king => C C C  
queen => c c c  
man => C C  
woman =>
```

### **O1-preview**

cc

### **Reference Answer**

cc

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

```
black => } ^ {  
gray => ^  
red => } * {  
door =>
```

### **O1-preview**



### **Reference Answer**



Figure 146: **Analogical Reasoning: Case 6-8.** A demonstration case for direct semantic structure mapping. Green denotes that the answers are correct.

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

motorcycle => \*\*

unicycle => !

tricycle => !!!

skateboard => !!!!

car => \*\*\*\*

bicycle =>

### **O1-preview**

!!

### **Reference Answer**

!!

### **Prompt**

We are conducting an experiment on general reasoning abilities. You will be presented with a series of words and drawings. Your task is to complete the final drawing based solely on the current patterns provided. Please respond concisely with only the final drawing.

Questions:

grandfather => !

grandmother => \*

mother => \*\*

father => !!

brother => !!!

sister =>

### **O1-preview**

\*\*\*

### **Reference Answer**

\*\*\*

Figure 147: **Analogical Reasoning: Case 9-10.** A demonstration case for multiple layers of semantic content and structure. Green denotes that the answers are correct.

## A.16 Sentiment Analysis

### Prompt

You are a sentiment analysis expert. Please analyze all the emotions expressed in the following text. For each emotion, provide the sentiment category, the relevant text snippet, and a brief explanation.

Text:

All the appetizers and salads were fabulous, the steak was mouth watering and the pasta was delicious!!!

### O1-preview

\*\*Emotion 1: Satisfaction\*\*

- \*\*Sentiment Category:\*\* Positive
- \*\*Relevant Text Snippet:\*\* "All the appetizers and salads were fabulous"
- \*\*Explanation:\*\* The word "fabulous" indicates a high level of satisfaction with the appetizers and salads.

---

\*\*Emotion 2: Delight\*\*

- \*\*Sentiment Category:\*\* Positive
- \*\*Relevant Text Snippet:\*\* "the steak was mouth watering"
- \*\*Explanation:\*\* Describing the steak as "mouth watering" conveys delight and enthusiasm about the steak.

\*\*Emotion 3: Excitement\*\*

- \*\*Sentiment Category:\*\* Positive
- \*\*Relevant Text Snippet:\*\* "the pasta was delicious!!!"
- \*\*Explanation:\*\* The use of "delicious" along with multiple exclamation marks expresses excitement and strong approval of the pasta.

---

\*\*Overall Emotion: Enthusiasm\*\*

- \*\*Sentiment Category:\*\* Positive
- \*\*Relevant Text Snippet:\*\* The entire sentence
- \*\*Explanation:\*\* The enthusiastic tone throughout the text reflects a highly positive dining experience, emphasizing enjoyment of all courses.

### Reference Answer

```
<sentence id="813">
    <text>All the appetizers and salads were fabulous, the steak was mouth watering and the pasta was
delicious!!!</text>
    <aspectTerms>
        <aspectTerm term="appetizers" polarity="positive" from="8" to="18"/>
        <aspectTerm term="salads" polarity="positive" from="23" to="29"/>
        <aspectTerm term="steak" polarity="positive" from="49" to="54"/>
        <aspectTerm term="pasta" polarity="positive" from="82" to="87"/>
    </aspectTerms>
    <aspectCategories>
        <aspectCategory category="food" polarity="positive"/>
    </aspectCategories>
</sentence>
```

Figure 148: **Sentiment analysis: Case 5.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

---

**Prompt**

You are a sentiment analysis expert. Please analyze all the emotions expressed in the following text. Its input is an E#A pair and a sentence, requiring us to output the emotional polarity of this E#A pair.

For example:

Judging from previous posts this used to be a good place, but not any longer.  
{target:"NULL" category:"RESTAURANT#GENERAL" polarity:"negative" from:"-" to="-"}

Text:

The food was lousy - too sweet or too salty and the portions tiny.

**O1-preview**

{target:"food" category:"FOOD#TASTE" polarity:"negative" from:"-" to:"-"}  
{target:"portions" category:"FOOD#SIZE" polarity:"negative" from:"-" to:"-")}

**Reference Answer**

{target:"food" category="FOOD#QUALITY" polarity="negative" from:"4" to:"8"}  
{target:"portions" category:"FOOD#STYLE\_OPTIONS" polarity:"negative" from:"52" to:"60")}

Figure 149: **Sentiment analysis: Case 6.** An example illustrating the capability of o1-preview in the field of sentiment analysis.

## A.17 Public Health Policy Analysis

### Prompt

What evidence exists that the ACA's Medicaid expansion improved access to care and health outcomes for low-income populations?

### O1-preview

The Affordable Care Act's (ACA) **Medicaid expansion** has been a subject of extensive research since its implementation, and a substantial body of evidence indicates that it has significantly improved access to care and health outcomes for low-income populations. Below is a summary of the key findings from various studies and analyses up to October 2023.

#### Increased Insurance Coverage

- Evidence: Kaiser Family Foundation (KFF): Reports have consistently shown that states opting for Medicaid expansion experienced substantial increases in Medicaid enrollment among low-income adults. For instance, between 2014 and 2021, expansion states saw an average increase of 35% in Medicaid enrollment compared to non-expansion states.

- National Health Interview Survey (NHIS): Data from the NHIS indicated that the uninsured rate among adults with incomes below 138% of the federal poverty level (FPL) dropped significantly in expansion states, from 35% in 2013 to 15% in 2021.

#### Improved Access to Care

- Evidence: Health Affairs (2022): A study found that Medicaid expansion was associated with a 5-11% increase in having a usual source of care and a 7% increase in cost-related barriers to care among low-income adults.

JAMA Internal Medicine (2017): Research indicated that low-income individuals in expansion states were more likely to receive preventive services, such as cancer screenings and routine check-ups, compared to those in non-expansion states.

#### Increased Utilization of Health Services

- Evidence: Annals of Internal Medicine (2017): The study found that Medicaid expansion led to increased access to primary care and specialist visits, particularly for chronic conditions such as diabetes and hypertension.

New England Journal of Medicine (NEJM, 2019): The study showed a 9% increase in outpatient visits and a 6% increase in prescriptions filled among Medicaid enrollees in expansion states.

The Commonwealth Fund (2021): Reports from the Commonwealth Fund showed that Medicaid expansion significantly improved access to specialty care for chronic diseases.

#### Reduced Health Disparities

- Evidence: Health Equity (2022): Studies demonstrated that Medicaid expansion helped reduce racial and ethnic disparities in access to care and health outcomes, particularly among African American and Hispanic populations.

#### Conclusion:

The overwhelming body of evidence up to October 2023 strongly suggests that the ACA's **Medicaid expansion** has led to:

- Increased insurance coverage among low-income populations.
- Improved access to preventive and other health services.
- Increased utilization of healthcare services.
- Reduced racial and ethnic disparities in health outcomes.
- Positive effects on mental health outcomes.

### Reference Answer

Medicaid expansion under the ACA significantly reduced the uninsured rate, particularly among Black and Hispanic populations. The expansion also improved timely access to care, as evidenced by reduced rates of conditions like ruptured appendicitis in these populations.

Figure 150: **Public Health Policy Analysis: Case 4.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

### **Prompt**

What was the effect of the ACA on the uninsured rate in the U.S., particularly among young adults and minority groups?

### **O1-preview**

The Affordable Care Act (ACA) had a significant impact on reducing the uninsured rate in the United States, particularly among young adults and minority groups.

Overall Reduction in Uninsured Rate:

- National Decline: The uninsured rate among non-elderly Americans dropped from about 16% in 2010 to around 9% by 2016 after the ACA's major provisions took effect.

Impact on Young Adults:

- **Extended Parental Coverage:** One of the ACA's early provisions allowed young adults up to age 26 to remain on their parents' health insurance plans.

- Decrease in Uninsured Young Adults: This led to a substantial decline in the uninsured rate for individuals aged 19 to 25, dropping from approximately 34% in 2010 to about 15% by 2016.

Effects on Minority Groups:

- **Medicaid Expansion and Subsidies:** The ACA expanded Medicaid eligibility and provided subsidies for private insurance, which disproportionately benefited minority populations who had higher uninsured rates.

- African Americans: The uninsured rate decreased from around 19% in 2010 to about 11% by 2016.

- Hispanic Americans: Experienced a reduction from approximately 32% uninsured to about 19% in the same period.

- Asian Americans and Other Minorities: Also saw significant decreases in uninsured rates due to increased access to affordable coverage.

Remaining Challenges:

- Persistent Disparities: Despite overall improvements, minority groups still have higher uninsured rates compared to whites.

- **Medicaid Expansion Gaps:** Not all states opted to expand Medicaid, limiting coverage gains in certain areas, particularly affecting minorities and low-income individuals.

- **Undocumented Immigrants:** Many remain uninsured due to ineligibility for ACA provisions.

Conclusion:

The ACA significantly lowered the uninsured rate across the nation, with young adults and minority groups seeing some of the most substantial gains in health insurance coverage. These improvements have enhanced access to healthcare services and contributed to better health outcomes for these populations. However, disparities persist, and ongoing efforts are necessary to address coverage gaps and ensure equitable access to healthcare for all.

### **Reference Answer**

The ACA significantly reduced the uninsured rate, particularly among young adults and minority populations. The ability for young adults to remain on their parents' insurance until age 26 was a key factor in reducing the uninsured rate in this group.

Figure 151: **Public Health Policy Analysis: Case 5.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

### **Prompt**

How did the ACA affect hospital readmission rates and the quality of care provided in surgical settings?

### **O1-preview**

The Affordable Care Act (ACA) significantly impacted hospital readmission rates and the quality of care in surgical settings through several key initiatives:

Reduction in Hospital Readmission Rates:

- **Hospital Readmissions Reduction Program (HRRP):** Implemented under the ACA, the HRRP penalizes hospitals with higher-than-expected 30-day readmission rates for specific conditions and procedures, including certain surgical interventions. This financial incentive encouraged hospitals to improve discharge planning, patient education, and post-discharge support to reduce unnecessary readmissions.

Enhanced Care Coordination:

- The ACA promoted better coordination among healthcare providers, leading to smoother transitions from hospital to home or other care settings. Improved communication and follow-up care helped identify and address postoperative complications earlier, decreasing the likelihood of readmissions.

Use of Evidence-Based Protocols:

- Hospitals adopted standardized clinical pathways and protocols for surgical care to minimize variations in practice, which contributed to lower complication rates and readmissions.

Improvement in Quality of Care in Surgical Settings:

- **Value-Based Purchasing Program:** The ACA introduced this program to reward hospitals for providing high-quality care. Payments are adjusted based on performance metrics, including surgical outcomes, patient safety indicators, and patient satisfaction scores. This incentivized hospitals to focus on quality improvements in surgical care.

• Emphasis on Patient Safety: The ACA supported initiatives to reduce hospital-acquired conditions and surgical site infections. Hospitals implemented safety checklists, infection control measures, and staff training programs to enhance patient safety in surgical settings.

• Accountable Care Organizations (ACOs): By fostering the development of ACOs, the ACA encouraged integrated care models where providers share responsibility for patient outcomes. This collaborative approach improved preoperative and postoperative care management for surgical patients.

• Expanded Access to Preventive Services: With increased insurance coverage, more patients received timely preventive care and elective surgeries before conditions worsened, leading to better outcomes and fewer complications.

Challenges and Considerations:

• Variable Impact: While overall readmission rates decreased, some hospitals, particularly those serving high-risk or low-income populations, faced challenges in meeting the targets due to factors beyond clinical care, such as socioeconomic determinants.

• **Documentation and Administrative Burden:** Hospitals had to invest in data collection and reporting systems to comply with ACA requirements, which sometimes diverted resources from direct patient care.

Conclusion:

The ACA led to a reduction in hospital readmission rates and enhanced the quality of care in surgical settings by incentivizing hospitals to adopt best practices, improve care coordination, and focus on patient safety. These efforts resulted in better surgical outcomes, fewer complications, and a more efficient healthcare system. However, disparities remain, and ongoing efforts are necessary to address challenges faced by certain hospitals and patient populations to ensure equitable improvements across the board.

### **Reference Answer**

The ACA's Hospital Readmissions Reduction Program reduced readmission rates for several conditions, including surgical procedures such as coronary artery bypass grafting and total hip/knee replacements. However, the program's overall impact on surgical care quality remains mixed, with no significant changes in some surgical outcomes.

Figure 152: **Public Health Policy Analysis: Case 6.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

**Prompt**

What are the long-term impacts of the ACA on health insurance marketplaces and the availability of affordable insurance plans? Return brief summary.

**O1-preview**

The Affordable Care Act (ACA) has had several long-term impacts on health insurance marketplaces and the availability of affordable insurance plans:

- **Expanded Access to Insurance:** The creation of health insurance marketplaces allowed millions of individuals and families to purchase insurance plans, many for the first time.
- **Financial Assistance:** Subsidies and tax credits were provided to lower- and middle-income individuals, making insurance plans more affordable and increasing enrollment.
- **Standardized Coverage:** The ACA mandated that all plans offer essential health benefits and prohibited denial of coverage for pre-existing conditions, improving the quality and comprehensiveness of available plans.
- **Market Stabilization Efforts:** Measures like the individual mandate (though later repealed) were introduced to encourage a balanced risk pool, aiming to keep premiums stable.
- **Challenges with Affordability and Choice:** Some regions experienced rising premiums and limited insurer participation over time, affecting the affordability and variety of plans available.

Overall, the ACA significantly increased the availability of affordable insurance plans through the health insurance marketplaces, improving access to healthcare. However, ongoing challenges such as premium costs and insurer participation continue to affect the long-term sustainability and affordability of these marketplaces.

**Reference Answer**

The ACA created health insurance marketplaces that provided access to affordable insurance plans for millions of Americans, supported by tax credits and subsidies. While the marketplaces remain stable, concerns about premium increases and funding cuts continue to pose challenges for long-term sustainability.

Figure 153: **Public Health Policy Analysis: Case 7.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

### **Prompt**

What challenges did the ACA face in achieving its goals related to reducing overall healthcare costs?

### **O1-preview**

The Affordable Care Act (ACA), enacted in 2010, sought to reduce overall healthcare costs while expanding insurance coverage and improving care quality. However, achieving cost reduction proved challenging due to several factors:

- **Rising Healthcare Costs Pre-Reforms:** Healthcare expenses continued to escalate due to an aging population, increased prevalence of chronic diseases, and the high cost of new medical technologies and prescription drugs. These systemic issues are difficult to address through policy changes alone.
  - **Limited Control Over Drug and Device Prices:** The ACA did not implement significant measures to regulate the prices of prescription drugs and medical devices. Pharmaceutical and medical device companies wield considerable influence and successfully lobbied against price controls, leaving these costs largely unchecked.
  - **Fee-for-Service Payment Model:** The dominant fee-for-service model incentivizes providers to perform more procedures and tests, focusing on quantity rather than quality of care. Although the ACA introduced alternative payment models emphasizing value over volume, the transition has been slow and met with resistance from some healthcare providers.
  - **Administrative Complexity and Costs:** The U.S. healthcare system's complexity leads to high administrative expenses. The ACA introduced new regulations and reporting requirements, which increased administrative burdens for providers and insurers, potentially offsetting some cost savings.
  - **Political Opposition and Legal Challenges:** Persistent political opposition and numerous legal challenges created uncertainty and hindered the full implementation of the ACA's provisions. Efforts to repeal or weaken the law disrupted initiatives designed to control costs, such as Medicaid expansion in certain states.
  - **Fragmented Care Delivery:** The lack of coordination among various healthcare providers results in inefficiencies and duplicate services. While the ACA promoted Accountable Care Organizations (ACOs) to improve care coordination, integrating these models into the existing system has been challenging.
  - **Consumer Behavior and Preventive Care:** Changing patient behavior is crucial for cost reduction. Despite the ACA's emphasis on preventive services, many individuals continue to underutilize these benefits, leading to higher costs associated with preventable illnesses.
  - **Lack of Price Transparency:** Without clear information on the cost of services, consumers cannot make informed decisions, and market forces cannot effectively drive down prices. The ACA's measures to improve transparency have been insufficient in overcoming this challenge.
  - **Slow Adoption of Health Information Technology:** Implementing electronic health records (EHRs) and other technologies can reduce costs by improving efficiency. However, adoption has been slow due to high initial costs, ongoing maintenance expenses, and concerns about workflow disruptions.
  - **Underfunding and Implementation Issues:** Some cost-saving programs under the ACA were underfunded or not fully implemented due to budget constraints and administrative challenges, limiting their potential impact on reducing expenses.
  - **High Cost of End-of-Life Care:** A significant portion of healthcare spending occurs in the final months of life. The ACA struggled to address these costs due to the sensitive nature of end-of-life care discussions and political controversies surrounding the topic.
  - **Variability in Medicaid Expansion:** Not all states chose to expand Medicaid, which limited the ACA's ability to reduce uncompensated care costs nationwide and affected the overall risk pool of insured individuals.
  - **Insurance Market Dynamics:** Regulations that prohibited denying coverage based on pre-existing conditions and required comprehensive benefits led some insurers to raise premiums. Higher premiums can make insurance less affordable, affecting the ACA's goal of cost reduction.
- In summary, the ACA faced multifaceted challenges in reducing healthcare costs, stemming from systemic issues within the healthcare system, political and legal obstacles, stakeholder resistance, and complexities in changing established practices. While the ACA made strides in certain areas, significantly lowering overall healthcare expenses remains a complex goal requiring ongoing efforts beyond the scope of the legislation.

### **Reference Answer**

The ACA reduced healthcare costs in certain areas, such as hospital readmissions and bundled payment initiatives. However, challenges remain, particularly with rising healthcare costs and the sustainability of programs like the Medicare Shared Savings Program (MSSP), which initially did not focus on surgical care.

Figure 154: **Public Health Policy Analysis: Case 8.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

### **Prompt**

How effective was the ACA in promoting the use of electronic health records and other healthcare technologies to improve care coordination?

### **O1-preview**

The Affordable Care Act (ACA), enacted in 2010, aimed to improve the U.S. healthcare system by expanding access, enhancing quality, and reducing costs. A significant component of this effort involved promoting the use of electronic health records (EHRs) and other healthcare technologies to improve care coordination among providers. Here's an analysis of how effective the ACA was in achieving this goal:

Promotion of EHR Adoption:

- Integration with Payment Models: The ACA introduced new care delivery models like Accountable Care Organizations (ACOs) and Patient-Centered Medical Homes (PCMHs). These models emphasize coordinated care and rely heavily on health information technology to share patient data among providers. The ACA incentivized providers to adopt EHRs by tying reimbursement to the use of these technologies in care coordination.
- Quality Reporting Requirements: The law mandated the reporting of certain quality measures that are easier to track and report using EHR systems. This pushed healthcare providers toward adopting electronic systems to meet compliance requirements.
- Support for Health Information Exchanges (HIEs): The ACA provided funding and support for the development of HIEs, which facilitate the secure sharing of patient information across different healthcare settings. This interoperability is crucial for effective care coordination.

Effectiveness in Improving Care Coordination:

- Increased EHR Adoption Rates: By 2020, EHR adoption among office-based physicians exceeded 85%, a significant increase from prior years. While the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 initially spurred this growth through financial incentives, the ACA reinforced the importance of EHRs in new payment and care delivery models.
- Enhanced Data Sharing: The promotion of HIEs and interoperability standards under the ACA has led to improved data sharing capabilities. Providers can now more easily access a patient's medical history, lab results, and medication lists, which is essential for coordinated care.
- Reduction in Duplicate Testing: Improved access to patient records has helped reduce unnecessary duplicate tests and procedures, saving costs and improving patient burden.
- Mixed Outcomes on Care Quality: While there have been improvements in certain areas like preventive care and chronic disease management, the impact on overall care quality and patient outcomes has been mixed. Some studies suggest modest improvements, while others indicate that simply adopting EHRs does not automatically translate to better care without proper implementation and use.

Challenges and Limitations:

- Interoperability Issues: Despite efforts, achieving seamless interoperability between different EHR systems remains a challenge. This limits the potential for full-scale care coordination across various providers and care settings.
- Administrative Burden: The transition to EHRs has introduced significant administrative tasks for healthcare providers, sometimes detracting from patient care. Clinician burnout related to EHR use has been a growing concern.
- Variable Implementation: The effectiveness of EHRs in improving care coordination often depends on how well they are integrated into clinical workflows. Inconsistent implementation and varying levels of user proficiency can hinder their potential benefits.

Conclusion:

The ACA was moderately effective in promoting the use of electronic health records and healthcare technologies to improve care coordination. It built upon previous legislation by integrating health IT into broader healthcare reform initiatives, encouraging providers to adopt and meaningfully use EHRs. While there have been notable improvements in EHR adoption rates and some aspects of care coordination, challenges like interoperability issues and administrative burdens have limited the full realization of these technologies' benefits. Continued efforts are needed to address these challenges and enhance the effectiveness of health IT in achieving coordinated, high-quality care.

### **Reference Answer**

The ACA's focus on improving care coordination through Accountable Care Organizations (ACOs) led to increased adoption of electronic health records and other technologies. ACOs have contributed to better tracking of health outcomes, although challenges remain in fully integrating these technologies into surgical care settings.

Figure 155: **Public Health Policy Analysis: Case 9.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

### **Prompt**

What evidence suggests that the ACA helped improve health outcomes related to chronic conditions such as diabetes and hypertension?

### **O1-preview**

Evidence suggests that the Affordable Care Act (ACA) has contributed to improved health outcomes for chronic conditions like diabetes and hypertension through several key mechanisms:

- **Increased Insurance Coverage:** The ACA expanded Medicaid eligibility and provided subsidies for private insurance, leading to millions of previously uninsured Americans gaining health coverage. This expansion improved access to healthcare services essential for managing chronic conditions.
- **Study Example:** A 2017 study published in the Journal of the American Medical Association (JAMA) found that Medicaid expansion under the ACA was associated with a significant increase in insurance coverage among low-income adults, which led to better access to care for chronic diseases like diabetes and hypertension.
- **Enhanced Access to Preventive Services:** The ACA mandated that most insurance plans cover preventive services without cost-sharing. This includes screenings for blood pressure and blood sugar levels, which are crucial for early detection and management of hypertension and diabetes.
- **Study Example:** Research in Health Affairs (2015) showed that the use of preventive services increased after the ACA's implementation, particularly among populations at risk for chronic conditions.
- **Improved Medication Adherence:** With better insurance coverage, more individuals could afford medications necessary for controlling chronic conditions.
- **Study Example:** A 2018 study in Diabetes Care reported that Medicaid expansion states saw a significant increase in prescriptions filled for diabetes medications compared to non-expansion states.
- **Better Disease Management:** The ACA encouraged coordinated care models, like Patient-Centered Medical Homes (PCMHs), which improve the management of chronic diseases through comprehensive care plans.
- **Study Example:** According to a 2016 report in the American Journal of Managed Care, patients with chronic conditions enrolled in PCMHs had better health outcomes, including improved blood pressure and blood sugar control.
- **Reduction in Health Disparities:** The ACA helped reduce racial and socioeconomic disparities in healthcare access, which disproportionately affect chronic disease outcomes.
- **Study Example:** A 2018 study in The Lancet found that the ACA narrowed the gap in access to care between minority and white patients, leading to earlier diagnosis and treatment of hypertension and diabetes in underserved populations.
- **Positive Trends in Clinical Outcomes:** Several studies have linked the ACA to measurable improvements in health markers associated with chronic conditions.
- **Study Example:** Research published in Circulation (2019) indicated that states with Medicaid expansion experienced greater reductions in uncontrolled hypertension rates than non-expansion states.
- **Decreased Hospitalizations and Emergency Visits:** Improved outpatient management of chronic conditions has led to fewer hospitalizations and emergency room visits.
- **Study Example:** A 2020 study in Health Services Research showed a decline in hospital admissions for diabetic complications in Medicaid expansion states.

### **Conclusion:**

The evidence indicates that the ACA has had a positive impact on health outcomes for chronic conditions such as diabetes and hypertension by increasing insurance coverage, enhancing access to preventive services, improving medication adherence, and promoting better disease management. These improvements have led to earlier detection, more effective treatment, and reductions in complications associated with these chronic diseases.

### **Reference Answer**

The ACA improved access to preventive services and management of chronic conditions, including diabetes and hypertension, by expanding Medicaid and ensuring coverage for preventive care. This led to earlier detection and better management of chronic diseases, particularly in low-income populations.

Figure 156: **Public Health Policy Analysis: Case 10.** A example illustrating the capability of o1-preview to respond to specialized questions of health policy in the domain of Public Health. Yellow indicates partial consistency between the o1-preview response and the reference answer, and red indicates fully consistency, while green signifies complete consistency.

## A.18 Medical Genetics and Genomics Reasoning

### Prompt

You are the genetic medical researcher, based on the disease description and gene description, reasoning if the gene is related to the disease, give your inference and argumentations.

#### Gene description:

In the vascular system, SOD3 appears to be located on the endothelial cell surface. The characteristic distinguishing SOD3 from SOD1 and SOD2 is the heparin-binding capacity. SOD3 binds on the surface of endothelial cells through the heparan sulfate proteoglycan and eliminates the oxygen radicals from the NADP-dependent oxidative system of neutrophils. Adachi et al. (1992) developed an immunoassay system for EC-SOD in order to measure EC-SOD levels in the serum of healthy subjects. They found that 6% of these persons had an SOD3 level that was 10- to 15-fold higher than the mean SOD3 level in all subjects.

#### Disease description:

The SOD1 gene encodes superoxide dismutase-1 (EC 1.15.1.1), a major cytoplasmic antioxidant enzyme that metabolizes superoxide radicals to molecular oxygen and hydrogen peroxide, thus providing a defense against oxygen toxicity (Niwa et al., 2007). Soluble cytoplasmic SOD1 is a copper- and zinc-containing enzyme; the SOD1 gene maps to chromosome 21q22 (Sherman et al., 1983). SOD2 (147460) is a distinct mitochondrial enzyme that contains manganese; the SOD2 gene maps to 6q25. SOD1 is a homodimer and SOD2 a tetramer (Beckman et al., 1973). Fridovich (1979) concluded that SOD1 and SOD2 evolved from different primordial genes, which is an example of analogy, not homology, and of convergent evolution. Doonan et al. (1984) cited the superoxide dismutases as an example of cytosolic and mitochondrial isoenzymes with no apparent evolutionary relationship.

### O1-preview

Based on the provided descriptions, there is a potential connection between the SOD3 gene and diseases associated with oxidative stress, similar to those involving SOD1 and SOD2.

#### \*\*Gene Description:\*\*

- \*\*SOD3 (Extracellular Superoxide Dismutase):\*\*
  - Located on the endothelial cell surface in the vascular system.
  - Distinct from SOD1 and SOD2 due to its heparin-binding capacity.
  - Binds to endothelial cells via heparan sulfate proteoglycans.
  - Eliminates oxygen radicals from the NADP-dependent oxidative system of neutrophils.
  - Approximately 6% of healthy individuals have SOD3 levels 10- to 15-fold higher than the average.

#### \*\*Disease Description:\*\*

- \*\*SOD1:\*\*
  - Encodes superoxide dismutase-1, a major cytoplasmic antioxidant enzyme.
  - Metabolizes superoxide radicals to molecular oxygen and hydrogen peroxide.
  - Provides defense against oxygen toxicity.
  - Located in the cytoplasm and contains copper and zinc.
- \*\*SOD2:\*\*
  - A mitochondrial enzyme containing manganese.
  - Functions similarly to SOD1 but within the mitochondria.

#### \*\*Evolutionary Context:\*\*

- SOD1 and SOD2 evolved from different primordial genes (convergent evolution).
- No apparent evolutionary relationship between cytosolic and mitochondrial isoenzymes.

#### \*\*Inference and Argumentation:\*\*

- \*\*Common Functionality:\*\*
  - All three SOD enzymes (SOD1, SOD2, SOD3) play crucial roles in detoxifying superoxide radicals, protecting cells from oxidative damage.
- \*\*Oxidative Stress and Disease:\*\*
  - Mutations or dysfunction in SOD1 are linked to diseases characterized by oxidative stress, such as amyotrophic lateral sclerosis (ALS).
  - Given their roles in managing reactive oxygen species, alterations in any SOD enzyme could contribute to disease pathology.

#### \*\*SOD3's Unique Role:\*\*

Figure 156: Medical Genetics and Genomics Reasoning: Case 3. Positive example illustrating the capability of the o1-preview to respond to gene-disease-related questions. Grey indicates the prompt instructions. Blue indicates the gene description and disease description. Green indicates the correct inferences. Yellow indicates the key deduction steps of the o1-preview.

## A.19 Medical Knowledge Question Answer

### Prompt

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "The part of the mandible which ossify in cartilage is:"

"opa", "Coronoid process", "opb", "Condylar process", "opc", "Upper half of ramus above the level of mandibular foramen", "opd", "All of the above"

### O1-preview

The correct answer is **D: All of the above**.

#### Explanation:

The mandible develops through both intramembranous and endochondral ossification processes.

1. **Intramembranous Ossification:** Most of the mandible, including the body of the mandible, develops directly from mesenchymal tissue through intramembranous ossification. This process does not involve a cartilage precursor.
2. **Endochondral Ossification:** Certain parts of the mandible develop from cartilage precursors through endochondral ossification. These are known as secondary cartilages and include:
  - **Condylar Process (Option B):** This is the main growth site of the mandible during development. The condylar cartilage contributes to the lengthening of the mandible and ossifies through endochondral ossification.
  - **Coronoid Process (Option A):** Although less significant than the condylar process, the coronoid process also develops from a secondary cartilage and ossifies endochondrally.
  - **Upper Half of the Ramus Above the Level of the Mandibular Foramen (Option C):** This area of the mandible also contains secondary cartilage that ossifies through endochondral ossification.

Since all the options listed involve parts of the mandible that ossify in cartilage (through endochondral ossification), the correct choice is:

**Answer: All of the above**

### Reference Answer

"exp": "The mandible is the second bone, next to the clavicle, to ossify in the body. Its greater part ossifies in membrane. The parts ossifying in cartilage include the incisive part below the incisor teeth, the coronoid and condyloid processes, and the upper half of the ramus above the level of the mandibular foramen.",

"cop": 4

Figure 157: **Medical Knowledge Question Answer: Case 3.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question":"MHC Class II proteins are expressed by:"

"opa":"B-cells, dendritic cells and macrophages","opb":"Platelets","opc":"T-cells","opd":"All nucleated cells"

**O1-preview**

**Answer:**

**Option A: B-cells, dendritic cells, and macrophages**

**Explanation:**

**Major Histocompatibility Complex (MHC) Class II Proteins:**

- MHC Class II molecules are essential for the immune system to recognize foreign particles.
- They present extracellularly derived peptides (antigens) to **CD4+ helper T cells**, initiating an immune response.
- Expression of MHC Class II is restricted to certain immune cells known as **professional antigen-presenting cells (APCs)**.

**Professional Antigen-Presenting Cells (APCs):**

1. **B-Cells:**

- Function as APCs by internalizing antigens through their B-cell receptors.
- Process and present antigen peptides on MHC Class II molecules to helper T cells.

2. **Dendritic Cells:**

- Highly specialized in capturing antigens and initiating T-cell responses.
- Express MHC Class II constitutively at high levels.
- Serve as a crucial link between innate and adaptive immunity.

3. **Macrophages:**

- Engulf pathogens and debris through phagocytosis.
- Present processed antigen peptides via MHC Class II to helper T cells.
- Activation enhances MHC Class II expression.

**Why the Other Options Are Incorrect:**

• **Option B: Platelets**

- **Platelets lack a nucleus and are not involved in antigen presentation.**
- They do not express MHC Class II molecules.

• **Option C: T-Cells**

- While helper T cells interact with MHC Class II molecules, they do not express MHC Class II themselves under normal circumstances.
- T cells primarily express MHC Class I molecules.

• Option D: All Nucleated Cells

- All nucleated cells express **MHC Class I** molecules, not MHC Class II.
- MHC Class I presents endogenous peptides to **CD8+ cytotoxic T cells**.

**Summary:**

MHC Class II proteins are expressed exclusively by professional antigen-presenting cells—**B-cells, dendritic cells, and macrophages**—which play a pivotal role in initiating and regulating immune responses by presenting extracellular antigens to helper T cells.

**Therefore, the correct answer is Option A.**

**Reference Answer**

"exp": "Answer- A. B-cells, dendritic cells and macrophages MHC Class II proteins are expressed by all antigen-presenting cells, which include B-cells, follicular dendritic cells and macrophages. MHC Class I proteins are **expressed by all nucleated cells, which excludes platelets and RBCs**. HLA class I antigens (A, B and C) are found on the surface of virtually all nucleated cells- They are the principal antigens involved in graft rejection and cell-mediated cytotoxicity. Class I molecules may function as components of hormonal receptors. HLA class II antigens are more restricted in distribution, being found only on cells of the immune system macrophages, dendritic cells, activated T cells, and particularly on B cells.",

"cop":1

Figure 158: **Medical Knowledge Question Answer: Case 4.** An example illustrating the capability of o1-preview to respond to questions that arise in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question":"Geniculate neuralgia is caused in the nerve"

"opa":"VII","opb":"IX","opc":"X","opd":"II"

**O1-preview**

The correct answer is A. VII (Facial Nerve).

**Explanation:**

Geniculate neuralgia is a rare condition characterized by severe, sudden, and repeated episodes of deep ear pain. The pain is often sharp, shooting, or burning and can be triggered by stimuli such as swallowing or touching specific areas around the ear.

This neuralgia is associated with the **geniculate ganglion**, which is a collection of nerve cell bodies associated with the **facial nerve (cranial nerve VII)**. The facial nerve has several functions, including controlling the muscles of facial expression, conveying taste sensations from the anterior two-thirds of the tongue, and providing parasympathetic fibers to certain glands.

The geniculate ganglion contains sensory fibers of the facial nerve, including those for taste and some sensation from the ear canal. When this ganglion is affected, it can lead to the symptoms observed in geniculate neuralgia. Conditions like **Ramsay Hunt syndrome type II** can involve the geniculate ganglion due to **reactivation of the varicella-zoster virus (shingles)**, leading to ear pain and facial paralysis.

**Other Options:**

- **IX (Glossopharyngeal Nerve):** Involved in glossopharyngeal neuralgia, which causes pain in the throat, tonsillar fossa, and the back of the tongue.
- **X (Vagus Nerve):** Rarely associated with neuralgia but can cause cardiac and gastrointestinal symptoms if affected.
- **II (Optic Nerve):** Associated with vision; lesions here lead to visual deficits, not ear pain.

Therefore, geniculate neuralgia is specifically associated with the facial nerve (cranial nerve VII).

**Answer:** VII

**Reference Answer**

"exp":"Geniculate neuralgia (Nervus intermedius neuralgia) results from **herpes zoster infection** of geniculate ganglion and nervus intermedius branch of seventh cranial nerve.",

"cop":1

Figure 159: **Medical Knowledge Question Answer: Case 5.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question":"Prucalopride drug is ?"

"opa":"5HT4 agonist","opb":"5HT2b agonist","opc":"5HT2b antagonist","opd":"5HT2a agonist"

**O1-preview**

**Answer:**

**A) 5HT4 agonist**

**Explanation:**

Prucalopride is a gastrointestinal prokinetic agent primarily used for the treatment of chronic constipation. Its mechanism of action is based on its high affinity and selectivity for the serotonin (5-hydroxytryptamine, 5-HT) subtype 4 receptors (5-HT4 receptors). By acting as a 5-HT4 receptor agonist, prucalopride stimulates colonic motility and accelerates intestinal transit.

Here's a detailed explanation:

**1. Mechanism of Action:**

- **5-HT4 Receptor Agonism:** Prucalopride binds selectively to 5-HT4 receptors located in the gastrointestinal (GI) tract. Activation of these receptors enhances the release of neurotransmitters like acetylcholine, which stimulates muscle contractions in the intestine, promoting bowel movements.

**2. Selectivity and Safety:**

- **High Selectivity:** Prucalopride has a high selectivity for 5-HT4 receptors over other serotonin receptor subtypes (such as 5-HT2A, 5-HT2B, and 5-HT3). This selectivity minimizes off-target effects and reduces the risk of adverse events associated with other receptors.
- **Cardiovascular Safety:** Unlike some earlier prokinetic agents (e.g., cisapride, which also affected 5-HT2B receptors and was associated with cardiac issues), prucalopride's selectivity for 5-HT4 receptors translates to a better cardiovascular safety profile.

**3. Clinical Use:**

- **Treatment of Chronic Constipation:** By enhancing GI motility through 5-HT4 receptor activation, prucalopride effectively relieves symptoms of chronic constipation without significant impact on other serotonin-mediated systems.

**4. Excluding Other Options:**

- **Option B (5-HT2B agonist):** Agonism at 5-HT2B receptors is associated with valvular heart disease; prucalopride does not activate these receptors.
- **Option C (5-HT2B antagonist):** Prucalopride does not act as an antagonist at 5-HT2B receptors.
- **Option D (5-HT2A agonist):** Activation of 5-HT2A receptors is linked to various central nervous system effects; prucalopride does not target these receptors.

**Conclusion:**

Prucalopride is a selective 5-HT4 receptor agonist, used to enhance gastrointestinal motility in patients with chronic constipation. Its selectivity ensures efficacy while minimizing adverse effects associated with other serotonin receptors.

---

**Reference:** - Tack J, Camilleri M. "Prucalopride: targeted stimulation of colonic motility." *Gut*. 2010;59(3):297-299. - Bouras EP, Camilleri M. "Prucalopride for constipation." *Expert Opin Pharmacother*. 2011;12(4):605-616.

**Reference Answer**

"exp": "Ans. A. 5HT4 agonist" Prucalopride is a drug acting as a selective, high affinity 5-HT4 receptor agonist which targets the impaired motility associated with chronic constipation, thus normalizing bowel movements.\* Approved for use in Europe in 2009.\* Prucalopride, a first in class dihydro-benzofuran-carboxamide, is a selective, high affinity serotonin (5-HT4) receptor agonist with enterokinetic activities.\* Prucalopride alters colonic motility patterns serotonin 5-HT4 receptor stimulation: it stimulates colonic mass movements, which provide the main propulsive force for defecation.\* The observed effects are exeed highly selective action on 5-HT4 receptor.\* Prucalopride has >150-fold higher affinity for 5-HT4 receptors than for other receptors.",

"cop":1

Figure 160: **Medical Knowledge Question Answer: Case 6.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "A 16-year old girl was brought with primary amenorrhea. Her mother mentioned that she started developing breast at the age of 12. She was prescribed OCPs 2 years back by a doctor with no effect. She was having normal stature and was a football player. On examination, breasts were well developed (Tanner's stage 5) and pubic hair was minimal (Tanner's stage 1). What is the most probable diagnosis?"

"opa": "Premature ovarian failure", "opb": "Turner's syndrome", "opc": "Müllerian agenesis", "opd": "Androgen insensitivity"

**O1-preview**

The most probable diagnosis for this 16-year-old girl is **Androgen Insensitivity Syndrome (Option D)**.

**Explanation:**

To determine the most probable diagnosis, let's analyze the key clinical features presented:

1. **Primary Amenorrhea:** She has not started menstruating by the age of 16.
2. **Breast Development:** She started developing breasts at age 12, and on examination, her breasts are well developed (Tanner Stage 5).
3. **Minimal Pubic Hair:** She has minimal pubic hair (Tanner Stage 1).
4. **Normal Stature and Athleticism:** She has a normal height and is a football player, indicating normal overall growth and development.
5. **Ineffectiveness of Oral Contraceptive Pills (OCPs):** She was prescribed OCPs two years ago with no effect, suggesting that hormonal therapy did not initiate menstruation.

**Interpreting the Clinical Signs:**

- **Breast Development with Minimal Pubic Hair:** This combination suggests that estrogen effect is present (breast development), but androgen effect is absent or minimal (lack of pubic hair).
- **Primary Amenorrhea with Normal External Female Genitalia and Absent Uterus:** In some conditions, the uterus may be absent or underdeveloped, leading to primary amenorrhea despite normal secondary sexual characteristics.

**Differential Diagnoses:**

Let's evaluate each option based on the clinical features:

**A) Premature Ovarian Failure:**

- **Expected Findings:**
  - Cessation of ovarian function before age 40.
  - Decreased estrogen production.
  - Lack of breast development or regression.
  - Presence of pubic and axillary hair (since adrenal androgens are unaffected).
- **Conclusion:** Unlikely, because the patient has well-developed breasts and minimal pubic hair.

**B) Turner's Syndrome (45,XO):**

• **Expected Findings:**

- Short stature.
- Lack of breast development (gonadal dysgenesis).
- Webbed neck, shield chest.
- Amenorrhea due to streak ovaries.
- Normal pubic hair (adrenal androgens are present).

• **Conclusion:** Unlikely, as the patient has normal stature and well-developed breasts.

**C) Müllerian Agenesis (Mayer-Rokitansky-Küster-Hauser Syndrome):**

• **Expected Findings:**

- Primary amenorrhea.
- Normal secondary sexual characteristics (normal breast and pubic hair development).
- Absent or underdeveloped uterus and upper two-thirds of the vagina.

• **Conclusion:** Unlikely, because pubic hair development is normal in Müllerian agenesis, whereas our patient has minimal pubic hair.

**D) Androgen Insensitivity Syndrome (Testicular Feminization) :**

• **Pathophysiology:**

- Genetic males (46,XY) with end-organ resistance to androgens due to mutations in the androgen receptor gene.
- Testes produce androgens, but target tissues cannot respond to them.
- Excess androgens are converted to estrogens, promoting breast development.

• **Expected Findings:**

- Phenotypically female external genitalia.
- Well-developed breasts (due to estrogen effect).
- Absent or sparse pubic and axillary hair (due to androgen insensitivity).
- Primary amenorrhea (absence of uterus and upper vagina).
- Normal or tall stature.

• **Conclusion:** Consistent with the patient's presentation of well-developed breasts, minimal pubic hair, and primary amenorrhea.

**Final Diagnosis:**

Given the combination of **primary amenorrhea**, **well-developed breasts**, **minimal pubic hair**, and the ineffectiveness of hormonal therapy (OCPs), **Androgen Insensitivity Syndrome** is the most probable diagnosis.

**Answer: Androgen insensitivity (Option D)**

**Reference Answer**

"exp": "Ans. D. Androgen insensitivity" (Ref: Shawl 16/e p141, 15/e p111-112; Novaks 14/1037-1038; Dutta Gvnae 6/e p424) Most likely diagnosis here is androgen insensitivity syndrome. Androgen Insensitivity Syndrome: Also referred as "**Testicular feminization**". An X-linked recessive condition.

Results in failure of normal masculinization of external genitalia in chromosomally male individuals. Testes produce normal amounts of mullerian-inhibiting factor (MIF), also known as mullerian-inhibiting substance (MIS) or anti-mullerian hormone/factor (AMH/AMF). Features: Absence of fallopian tubes, a uterus, or proximal (upper) vagina. Identified in newborn period - By presence of inguinal masses & later identified as testes during surgery. Sometimes diagnosed in teenage years during primary amenorrhea evaluation. Absence of pubic & axillary hair in adolescent patients. Scanty body hair & lack of acne. Yet normal breast - Due to testosterone to estradiol conversion.",

"cop":4

Figure 161: **Medical Knowledge Question Answer: Case 7.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "A 42 years old male patient presented with jaundice. His AST was 48 U, ALT was 51 U, ALP, GGT were normal. Ultrasound of liver was suggestive of cirrhosis. Viral markers were done and the following results were obtained.

TestResult

Anti-HAV: Negative  
Anti-HBsAg: Negative  
Anti-HBeAg: Negative  
Anti-HBcIgG: Positive  
Anti-HBcIgM: Negative  
HBsAg: Negative  
Anti-HCV: Positive  
Anti-HEV: Negative

What is the next best step in management of this patient?"

"opa": "Sta interferon therapy", "opb": "Liver biopsy", "opc": "-PCR for hepatitis C virus", "opd": "PCR for HBV-DNA"

**O1-preview**

To determine the best next step in managing this patient, let's analyze the provided information systematically.

**Patient Profile:** - Age and Gender: 42-year-old male - **Symptoms:** Presented with jaundice - **Laboratory Findings:** - **AST:** 48 U/L (mildly elevated) - **ALT:** 51 U/L (mildly elevated) - **ALP and GGT:** Normal (indicating no significant cholestasis) - **Ultrasound:** Suggestive of cirrhosis (indicating chronic liver disease)

**Viral Hepatitis Markers:****1. Hepatitis A Virus (HAV):**

- Anti-HAV: Negative

**2. Hepatitis B Virus (HBV):**

- HBsAg (Hepatitis B surface antigen): Negative
- Anti-HBs (antibody to HBsAg): Negative
- Anti-HBc IgG (antibody to HBcAg, IgG type): Positive
- Anti-HBc IgM (antibody to HBcAg, IgM type): Negative
- HBeAg (Hepatitis B e antigen): Negative
- Anti-HBeAg: Not provided

**• Interpretation of HBV Markers:**

- **Isolated Anti-HBc IgG Positive:** This can indicate a resolved past infection, occult HBV infection, or a false-positive result. However, without HBsAg or anti-HBs, it's unclear whether active infection is present.
- **Occult HBV Infection:** Possible, but less likely to be the primary cause given the serology.

**3. Hepatitis C Virus (HCV):**

- Anti-HCV: Positive

- **Interpretation:** Indicates exposure to HCV but does not distinguish between past or current infection. Active infection requires detection of HCV RNA.

**4. Hepatitis E Virus (HEV):**

- Anti-HEV: Negative

**Analysis:**

- **Most Likely Cause of Liver Disease:** Given the positive anti-HCV antibody and signs of chronic liver disease (cirrhosis on ultrasound), chronic hepatitis C infection is the probable cause.
- **Next Best Step:**
  - **Confirm Active HCV Infection:** Determine if the HCV infection is active by detecting HCV RNA in the blood.
  - **PCR for HCV RNA:** This diagnostic test detects the presence of HCV RNA, confirming active viral replication and current infection.

**Why Not the Other Options?**

- **Option A (Start interferon therapy):** Initiating antiviral therapy without confirming active infection is premature. Treatment should be based on confirmed active infection and specific viral load measurements.
- **Option B (Liver biopsy):** While a liver biopsy can assess the degree of liver damage, it is invasive. Non-invasive tests should be performed first to confirm active infection before deciding on biopsy.
- **Option D (PCR for HBV-DNA):** Although there's a possibility of occult HBV infection due to isolated anti-HBc IgG positivity, the lack of other HBV markers and presence of anti-HCV suggests HCV is the more pressing issue to address first.

**Conclusion:**

The most appropriate next step is to perform a PCR test for hepatitis C virus to confirm active infection and guide further management.

**Answer: Option C: Order a PCR test for hepatitis C virus to confirm active infection.**

**Reference Answer**

"exp": "Ans: C. -PCR for hepatitis C virus (Ref Harrison 19/e p2017, 2018: 18/e p2551)Positive Anti-HBcAg IgG: Mere indicator of hepatitis B infection in past. Positive Anti-HCV antibody: Most sensitive indicator of HCV infection - Presence of HCV. Presence of Anti-HCV antibody points to hepatitis C infection a Hence HCV-RNA levels determined to establish chronic hepatitis C diagnosis before staing Interferon therapy. Requires molecular amplification by PCR. Testing for HCV RNA or repeat anti-HCV testing later stages - Necessary to establish diagnosis.",  
"cop":3

Figure 162: **Medical Knowledge Question Answer: Case 8.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "A 22 years old male presented with history of fever, sore throat and enlarged neck lymph nodes. He was ordered a Paul-Bunnell test with a suspicion of Infectious mononucleosis. What is the immunological basis behind the use of this test?"

"opa": "Heterophile antibody test", "opb": "Complement mediated agglutination reaction", "opc": "Homophile antibody test", "opd": "Latex agglutination test"

**O1-preview**

**Answer:**

**A. Heterophile antibody test**

**Explanation:**

The Paul-Bunnell test is a serological test used to detect heterophile antibodies in the patient's serum, which is a characteristic finding in infectious mononucleosis caused by the Epstein-Barr virus (EBV).

**Heterophile Antibodies:**

- **Definition:** Heterophile antibodies are antibodies that react with antigens from unrelated species. In the case of infectious mononucleosis, these antibodies can agglutinate red blood cells from animals such as sheep, horses, or cows.
- **Formation in EBV Infection:** During an EBV infection, the virus stimulates B lymphocytes to produce a broad range of antibodies, including heterophile antibodies.
- **Detection in Paul-Bunnell Test:** The test mixes the patient's serum with animal red blood cells and looks for agglutination (clumping). A positive result indicates the presence of heterophile antibodies.

**Why Other Options Are Incorrect:**

- **B. Complement mediated agglutination reaction:**
  - This reaction involves the complement system, a part of the immune system that enhances the ability of antibodies to clear pathogens. It is not the basis of the Paul-Bunnell test.
- **C. Homophile antibody test:**
  - The term "homophile antibodies" is not standard in immunology. It may be a distractor term. The correct term is "heterophile antibodies."
- **D. Latex agglutination test:**
  - Latex agglutination tests use latex particles coated with antigens or antibodies to detect their counterparts in a sample. While this method is used in various diagnostic tests, it is not the principle behind the Paul-Bunnell test.

**Summary:**

The Paul-Bunnell test detects heterophile antibodies produced during an EBV infection, which cause agglutination of animal red blood cells. Therefore, the immunological basis of the test is the presence of heterophile antibodies.

**Reference Answer**

"exp": "Answer- A. Heterophile antibody test" Epstein-Bat virus (EBV) infection includes specific antibodies to EBV and various unrelated non-EBV heterophile antibodies. These heterophile

---

antibodies react to antigens from animal RBCs. Sheep RBCs agglutinate in the presence of heterophile antibodies and are the basis for the Paul-Butrnell test.",  
"cop":1

---

Figure 163: **Medical Knowledge Question Answer: Case 9.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

**Prompt**

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

"question": "A 28 year old man has lenticonus and end stage renal disease now. His maternal uncle also died of the same illness. What is the most likely diagnosis?",  
"opa": "Autosomal dominant polycystic kidney disease", "opb": "Autosomal recessive polycystic kidney disease", "opc": "Oxalosis", "opd": "Alport syndrome"

**O1-preview**

**Answer:** Alport syndrome (Option D)

**Explanation:**

The patient is a 28-year-old man presenting with:

- End-stage renal disease (ESRD)
- Lenticonus
- A family history of similar illness in his maternal uncle

These clinical features point toward **Alport syndrome**. Here's why:

**Key Features Explained:****1. Lenticonus:**

- Lenticonus refers to a conical protrusion of the lens capsule and cortex, typically anteriorly (anterior lenticonus).
- **Alport syndrome** is characteristically associated with anterior lenticonus.
- This ocular manifestation is due to a defect in the type IV collagen found in the basement membranes of the lens.

**2. End-stage Renal Disease (ESRD):**

- Alport syndrome leads to progressive renal failure due to defective glomerular basement membrane.
- Patients often develop hematuria, proteinuria, and ultimately ESRD in early adulthood.

**3. Family History Through Maternal Line:**

- The patient's **maternal uncle** (mother's brother) also had the same illness and died from it.
- Alport syndrome is most commonly inherited in an **X-linked dominant** pattern.
- In X-linked inheritance:
  - Males are more severely affected because they have only one X chromosome.
  - Carrier mothers can pass the mutated gene to their sons (affected) and daughters (carriers or less severely affected).

**Exclusion of Other Options:**

- **Option A: Autosomal Dominant Polycystic Kidney Disease (ADPKD)**
  - ADPKD presents with multiple bilateral renal cysts, hypertension, and possible cysts in other organs (e.g., liver).
  - **Not associated with lenticonus.**

- Inherited in an autosomal dominant pattern, which doesn't fit the maternal lineage described.
- **Option B: Autosomal Recessive Polycystic Kidney Disease (ARPKD)**
  - ARPKD typically presents in infancy or early childhood with renal cysts and liver fibrosis.
  - Not associated with lenticonus.
  - The inheritance pattern does not match the family history.
- **Option C: Oxalosis (Primary Hyperoxaluria)**
  - A metabolic disorder leading to excessive oxalate production, renal stones, nephrocalcinosis, and renal failure.
  - Not associated with lenticonus or the described inheritance pattern.

**Conclusion:**

Given the combination of **lenticonus**, **early-onset ESRD**, and a **family history consistent with X-linked inheritance**, **Alport syndrome** is the most probable diagnosis. Therefore, the most likely diagnosis is **Alport syndrome (Option D)**.

**Reference Answer**

"exp": "Presentation of male patient with **lenticonus and end stage renal disease with a family history of renal disease is highly suggestive of Alport syndrome**. AR polycystic kidney is ruled out because the **age of presentation in ARPKD is childhood** and **most of the affected children do not survive beyond their childhood**. AD polycystic kidney is ruled out because there is **no association of ADPKD with lenticonus as is mentioned in our question**. Alport syndrome is manifest by hematuria with progression to chronic renal failure, accompanied by nerve deafness and various eye disorders, including lens dislocation, posterior cataracts, and corneal dystrophy. In about 85% cases, it is inherited as an X-linked trait. So, males express the full syndrome, and females are carriers in whom manifestations of disease are typically limited to hematuria. Autosomal recessive and autosomal dominant forms also exist in which both the sexes are equally susceptible. In Alport syndrome, Hematuria is the earliest manifestation and the sensorineural deafness is the commonest extra renal abnormality.",  
 "cop": 4

Figure 164: **Medical Knowledge Question Answer: Case 10.** A example illustrating the capability of o1-preview to respond to questions that arises in medical exams. Yellow indicates partial consistency between the o1-preview response and the reference answer, while green signifies complete consistency.

## A.20 Art Education

### Prompt

Make a list of non-human objects that represent some aspect of what you consider to be "your own" cultural community. Or choose an object from a cultural community where you have studied or spent time. Freewrite this list for at least 1–3 minutes without censoring yourself.

Next, review the list and choose the non-human object that resonates the most with layered meaning or significance for you. You do not need to know what or why yet!

Now, try writing a dramatic monologue from that object's point of view. Animate the object. How does it think? What is its history? Does the object consider itself to be part of a family? What losses or near-death experiences has the object encountered? Try doing some research about the object to help to create a voice that is more honest and convincing. For vocabulary, use terms that you find on pages that discuss the object's manufacturing or materials. What idioms and vernacular does your object use? What is the object's age and how is that reflected in the diction and syntax that you use to convey the object's voice?

### O1-preview

List of Non-Human Objects Representing My Cultural Community:

- Worn leather-bound journal
- Antique pocket watch
- Wooden rocking chair
- Cast iron skillet
- Quilted patchwork blanket
- Mason jar
- Old family piano
- Handcrafted walking cane
- Rustic farmhouse table
- Weathered fishing rod
- Stained coffee mug
- Silver locket necklace
- Vintage record player
- Ceramic cookie jar
- Tin oil lamp
- Iron horseshoe
- Embroidered handkerchief
- Harvest basket
- Porcelain tea set
- Woolen scarf

Monologue from the Quilted Patchwork Blanket's Point of View: I am the quilted patchwork blanket, draped over the back of the old sofa. Each square of my fabric holds a story, stitched together with care by Miss Eleanor decades ago. Her hands pieced me from worn cotton and faded calico—dresses and shirts long outgrown. Threads of crimson and indigo weave through me, preserving memories of simpler times. The floral print? From a dress worn to a spring picnic. The plaid? A flannel shirt that braved the winter cold. I've offered comfort through the years, cradling newborns and sheltering children from stormy nights. Once, an ember burned my fabric, but Miss Eleanor patched me with a piece of her wedding gown, making me whole again. Though my colors have faded and my edges frayed, I still hold memories within my seams. I think of the old saying, "We are all cut from the same cloth." As long as there are stories to hold and hearts that cherish the past, I am fulfilled, a guardian of family lore.

### Reference Answer

Say chair and feel abstraction's weight, not classic mission, fabric, or rattan cane, not head or director, but this decorative art unmoored from its coupling: rocking\_\_\_\_\_, electric \_\_\_\_\_. Hear its strained, cherry purpose, scuffed loosening from rooms, as if ch might fall, leaving air stacked in broken corners. Tilt it to cheer and clink glasses while the meat chars, left too long by the sitter, the soldier, the rocker, the curator, the one who certifies preservation and forbids touch. Survivor of the War of Jenkins' Ear, this wooden Spaniard is severed, too, from use, a stretched rawhide reminder, shrunk to sit out dispute. See ya, one might say in its native tongue and we do not sit on la silla, we don't recline in others' tongues because we won an ordinary household article, our museums tell us so.

Figure 165: **Art Education: Case 3.** This case examines o1-preview's ability to create a dramatic monologue from the perspective of a non-human object. The prompt, drawn from Cahnmann-Taylor, Melisa's *Giving Meaning and Voice to Objects* in her work *Arts-based Research in Education*, encourages reflective storytelling through objects [15]. The reference answer is the poem written by Cahnmann-Taylor herself.

### **Prompt**

How do the artist Nick Cave's Soundsuits influence high school students' exploration of identity, creativity, and self-expression, particularly in challenging societal norms and fostering critical thinking?

### **O1-preview**

Nick Cave's \*\*Soundsuits\*\* profoundly influence high school students by providing a dynamic platform for exploring identity, fostering creativity, encouraging self-expression, challenging societal norms, and promoting critical thinking.

#### **Exploring Identity:**

The Soundsuits are elaborate, wearable sculptures that completely conceal the wearer's body, masking all identifiers such as race, gender, and class. For high school students grappling with their own identities, this concept introduces the idea of "anonymity as empowerment". By obscuring external identifiers, the Soundsuits invite individuals to consider who they are beyond societal labels and stereotypes. This allows students to:

- Reflect on Personal Identity: Encourage introspection about their own identities without the influence of societal expectations.
- Empathize with Others: Gain a deeper understanding of the experiences of those who may feel unseen or marginalized.
- Discuss Diversity and Inclusion: Open dialogues about acceptance and understanding of different identities in their communities.

#### **Fostering Creativity:**

Nick Cave's use of found objects and unconventional materials stimulates students' imaginations. The vibrant colors, textures, and sounds generated by the suits inspire students to:

- Experiment with Materials: Encourage the use of everyday objects in art, promoting sustainability and innovation.
- Explore Multimedia Art Forms: Combine visual arts with performance, music, and movement.
- Develop Original Artwork: Inspire the creation of personal art projects that reflect their unique perspectives.

#### **Encouraging Self-Expression:**

The performative aspect of the Soundsuits allows students to express themselves beyond verbal communication. By engaging with movement and sound, students can:

- Express Emotions Physically: Use body language and dance to convey feelings and stories.
- Build Confidence: Perform in a way that feels safe and liberating due to the anonymity provided by the suits.
- Collaborate with Peers: Work together to create performances, fostering teamwork and communication skills.

#### **Challenging Societal Norms:**

Nick Cave created the first Soundsuit in response to racial violence and profiling, specifically the beating of Rodney King in 1991. The Soundsuits challenge viewers to confront issues of race, identity, and prejudice. For high school students, engaging with this work encourages them to:

- Question Prejudices: Examine their own biases and the stereotypes prevalent in society.
- Understand Social Justice Issues: Learn about historical and contemporary social movements.

- Advocate for Change: Empower them to use art as a form of protest and social commentary.

#### Promoting Critical Thinking:

Analyzing the Soundsuits involves interpreting complex themes and symbols. This process enhances students' critical thinking skills by encouraging them to:

- Interpret Artistic Meaning: Delve into the symbolism and messages conveyed through the artwork.
- Connect Art to Social Contexts: Understand how art reflects and influences societal issues.
- Engage in Thoughtful Discussion: Participate in debates and discussions that challenge their perspectives.

#### Educational Applications:

Teachers can incorporate Nick Cave's Soundsuits into the curriculum through interdisciplinary projects that combine art, history, social studies, and literature. Examples include:

- Art Projects: Students create their own Soundsuits or wearable art pieces that represent aspects of their identity or social issues important to them.
- Performances: Students choreograph and perform dances or movements while wearing their creations, exploring the relationship between art and performance.
- Research Assignments: Investigate the historical and cultural context of the Soundsuits, including studies on race relations and the role of art in activism.
- Reflective Writing: Compose essays or journals reflecting on their experiences and the impact of societal norms on individual expression.

#### Conclusion:

Nick Cave's Soundsuits serve as a powerful educational tool that resonates with high school students. They offer a tangible way to explore and communicate complex ideas about identity and society. By engaging with the Soundsuits, students develop a deeper understanding of themselves and the world around them, becoming more empathetic, creative, and critically minded individuals. This engagement not only enriches their educational experience but also empowers them to become agents of change in their communities.

#### **Reference Answer**

Nick Cave's art, particularly his "Soundsuits," has a significant influence on high school students by offering them a powerful tool for exploring identity, transformation, and self-expression. Through Cave's wearable sculptures, students engage with the concept of persona, allowing them to explore various facets of their identity in an artistic and experiential way. By creating their own Soundsuits, students are encouraged to embody alter egos, gaining firsthand experience in how artistic expression can transcend societal norms and personal boundaries.

This process fosters critical thinking, creativity, and introspection as students use Cave's work to challenge societal constructs around race, gender, and individuality. It also allows them to experiment with materials, performance, and storytelling, helping them understand how art can be a medium for self-discovery and social commentary. Nick Cave's work not only enhances students' artistic skills but also deepens their empathy and awareness of diverse perspectives, empowering them to use art as a means of engaging with the world around them.

Figure 166: **Art Education: Case 4.** This case evaluates o1-preview's ability to analyze the influence of Nick Cave's *Soundsuits* on high school students' exploration of identity, creativity, and self-expression [24].

**Prompt**

What is the relationship between art education and disability studies, and how does integrating disability studies enhance the inclusivity and critical inquiry in art education?

**O1-preview**

The relationship between art education and disability studies is an evolving and increasingly significant intersection that enriches both fields. Art education traditionally focuses on the development of artistic skills, appreciation of art history, and fostering creativity. Disability studies, on the other hand, is an interdisciplinary field that examines disability from a social, cultural, and political perspective, moving beyond mere medical or individual models of disability. Integrating disability studies into art education creates a dynamic synergy that enhances inclusivity and promotes critical inquiry within the educational landscape.

**\*\*Understanding Disability Through Art\*\***

By incorporating disability studies into art education, educators and students gain a deeper understanding of disability as a social construct rather than just a physical or medical condition. This integration allows for exploration of themes such as representation, identity, and accessibility within the arts. It challenges preconceived notions and stereotypes about disability, encouraging a more nuanced and empathetic perspective.

**\*\*Enhancing Inclusivity\*\***

1. **\*\*Inclusive Curriculum Development\*\*:** Integrating disability studies into the art curriculum ensures that the content is reflective of diverse experiences. This includes studying artworks by artists with disabilities, exploring themes of disability in art history, and recognizing the contributions of artists who have navigated and challenged societal barriers.
2. **\*\*Representation Matters\*\*:** When students see disability represented authentically and respectfully in the arts, it fosters a sense of belonging and validation. This representation helps students with disabilities to see themselves reflected in the curriculum and encourages all students to appreciate diversity.
3. **\*\*Accessible Art Practices\*\*:** Disability studies advocate for accessibility, which translates into art education through adaptive tools, materials, and methods that accommodate various needs. This could involve using tactile art forms, incorporating assistive technologies, or adapting equipment to ensure all students can participate fully.
4. **\*\*Universal Design for Learning (UDL)\*\*:** Applying UDL principles means designing art lessons that are accessible to all learners from the outset. This proactive approach minimizes barriers and promotes equal opportunities for engagement and expression.

**\*\*Promoting Critical Inquiry\*\***

1. **\*\*Challenging Ableism\*\*:** Integrating disability studies prompts critical examination of ableist attitudes and practices within the arts and society. Students are encouraged to question normative standards of ability and beauty, and to recognize how these standards marginalize individuals with disabilities.
2. **\*\*Interdisciplinary Connections\*\*:** Disability studies bridge art education with sociology, ethics, history, and other disciplines. This interdisciplinary approach enriches students' understanding and promotes critical thinking about broader social issues.
3. **\*\*Analyzing Representation\*\*:** Students critically analyze how disability has been depicted in different art forms and historical contexts. This analysis reveals the evolving perceptions of disability and stimulates discussions about agency, voice, and representation in the arts.

4. \*\*Empowering Voices\*\*: Encouraging students with disabilities to express their experiences through art empowers them and adds valuable perspectives to the classroom. It fosters an environment where diverse narratives are not only accepted but celebrated.

#### **\*\*Fostering Empathy and Social Justice\*\***

Integrating disability studies into art education cultivates empathy by allowing students to engage with experiences different from their own. It promotes social justice by addressing inequalities and advocating for the rights and inclusion of individuals with disabilities. Through art, students can explore and challenge the societal structures that contribute to discrimination and exclusion.

#### **\*\*Encouraging Innovation and Creativity\*\***

Accessibility often necessitates creative problem-solving, which can lead to innovative artistic practices. Artists with disabilities have developed unique methods and styles, enriching the art world. Exposure to these practices inspires students to think creatively and pushes the boundaries of traditional art forms.

#### **\*\*Conclusion\*\***

The relationship between art education and disability studies is mutually beneficial. Integrating disability studies into art education not only enhances inclusivity by ensuring all students are represented and can participate meaningfully but also enriches critical inquiry by challenging students to think deeply about social constructs, representation, and justice. It prepares students to be more empathetic, socially conscious individuals who value diversity and are equipped to contribute positively to an inclusive society.

By embracing disability studies, art education becomes a transformative space where all students can explore, create, and learn from one another's experiences, ultimately enriching the educational experience for everyone involved.

#### **Reference Answer**

Art education, traditionally aligned with special education, often follows a medical model that views disability as a deficiency requiring therapeutic intervention. However, disability studies offer a more expansive and inclusive perspective, recognizing disability as a cultural and social identity rather than merely a medical condition.

The interdisciplinary dialogue between art education and disability studies can challenge outdated metaphors and stereotypes associated with disability. By incorporating disability studies into art education, educators can promote critical thinking, self-awareness, and social justice. This integration encourages students to explore disability through creative expression, visual culture, and performance, allowing for more nuanced representations of disability in art and fostering an inclusive environment that acknowledges the lived experiences of disabled individuals.

In conclusion, the relationship between art education and disability studies lies in their shared goal of challenging societal norms and advocating for inclusivity, critical inquiry, and the representation of diverse bodies and identities through art. By merging these fields, art educators can better serve all students, including those with disabilities, while contributing to broader cultural and social discourses surrounding disability.

Figure 167: **Art Education: Case 5.** This case explores the relationship between art education and disability studies, examining how integrating disability studies enhances inclusivity and promotes critical inquiry in art education. The prompt is based on Derby's *Disability Studies and Art Education* [25], which frames disability as a cultural and social identity.

**Prompt**

How do you think art education contributes to the development of an individual's personality?

**O1-preview**

Art education plays a significant role in shaping an individual's personality by nurturing a wide range of cognitive, emotional, and social skills. Here are several ways in which art education contributes to personal development:

## 1. Fosters Creativity and Imagination:

Art encourages individuals to think outside the box and explore new ideas. Engaging in artistic activities stimulates the imagination, allowing people to approach problems creatively and consider multiple solutions.

## 2. Enhances Emotional Expression:

Art provides a medium for expressing feelings that might be difficult to articulate verbally. Through creating art, individuals can convey emotions and thoughts, leading to improved emotional intelligence and self-awareness.

## 3. Develops Critical Thinking and Problem-Solving Skills:

Artistic creation often involves planning, experimenting, and making decisions about composition, color, and technique. This process enhances critical thinking and problem-solving abilities that are transferable to other areas of life.

## 4. Builds Confidence and Self-Esteem:

Completing an artwork can give individuals a sense of accomplishment. Positive feedback and personal satisfaction from artistic endeavors boost confidence and encourage a willingness to take on new challenges.

## 5. Improves Fine Motor Skills and Coordination:

Engaging in activities like drawing, painting, or sculpting refines fine motor skills and hand-eye coordination. This physical development is especially beneficial in early childhood but continues to enhance dexterity at any age.

## 6. Encourages Patience and Perseverance:

Creating art often requires time and repeated efforts to achieve the desired result. This teaches patience and the value of perseverance, fostering a growth mindset that views challenges as opportunities to learn.

## 7. Promotes Cultural Awareness and Empathy:

Studying art exposes individuals to different cultures, histories, and perspectives. Understanding the context behind various artworks cultivates empathy and appreciation for diversity.

## 8. Enhances Communication Skills:

Art is a universal language that transcends words. Learning to interpret and create visual messages improves overall communication skills, enabling individuals to convey complex ideas and emotions effectively.

## 9. Provides Stress Relief and Emotional Well-Being:

Artistic activities can be therapeutic. Focusing on creating art allows individuals to enter a state of flow, reducing stress and promoting relaxation. This contributes positively to mental health and emotional well-being.

## 10. Stimulates Academic Achievement:

Research suggests that involvement in the arts can improve performance in other academic subjects. Skills developed through art—such as critical thinking, attention to detail, and persistence—enhance learning across disciplines.

## 11. Cultivates Social Skills and Collaboration:

Participating in group art projects or classes fosters cooperation and communication. Sharing ideas and giving and receiving feedback enhance interpersonal skills and the ability to work effectively with others.

## 12. Sparks Curiosity and Lifelong Learning:

Art education ignites curiosity about the world, encouraging individuals to ask questions and seek new knowledge. This inquisitive mindset supports continuous personal growth and adaptability. In summary, art education contributes significantly to personal development by enriching an individual's emotional depth, intellectual capabilities, and social interactions. By engaging with art, people develop a more nuanced understanding of themselves and the world around them, leading to a well-rounded and resilient personality.

**Reference Answer**

Aesthetic and Emotional Growth: Art education enhances an individual's ability to appreciate beauty, fostering a deeper emotional intelligence. Through engaging with art, students learn to interpret emotions and experiences, which strengthens their capacity for empathy and emotional expression.

Creativity and Critical Thinking: Art education encourages creativity and problem-solving. Students learn to think critically about social, cultural, and personal issues, which broadens their perspectives and helps them find innovative solutions in various aspects of life.

Moral and Ethical Development: Art education often involves exploring themes of morality, ethics, and justice. By engaging with these themes, students develop a sense of social responsibility and personal values, which are crucial for becoming a thoughtful and active member of society.

Socialization and Communication: Art provides a means for students to communicate and connect with others, both within and across cultures. This strengthens their social skills and fosters a greater understanding of diversity and inclusion.

Self-expression and Confidence: Art education gives students a platform to express their personal ideas and feelings, boosting their self-confidence and helping them develop a strong sense of identity.

Figure 168: **Art Education: Case 6.** This case explores the impact of art education on personal development. The reference answer emphasizes art's role in moral and ethical development, as well as its contribution to socialization and self-expression [160].

## A.21 Content Summarization

### **Prompt**

Please summarize the content in one sentence: Here are the best bits we heard backstage at the Royal Albert Hall in London. In La La Land, Emma Stone plays aspiring actress Mia Dolan. So does the best actress winner think Mia could ever win a Bafta or Oscar? After a dramatic pause, she delivered her verdict. "Hopefully!" "It's been like a whirlwind," said Spider-Man actor Tom Holland, winner of the Rising Star Award. "I'm so happy everything seems to be working out." The 20-year-old Brit isn't wrong there. Having made his screen debut in tsunami disaster movie The Impossible, he first appeared as Spidey in Captain America: Civil War and recently finished filming for Spider-Man: Homecoming. But Tom admitted the super hero role has its drawbacks. "Ever since I was a kid I've wanted to be Spider-Man. It's been so fun, but my legs are in bits right now from spider poses after spider poses. "So I'll be sitting down for the rest of the evening!" It was perhaps inevitable that Meryl Streep's speech at the Golden Globes would come up at the Baftas. Asked about Donald Trump's response, supporting actress winner Viola Davis didn't hold back. "Anyone who labels Meryl Streep an 'overrated' actress obviously doesn't know anything about acting," she said. "That's not just directed towards Donald Trump - that's directed towards anyone." She didn't stop there. "This is someone who is the master at her skill and she has lasted for 40 years in a very difficult profession," Viola continued. "One of the things people have to know about this woman is that she is the most honourable, accessible human being you could possibly want to meet." Best actor winner Casey Affleck said he'd grabbed a word with Meryl Streep after the Baftas ceremony. "She was taking pictures of people... and I told her how much her speech at the Golden Globes meant to all of us and how grateful I was that she did it and kicked in the door a little bit," he said. The Manchester by the Sea star said he hoped more actors would speak out. "There is a big audience for these awards shows... and I have to say I'm very proud to be a part of the arts community. "I don't always say some of the things I would like to say in those opportunities because there are people like Meryl Streep who say them much better than I can - and if they are going to be said it should be said very, very well because they are important." "I'm a little bit wobbly," confessed Dev Patel a few minutes after winning the supporting actor prize for Lion. "I really did not expect it, we have gone to so many awards ceremony and this one is where everything changed, on home turf with my family." Dev recalled how he had made his acting debut a decade ago on E4's teen drama Skins. "I remember the first time I ever stepped on a film set, I never knew what a boom mic was," he laughed. "My mum took me to an open casting of Skins after she saw an advert in Metro newspaper and 10 years on we are here at the Baftas - that is pretty amazing." Justin Hurwitz, who won the Bafta for his original music in La La Land, said that he'd written a title song that didn't end up in the movie. "It was going to be in the middle of the movie but we nixed that, and then we toyed with using it in the end credits, and we nixed that at the last minute." And how about this? Another Day of Sun, from the famous traffic jam scene, was cut "for many months" before being put back in as the big opening number. Mel Brooks, who was awarded the prestigious Bafta Fellowship, was asked about whether he might write a comedy about Donald Trump. "I'm not afraid of him, I don't think he's dangerous," said the 90-year-old writer, actor and producer. "I think he's mostly an entertainer, a guy who wants audiences to love him. "What I'm afraid of are all the guys around him, the people who whisper in his ears, like the people who whispered in George W Bush's ears and we got the Iraq War... "I just hope that Trump stays the egomaniac he is, listens to no-one and then we'll all be safe. But if he believes these guys we're all in trouble." Mel was in a lighter mood as he plugged his new musical Young Frankenstein, which will open in London's West End this year after a run in Newcastle. "I think, modestly speaking, it will be sensational," he chuckled. Follow us on Facebook, on Twitter @BBCNewsEnts, or on Instagram at bbcnewsents. If you have a story suggestion email entertainment.news@bbc.co.uk.

### **O1-preview**

At the Baftas in London, stars including Emma Stone, Tom Holland, Viola Davis, Casey Affleck, Dev Patel, Justin Hurwitz, and Mel Brooks shared backstage reactions to their awards, discussed industry topics like Meryl Streep's Golden Globes speech and Donald Trump, and reflected on their careers.

### **Reference Answer**

The Bafta awards had laughter, passion and plenty of politics.

Figure 169: **Content Summarization: Case 3.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

**Prompt**

Please summarize the content in one sentence: More than 300 people entered Alternative Model of the Year with the winner being decided by judges in Newcastle on Saturday. Vintage-fan Zoe Waters, 19, from Bishop Auckland, is a student at Newcastle University. Ms Waters said she was shocked to be declared the winner. "I've always dressed a bit crazy," she said. "I don't know why I wanted to be different, I guess I just always choose to wear what I like and look how I want to look rather than worrying about what's in fashion." She said she now plans to pursue a career in modelling. "I'm only 5ft 3in tall and I never wanted to look normal so I never thought I could actually be a model, I hadn't realised how big the alternative modelling world is." Organiser Kieran Martin said: "There is a lot of prejudice and hatred for people from sub-cultures so we wanted to create something that would celebrate the different lives we have, we are proud of what we are achieving." The final was held at Northumbria University in Newcastle.

**O1-preview**

Nineteen-year-old **vintage** enthusiast and Newcastle University student, Zoe Waters from Bishop Auckland, **won** the **Alternative Model** of the Year competition in Newcastle and **now plans to pursue a modeling career embracing her unique style**.

**Reference Answer**

A philosophy student who wears **vintage clothing** has **won** a national contest for **alternative models**.

Figure 170: **Content Summarization: Case 4.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

### **Prompt**

Please summarize the content in one sentence: The Association of School and College Leaders says England's schools have had to make more than £1bn savings this year, rising to £3bn by 2020. The government says school funding is at a record £40bn, with rises ahead. Education Secretary Justine Greening will hear heads' cash grievances at Friday's ASCL conference in Birmingham. She is due to address the union, which has published a survey of its members on the issue. It suggests schools are finding it difficult to make savings without cutting provision and that things are predicted to get worse over the next two years. Cost pressures are rising as greater pay, pension and national insurance costs are having to be covered from school budgets. ASCL complains a new funding formula for schools has reduced the basic level of school funding going forwards by too much. The meeting comes two days after requests for more money to spend on daily school costs were ignored by the chancellor in the Budget. Philip Hammond however did pledge £500m for school buildings, mainly new free schools - some of which could be grammar schools. One respondent said his school was moving to a "bare bones education", in which "the components that make education special and enjoyable are being eroded away". Some 95% of the 1,054 heads, deputies and senior teachers responding to the survey said they had cut back on support services - including equipment and materials, as well as mental health and special needs support. More than eight out of 10 said class sizes had increased - a claim strongly refuted by the Department for Education. And more than two-thirds said they had cut back on activities like clubs and trips. Just under three-quarters of respondents with GCSE-level classes said they had cut courses and just over three-quarters of heads with A-level students said they had also reduced subjects. Foreign modern languages, music, arts and drama were among subjects removed at A-level. Another said: "Through no fault of their own, students will have restricted subject choices, in larger class sizes with less pastoral support, whilst still being expected to perform at the highest of standards - nonsense!" One head said his school may have to axe its sixth form provision for next year and another said his school was starting to "creak" with all staff working to full capacity. Interim general secretary, Malcolm Trobe, said: "School leaders will do their utmost to protect provision, as they always do, but they cannot provide everything that is asked of them without the resources they need. "Unless the government invests more in the education system, there will be a significant impact on the lives and life chances of young people." A spokesman for the DfE said: "As this week's Budget demonstrates, the government is determined to ensure every child has access to a good school place and is given the opportunity to fulfil their potential. "The government has protected the core schools budget in real terms since 2010, with school funding at its highest level on record at more than £40bn in 2016-17 - and that is set to rise, as pupil numbers rise over the next two years, to £42bn by 2019-20."

### **O1-preview**

The Association of School and College Leaders warns that schools in England face significant funding shortfalls requiring billions in savings—leading to cuts in educational provision—while the government asserts that school funding is at record levels and will continue to rise.

### **Reference Answer**

Head teachers say they are axing GCSE and A-level subjects, increasing class sizes and cutting support services as they struggle with school funding.

Figure 171: **Content Summarization: Case 5.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

**Prompt**

Please summarize the content in one sentence: Wholesale prices have dropped and motoring organisations have suggested this merits reduced pump prices. Tesco and Morrisons are cutting prices by 2p a litre on Friday, and Asda and Sainsbury's said they would reduce prices by up to 2p a litre on Saturday. But drivers are being warned that prices can vary in different areas. The average UK price for a litre of unleaded petrol was 118.83p on Thursday, and 120.88p for a litre of diesel, according to the latest figures from Experian Catalyst. Ten days earlier, the average unleaded price was 120p a litre, and diesel cost 122.06p. The oil price and wholesale prices fell sharply at the start of March, and motoring organisations have argued that this should have been feeding through to another 2p cut per litre in prices at the pumps. The RAC said motorists could feel "aggrieved" that prices had not fallen further, earlier. Asda said it had dropped prices twice in two weeks, and had a national price cap to ensure motorists were dealt with equitably. Morrisons said it had also made a second cut in two weeks. Tesco is dropping petrol and diesel prices by 2p a litre over the course of Friday afternoon at all its outlets, followed by Sainsbury's on Saturday. Luke Bosdet, of the AA, said that there was general concern that so-called supermarket fuel price wars did not actually benefit all drivers across the country. He claimed prices fell the most in areas where there were a range of supermarkets located close to each other. In other areas with less competition, the same cuts were not as likely to be seen. "We would urge motorists to look around to find a better price. There are petrol price apps that can help," he said.

**O1-preview**

UK supermarkets are cutting fuel prices by up to 2p per litre after a drop in wholesale oil prices, but motoring organizations argue the reductions should be greater and warn that not all areas may benefit equally due to varying competition.

**Reference Answer**

Motorists will see an acceleration in fuel price cuts over the weekend as supermarkets take up to 2p off a litre of petrol and diesel.

Figure 172: **Content Summarization: Case 6.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

---

**Prompt**

Please summarize the content in one sentence: Police were alerted to an incident in Coilt Crescent in the Highland village at about 22:00 on Sunday. A man wearing dark clothing and carrying what appeared to be a gun entered the house and demanded money from the homeowner, police said. A man was later arrested in Cumbria and has been assisting officers investigating the incident. A four-figure sum of money was stolen from the house in Drumnadrochit's Kilmore area. Police Scotland said: "A high visibility police presence remains in area at this time to provide reassurance to the local community. "Police are keen to hear from members of the public who may have seen any suspicious activity in the area prior to the incident. "In addition, officers would like to hear from anyone who may have seen a silver VW Passat in the area earlier that day."

**O1-preview**

Police are investigating **an armed robbery in Drumnadrochit's Kilmore area**, where a man carrying what appeared to be a gun stole a four-figure sum from a house, **with a man later arrested in Cumbria** assisting their inquiries, **and are appealing for witnesses, especially anyone who saw a silver VW Passat.**

**Reference Answer**

**A man has been arrested in Cumbria** following reports of **an armed robbery at a house in Drumnadrochit**.

Figure 173: **Content Summarization: Case 7.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

### **Prompt**

Please summarize the content in one sentence: The 48-year-old man suffered a suspected heart attack five minutes into the second half of Kilmarnock's home match against Hibernian. Paramedics and medical staff from both clubs treated the man until an ambulance arrived at Rugby Park stadium and he was taken to a nearby hospital. Both managers agreed to abandon the match out of respect for the fan. The stadium was hushed into silence when medical staff and club doctors attended the stricken fan in the west stand. The referee then stopped the match with the teams drawing 1-1. Players left the pitch as the gravity of the situation became evident. Kilmarnock chairman Michael Johnston said: "The length of time the players were off the park was a problem but the overriding factor was respect for the fan and his family. "He collapsed and was tended to by medical staff of both clubs and was taken to Crosshouse Hospital by ambulance." He added that he could not reveal any more information about the fan's identity other than to say he was a long-standing season-ticket holder. Later, a Police Scotland spokesman said the man died in hospital. Kilmarnock manager Kenny Shiels pointed out that it was the third time he had been present at such an incident while manager at the Rugby Park side. In March 2012, midfielder Liam Kelly's father died following the Ayrshire side's League Cup final win over Celtic. A month later, an Inverness Caledonian Thistle fan died during that club's visit to Rugby Park.

### **O1-preview**

During Kilmarnock's home match against Hibernian, a 48-year-old fan suffered a fatal heart attack five minutes into the second half, leading both teams to agree to abandon the game out of respect.

### **Reference Answer**

A man has died in hospital after collapsing half-way through a Scottish Premier League match, police say.

Figure 174: **Content Summarization: Case 8.** A example illustrating the capability of o1-preview to respond to summarize news content. Yellow indicates partial consistency between the o1-preview's response and the reference answer, while green signifies complete consistency.

---

## References

- [1] 1988 IMO Problems/Problem 6. [https://artofproblemsolving.com/wiki/index.php?title=1988\\_IMO\\_Problems/Problem\\_6](https://artofproblemsolving.com/wiki/index.php?title=1988_IMO_Problems/Problem_6). [Online; accessed 19-Sep-2024]. Art of Problem Solving, 1988.
- [2] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).
- [3] Jean-Baptiste Alayrac et al. “Flamingo: a Visual Language Model for Few-Shot Learning”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23716–23736.
- [4] Suzi A Aleksander et al. “The gene ontology knowledgebase in 2023”. In: *Genetics* 224.1 (2023), iyad031.
- [5] Joanna S Amberger et al. “OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders”. In: *Nucleic acids research* 43.D1 (2015), pp. D789–D798.
- [6] Claudine Badue et al. *Self-Driving Cars: A Survey*. 2019. arXiv: [1901.04407 \[cs.RQ\]](https://arxiv.org/abs/1901.04407). URL: <https://arxiv.org/abs/1901.04407>.
- [7] DL Bandalos. *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.
- [8] SHIV BANSAL. “The Impact of Artificial Intelligence and Machine Learning on Financial Markets”. In: *Innovative Research Thoughts* 7.4 (2021), pp. 217–222.
- [9] Francesco Barbieri et al. “TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification”. In: *Proceedings of Findings of EMNLP*. 2020.
- [10] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.
- [11] Yoshua Bengio. “From systematic generalization to machine intelligence: Benefits and challenges”. In: *arXiv preprint arXiv:1911.01547* (2019).
- [12] Maciej Besta et al. “Graph of Thoughts: Solving Elaborate Problems with Large Language Models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 17682–17690.
- [13] Kevin Black et al. “Training diffusion models with reinforcement learning”. In: *arXiv preprint arXiv:2305.13301* (2023).
- [14] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [15] Melisa Cahnmann-Taylor and Richard Siegesmund. “Arts-based research in education”. In: *Foundations for practice. New York: Routledge* (2008).
- [16] Hongmin Cai et al. “Exploring Multimodal Approaches for Alzheimer’s Disease Detection Using Patient Speech Transcript and Audio Data”. In: *arXiv preprint arXiv:2307.02514* (2023).
- [17] Wenhui Chen et al. “MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training”. In: *arXiv preprint arXiv:2403.09611* (2024).
- [18] Paul F Christiano et al. “Deep reinforcement learning from human preferences”. In: *Advances in neural information processing systems* 30 (2017).
- [19] Zheng Chu et al. “Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1173–1203. URL: <https://aclanthology.org/2024.acl-long.65>.

- 
- [20] Haixing Dai et al. “AD-AutoGPT: An Autonomous GPT for Alzheimer’s Disease Infodemiology”. In: *arXiv preprint arXiv:2306.10095* (2023).
- [21] Haixing Dai et al. “Chataug: Leveraging chatgpt for text data augmentation”. In: *arXiv preprint arXiv:2302.13007* (2023).
- [22] Paulo Ney De Souza and Jorge-Nuno Silva. *Berkeley Problems in Mathematics*. Springer, 1998.
- [23] Mostafa Dehghani et al. “Scaling vision transformers to 22 billion parameters”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 7480–7512.
- [24] Lizz Denneau. *Unveiling the Power of Persona*. en. <http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm>.
- [25] John Derby. “Disability studies and art education”. In: *Studies in Art Education* 52.2 (2011), pp. 94–111.
- [26] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [27] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2020.
- [28] Fei Dou et al. “Towards Artificial General Intelligence (AGI) in the Internet of Things (IoT): Opportunities and Challenges”. In: *arXiv preprint arXiv:2309.07438* (2023).
- [29] Saadia Gutta Essa, Turgay Celik, and Nadia Emelia Human-Hendricks. “Personalized adaptive learning technologies based on machine learning techniques to identify learning styles: A systematic literature review”. In: *IEEE Access* 11 (2023), pp. 48392–48409.
- [30] Ying Fan et al. “Reinforcement learning for fine-tuning text-to-image diffusion models”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [31] Steven Y. Feng et al. *A Survey of Data Augmentation Approaches for NLP*. 2021. arXiv: [2105.03075 \[cs.CL\]](https://arxiv.org/abs/2105.03075). URL: <https://arxiv.org/abs/2105.03075>.
- [32] Enrique Fernandez et al. *Learning ROS for robotics programming*. Packt Publishing Ltd, 2015.
- [33] Simon Frieder et al. “Mathematical capabilities of chatgpt”. In: *Advances in neural information processing systems* 36 (2024).
- [34] Huan Fu et al. “3d-front: 3d furnished rooms with layouts and semantics”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10933–10942.
- [35] Luyu Gao et al. “Pal: Program-aided language models”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 10764–10799.
- [36] Tianyu Gao, Adam Fisch, and Danqi Chen. “Making pre-trained language models better few-shot learners”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 2021, pp. 3816–3830.
- [37] Xinyu Gong et al. “Advancing microbial production through artificial intelligence-aided biology”. In: *Biotechnology Advances* (2024), p. 108399.
- [38] Zihan Guan et al. “Cohortgpt: An enhanced gpt for participant recruitment in clinical study”. In: *arXiv preprint arXiv:2307.11346* (2023).
- [39] Ying Guo and Daniel Lee. “Leveraging chatgpt for enhancing critical thinking skills”. In: *Journal of Chemical Education* 100.12 (2023), pp. 4876–4883.
- [40] Kelvin Guu et al. *REALM: Retrieval-Augmented Language Model Pre-Training*. 2020. arXiv: [2002.08909 \[cs.CL\]](https://arxiv.org/abs/2002.08909). URL: <https://arxiv.org/abs/2002.08909>.
- [41] Zeyu Han et al. “Parameter-efficient fine-tuning for large models: A comprehensive survey”. In: *arXiv preprint arXiv:2403.14608* (2024).
- [42] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. “Definition, structure, content, use and impacts of electronic health records: a review of the research literature”. In: *International journal of medical informatics* 77.5 (2008), pp. 291–304.

- 
- [43] Kaiming He et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009.
- [44] US Department of Health, Office of the Assistant Secretary for Planning Human Services, and Evaluation. *Compilation of State Data on the Affordable Care Act*. Accessed: September 19, 2024. 2024. URL: <https://aspe.hhs.gov/compilation-state-data-affordable-care-act>.
- [45] Dan Hendrycks et al. “Measuring massive multitask language understanding”. In: *arXiv preprint arXiv:2009.03300* (2020).
- [46] Dan Hendrycks et al. “Measuring Mathematical Problem Solving With the MATH Dataset”. In: *NeurIPS* (2021).
- [47] William Hersh and David Hickam. “Use of a multi-application computer workstation in a clinical setting.” In: *Bulletin of the Medical Library Association* 82.4 (1994), p. 382.
- [48] William Hersh et al. “OHSUMED: An interactive retrieval evaluation and new large test collection for research”. In: *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer. 1994, pp. 192–201.
- [49] Jason Holmes et al. “Benchmarking a Foundation Large Language Model on its Ability to Relabel Structure Names in Accordance With the American Association of Physicists in Medicine Task Group-263 Report”. In: *Practical Radiation Oncology* (2024).
- [50] Jason Holmes et al. “Benchmarking a foundation LLM on its ability to re-label structure names in accordance with the AAPM TG-263 report”. In: *arXiv preprint arXiv:2310.03874* (2023).
- [51] Jason Holmes et al. “Evaluating Large Language Models on a Highly-specialized Topic, Radiation Oncology Physics”. In: *arXiv preprint arXiv:2304.01938* (2023).
- [52] Wenpin Hou and Zhicheng Ji. “GeneTuring tests GPT models in genomics”. In: *BioRxiv* (2023).
- [53] Edward J Hu et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).
- [54] Xiaoyang Hu et al. “In-context analogical reasoning with pre-trained language models”. In: *arXiv preprint arXiv:2305.17626* (2023).
- [55] Shima Imani, Liang Du, and Harsh Shrivastava. *MathPrompter: Mathematical Reasoning using Large Language Models*. 2023. arXiv: 2303.05398 [cs.CL]. URL: <https://arxiv.org/abs/2303.05398>.
- [56] AQ Jiang et al. “Mistral 7B (2023)”. In: *arXiv preprint arXiv:2310.06825* (2023).
- [57] Qiao Jin et al. “Genegpt: Augmenting large language models with domain tools for improved access to biomedical information”. In: *Bioinformatics* 40.2 (2024), btae075.
- [58] Alistair EW Johnson et al. “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports”. In: *Scientific data* 6.1 (2019), p. 317.
- [59] Alistair EW Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3.1 (2016), pp. 1–9.
- [60] Sarah Johnson et al. “Ethical Considerations in Multimodal Large Language Models: A Framework for Bias Evaluation and Mitigation”. In: *AI Ethics Journal* 5.1 (2024), pp. 1–15.
- [61] Takeshi Kojima et al. “Large language models are zero-shot reasoners”. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.
- [62] Anis Koubaa et al. “Exploring ChatGPT Capabilities and Limitations: A Survey”. In: *IEEE Access* 11 (2023), pp. 118698–118721. DOI: [10.1109/ACCESS.2023.3326474](https://doi.org/10.1109/ACCESS.2023.3326474).
- [63] Clemens Scott Kruse et al. “The use of electronic health records to support population health: a systematic review of the literature”. In: *Journal of medical systems* 42.11 (2018), p. 214.

- 
- [64] Tom Kubiszyn and Gary D Borich. *Educational testing and measurement*. John Wiley & Sons, 2024.
- [65] Brenden M Lake et al. “Building machines that learn and think like people”. In: *Behavioral and Brain Sciences* 40 (2017).
- [66] Nathan Lambert. *Reverse engineering OpenAI’s o1 — interconnects.ai*. <https://www.interconnects.ai/p/reverse-engineering-openai-o1>. [Accessed 23-09-2024].
- [67] Serge Lang. *Real and Functional Analysis*. Springer, 1993.
- [68] Ehsan Latif, Xiaoming Zhai, and Lei Liu. “Ai gender bias, disparities, and fairness: Does training data matter?” In: *arXiv preprint arXiv:2312.10833* (2023).
- [69] Hung Le et al. “CodeChain: Towards Modular Code Generation Through Chain of Self-revisions with Representative Sub-modules”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=vYhglxSj8j>.
- [70] Gyeong-Geon Lee et al. “Multimodality of ai for education: Towards artificial general intelligence”. In: *arXiv preprint arXiv:2312.06037* (2023).
- [71] Kimin Lee et al. “Aligning text-to-image models using human feedback”. In: *arXiv preprint arXiv:2302.12192* (2023).
- [72] Brian Lester, Rami Al-Rfou, and Noah Constant. “The Power of Scale for Parameter-Efficient Prompt Tuning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 3045–3059.
- [73] Brian C Leventhal and Kathryn N Thompson. “Surveying the measurement profession to assist recruiting in the United States”. In: *Educational Measurement: Issues and Practice* 40.3 (2021), pp. 83–95.
- [74] Martha Lewis and Melanie Mitchell. “Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models”. In: *arXiv preprint arXiv:2402.08955* (2024).
- [75] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: [2005.11401 \[cs.CL\]](https://arxiv.org/abs/2005.11401). URL: <https://arxiv.org/abs/2005.11401>.
- [76] Lgw. *GitHub - lgw863/LogiQA-dataset*. en. URL: <https://github.com/lgw863/LogiQA-dataset>.
- [77] Feng Li et al. “Deep learning-based automated detection of retinal diseases using optical coherence tomography images”. In: *Biomedical optics express* 10.12 (2019), pp. 6204–6226.
- [78] Lingyao Li et al. “A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs)”. In: *arXiv preprint arXiv:2405.03066* (2024).
- [79] Pin-Hui Li et al. “Solving the Self-regulated Learning Problem: Exploring the Performance of ChatGPT in Mathematics”. In: *Innovative Technologies and Learning*. Ed. by Yueh-Min Huang and Tânia Rocha. Cham: Springer Nature Switzerland, 2023, pp. 77–86. ISBN: 978-3-031-40113-8.
- [80] Wenhui Li et al. “CogVLM: Visual Expert for Pretrained Language Models”. In: *arXiv preprint arXiv:2311.03079* (2023).
- [81] Xian Li and Percy Liang. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 2021, pp. 4582–4597.
- [82] Xiang Li et al. “Artificial General Intelligence for Medical Imaging”. In: *arXiv preprint arXiv:2306.05480* (2023).
- [83] Xiaoyuan Li et al. *Evaluating Mathematical Reasoning of Large Language Models: A Focus on Error Identification and Correction*. 2024. arXiv: [2406.00755 \[cs.CL\]](https://arxiv.org/abs/2406.00755). URL: <https://arxiv.org/abs/2406.00755>.
- [84] Xiujun Li et al. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 121–137.

- 
- [85] Zhening Li et al. *LEMMA: Bootstrapping High-Level Mathematical Reasoning with Learned Symbolic Abstractions*. 2022. arXiv: [2211.08671 \[cs.AI\]](https://arxiv.org/abs/2211.08671). URL: <https://arxiv.org/abs/2211.08671>.
- [86] Percy Liang et al. “Holistic Evaluation of Language Models”. In: *arXiv preprint arXiv:2211.09110* (2022).
- [87] Wenxiong Liao et al. “Differentiating ChatGPT-generated and human-written medical texts: quantitative study”. In: *JMIR Medical Education* 9.1 (2023), e48904.
- [88] Wenxiong Liao et al. “Mask-guided BERT for few-shot text classification”. In: *Neurocomputing* (2024), p. 128576.
- [89] Chengyuan Liu et al. “More Than Catastrophic Forgetting: Integrating General Capabilities For Domain-Specific LLMs”. In: *arXiv preprint arXiv:2405.17830* (2024).
- [90] Mingjie Liu et al. “Chipnemo: Domain-adapted llms for chip design”. In: *arXiv preprint arXiv:2311.00176* (2023).
- [91] Yiheng Liu et al. “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models”. In: *arXiv preprint arXiv:2304.01852* (2023).
- [92] Yiheng Liu et al. *Understanding LLMs: A Comprehensive Overview from Training to Inference*. 2024. arXiv: [2401.02038 \[cs.CL\]](https://arxiv.org/abs/2401.02038). URL: <https://arxiv.org/abs/2401.02038>.
- [93] Yiheng Liu et al. “Understanding llms: A comprehensive overview from training to inference”. In: *arXiv preprint arXiv:2401.02038* (2024).
- [94] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [95] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.
- [96] Zhengliang Liu et al. “Context Matters: A Strategy to Pre-train Language Model for Science Education”. In: *arXiv preprint arXiv:2301.12031* (2023).
- [97] Zhengliang Liu et al. “DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4”. In: *arXiv preprint arXiv:2303.11032* (2023).
- [98] Zhengliang Liu et al. “Evaluating large language models for radiology natural language processing”. In: *arXiv preprint arXiv:2307.13693* (2023).
- [99] Zhengliang Liu et al. “FINE-TUNING LARGE LANGUAGE MODELS FOR RADIATION ONCOLOGY, A HIGHLY SPECIALIZED HEALTHCARE DOMAIN”. In: *International Journal of Particle Therapy* 12 (2024), p. 100428.
- [100] Zhengliang Liu et al. *Holistic Evaluation of GPT-4V for Biomedical Imaging*. 2023. arXiv: [2312.05256 \[eess.IV\]](https://arxiv.org/abs/2312.05256). URL: <https://arxiv.org/abs/2312.05256>.
- [101] Zhengliang Liu et al. “Holistic evaluation of gpt-4v for biomedical imaging”. In: *arXiv preprint arXiv:2312.05256* (2023).
- [102] Zhengliang Liu et al. “LLM-POTUS Score: A Framework of Analyzing Presidential Debates with Large Language Models”. In: *arXiv preprint arXiv:2409.08147* (2024).
- [103] Zhengliang Liu et al. “Pharmacygpt: The ai pharmacist”. In: *arXiv preprint arXiv:2307.10432* (2023).
- [104] Zhengliang Liu et al. “Radiology-llama2: Best-in-class large language model for radiology”. In: *arXiv preprint arXiv:2309.06419* (2023).
- [105] Zhengliang Liu et al. “RadOnc-GPT: A Large Language Model for Radiation Oncology”. In: *arXiv preprint arXiv:2309.10160* (2023).
- [106] Zhengliang Liu et al. “Surviving ChatGPT in healthcare”. In: *Frontiers in Radiology* 3 (2024), p. 1224682.

- 
- [107] Zhengliang Liu et al. “The Radiation Oncology NLP Database”. In: *arXiv preprint arXiv:2401.10995* (2024).
- [108] Zhengliang Liu et al. “Transformation vs Tradition: Artificial General Intelligence (AGI) for Arts and Humanities”. In: *arXiv preprint arXiv:2310.19626* (2023).
- [109] Penfold Louisa. *Make a cardboard construction*. 2020. URL: <https://louisapenfold.com/cardboard-art/>.
- [110] Jiasen Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Advances in Neural Information Processing Systems*. 2019.
- [111] Yanjun Lyu et al. “GP-GPT: Large Language Model for Gene-Phenotype Mapping”. In: *arXiv preprint arXiv:2409.09825* (2024).
- [112] Chong Ma et al. “An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT”. In: *IEEE Transactions on Artificial Intelligence* (2024).
- [113] Jianzhu Ma et al. “Using deep learning to model the hierarchical structure and function of a cell”. In: *Nature methods* 15.4 (2018), pp. 290–298.
- [114] Andrew Maas et al. “Learning word vectors for sentiment analysis”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011, pp. 142–150.
- [115] Aman Madaan et al. “Self-Refine: Iterative Refinement with Self-Feedback”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: <https://openreview.net/forum?id=S37h0erQLB>.
- [116] Timothy R McIntosh et al. “Inadequacies of large language model benchmarks in the era of generative artificial intelligence”. In: *arXiv preprint arXiv:2402.09880* (2024).
- [117] Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. 2022. arXiv: [1908.09635 \[cs.LG\]](https://arxiv.org/abs/1908.09635). URL: <https://arxiv.org/abs/1908.09635>.
- [118] Metadocs. *How OpenAI o1 works in a simple way and why it matters for RAG and Agentic - Use AI the right way — metadocs.co*. <https://www.metadocs.co/2024/09/17/how-openai-o1-works-in-a-simple-way-and-why-it-matters-for-rag-and-agentic/>. [Accessed 23-09-2024].
- [119] Shubhra Mishra et al. *MathCAMPS: Fine-grained Synthesis of Mathematical Problems From Human Curricula*. 2024. arXiv: [2407.00900 \[cs.AI\]](https://arxiv.org/abs/2407.00900). URL: <https://arxiv.org/abs/2407.00900>.
- [120] Melanie Mitchell. “Abstraction and analogy-making in artificial intelligence”. In: *Annals of the New York Academy of Sciences* 1505.1 (2021), pp. 79–101.
- [121] Moran Mizrahi et al. “State of what art? a call for multi-prompt llm evaluation”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 933–949.
- [122] Saif Mohammad et al. “SemEval-2018 Task 1: Affect in Tweets”. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1–17. DOI: [10.18653/v1/S18-1001](https://doi.org/10.18653/v1/S18-1001). URL: <https://aclanthology.org/S18-1001>.
- [123] Sam Musker et al. *Semantic Structure-Mapping in LLM and Human Analogical Reasoning*. 2024. arXiv: [2406.13803 \[cs.CL\]](https://arxiv.org/abs/2406.13803). URL: <https://arxiv.org/abs/2406.13803>.
- [124] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization”. In: *ArXiv* abs/1808.08745 (2018).
- [125] Yixin Nie et al. “Adversarial NLI: A new benchmark for natural language understanding”. In: *arXiv preprint arXiv:1910.14599* (2019).
- [126] Tim Nuttle et al. *Cherokee Corpus*. <https://cherokeedictionary.net/corpus/corpusMain>. Accessed: 2024-09-17. 2024.

- 
- [127] OpenAI. “GPT-4 Technical Report”. In: *arXiv preprint arXiv:2303.08774* (2023).
- [128] OpenAI. *Introducing ChatGPT — openai.com*. <https://openai.com/blog/chatgpt>. [Accessed 28-08-2023].
- [129] OpenAI. *Learning to Reason with LLMs*. <https://openai.com/index/learning-to-reason-with-lmms/>. [Accessed 19-09-2024]. 2024.
- [130] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.
- [131] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. “Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering”. In: *Conference on health, inference, and learning*. PMLR. 2022, pp. 248–260.
- [132] Ronald Carl Petersen et al. “Alzheimer’s disease neuroimaging initiative (ADNI): clinical characterization”. In: *Neurology* 74 (2010).
- [133] Pikachu345. *[Math Talk] Is the latest AI model’s capability approaching that of a PhD in Mathematics?* <https://www.bilibili.com/video/BV1T644e1EMN/>. [Online; accessed 19-Sep-2024]. Bilibili, 2024.
- [134] William F Pinar. “Currere”. In: *Key concepts in curriculum studies*. Routledge, 2019, pp. 50–52.
- [135] Silviu Pitis et al. “Boosted Prompt Ensembles for Large Language Models”. In: *arXiv e-prints* (2023), arXiv–2304.
- [136] Vagelis Plevris, George Papazafeiropoulos, and Alejandro Jiménez Rios. “Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard”. In: *AI* 4.4 (2023), pp. 949–969. ISSN: 2673-2688. DOI: [10.3390/ai4040048](https://doi.org/10.3390/ai4040048). URL: <https://www.mdpi.com/2673-2688/4/4/48>.
- [137] Gabriel Poesia and Noah D. Goodman. “Peano: learning formal mathematical reasoning”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381.2251 (2023), p. 20220044. DOI: [10.1098/rsta.2022.0044](https://doi.org/10.1098/rsta.2022.0044). eprint: <https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2022.0044>. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2022.0044>.
- [138] Gabriel Poesia et al. *Learning Formal Mathematics From Intrinsic Motivation*. 2024. arXiv: [2407.00695 \[cs.AI\]](https://arxiv.org/abs/2407.00695). URL: <https://arxiv.org/abs/2407.00695>.
- [139] Maria Pontiki et al. “SemEval-2014 Task 4: Aspect Based Sentiment Analysis”. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Aug. 2014, pp. 27–35.
- [140] Maria Pontiki et al. “SemEval-2015 Task 12: Aspect Based Sentiment Analysis”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, June 2015, pp. 486–495.
- [141] Indrasen Poola. *Tuning ChatGPT Mathematical Reasoning Limitations and Failures with Process Supervision*. Zenodo. June 2023. DOI: [10.5281/zenodo.8296440](https://doi.org/10.5281/zenodo.8296440). URL: <https://doi.org/10.5281/zenodo.8296440>.
- [142] Guanghui Qin and Jason Eisner. “Learning How to Ask: Querying LMs with Mixtures of Soft Prompts”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. 2021, pp. 5203–5212.
- [143] Alec Radford et al. “Improving language understanding by generative pre-training”. In: *OpenAI* (2018).
- [144] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [145] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

- 
- [146] Jack W Rae et al. “Scaling language models: Methods, analysis & insights from training gopher”. In: *arXiv preprint arXiv:2112.11446* (2021).
- [147] Rafael Rafailov et al. “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [148] Steve Rathje et al. “GPT is an effective tool for multilingual psychological text analysis”. In: *Proceedings of the National Academy of Sciences* 121.34 (2024), e2308950121.
- [149] Saed Rezayi et al. “AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition”. In: *International Joint Conference on Artificial Intelligence*. 2022.
- [150] Saed Rezayi et al. “ClinicalRadioBERT: Knowledge-Infused Few Shot Learning for Clinical Notes Named Entity Recognition”. In: *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. Springer. 2022, pp. 269–278.
- [151] Saed Rezayi et al. “Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications”. In: *IEEE Transactions on Big Data* (2024).
- [152] Alexey Romanov and Chaitanya Shivade. “Lessons from natural language inference in the clinical domain”. In: *arXiv preprint arXiv:1808.06752* (2018).
- [153] Sara Rosenthal, Noura Farra, and Preslav Nakov. “SemEval-2017 Task 4: Sentiment Analysis in Twitter”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518. DOI: [10.18653/v1/S17-2088](https://doi.org/10.18653/v1/S17-2088). URL: <https://aclanthology.org/S17-2088>.
- [154] Sheldon M Ross. *Stochastic processes*. John Wiley & Sons, 1995.
- [155] Zhao Ruixue et al. “Insights and reflections of the impact of ChatGPT on intelligent knowledge services in libraries”. In: *Journal of Library and Information Sciences in Agriculture* 35.1 (2023), p. 29.
- [156] Luis M. Sánchez-Ruiz et al. “ChatGPT Challenges Blended Learning Methodologies in Engineering Education: A Case Study in Mathematics”. In: *Applied Sciences* 13.10 (2023). ISSN: 2076-3417. DOI: [10.3390/app13106039](https://doi.org/10.3390/app13106039). URL: <https://www.mdpi.com/2076-3417/13/10/6039>.
- [157] Gargi Sathe et al. “Cross-Language Assessment of Mathematical Capability of ChatGPT”. In: *Intelligent Systems Design and Applications*. Ed. by Ajith Abraham et al. Cham: Springer Nature Switzerland, 2024, pp. 468–479. ISBN: 978-3-031-64779-6.
- [158] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).
- [159] Emily Sheng et al. “Revealing the Biases in Large Language Models”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 932–947.
- [160] Olena Shevtsova et al. “The role of art education in personality development”. In: *Amazonia Investiga* 13.74 (2024), pp. 143–155.
- [161] Yucheng Shi et al. “MedEdit: Model Editing for Medical Question Answering with External Knowledge Bases”. In: *arXiv preprint arXiv:2309.16035* (2023).
- [162] Peng Shu et al. “LLMs for Coding and Robotics Education”. In: *arXiv preprint arXiv:2402.06116* (2024).
- [163] Kashun Shum, Shizhe Diao, and Tong Zhang. “Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12113–12139. DOI: [10.18653/v1/2023.findings-emnlp.811](https://doi.org/10.18653/v1/2023.findings-emnlp.811). URL: <https://aclanthology.org/2023.findings-emnlp.811>.
- [164] Shelley Staples and Bradley Dilger. *Corpus and repository of writing*. <https://crow.corporaproject.org>. Learner corpus articulated with repository. 2018.

- 
- [165] Nisan Stiennon et al. “Learning to summarize with human feedback”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.
  - [166] Amber Stubbs and Özlem Uzuner. “Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus”. In: *Journal of biomedical informatics* 58 (2015), S20–S29.
  - [167] Lichao Sun et al. “Trustllm: Trustworthiness in large language models”. In: *arXiv preprint arXiv:2401.05561* (2024).
  - [168] Colleen M. Suter et al. “The Affordable Care Act at 10 Years: Evaluating the Evidence and Navigating an Uncertain Future”. In: *Journal of Surgical Research* 258 (2021), pp. 207–220. URL: [https://www.journalofsurgicalresearch.com/article/S0022-4804\(21\)00016-0/fulltext](https://www.journalofsurgicalresearch.com/article/S0022-4804(21)00016-0/fulltext).
  - [169] Chenjiao Tan et al. “On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications”. In: *arXiv preprint arXiv:2312.17016* (2023).
  - [170] Songbo Tan. “Large margin DragPushing strategy for centroid text categorization”. In: *Expert Systems with Applications* 33.1 (2007), pp. 215–220.
  - [171] Chenhao Tang et al. “PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models”. In: *arXiv preprint arXiv:2309.10238* (2023).
  - [172] Jie Tian et al. “Assessing Large Language Models in Mechanical Engineering Education: A Study on Mechanics-Focused Conceptual Understanding”. In: *arXiv preprint arXiv:2401.12983* (2024).
  - [173] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
  - [174] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.
  - [175] Cynthia Van Hee, Els Lefever, and Véronique Hoste. “Semeval-2018 task 3: Irony detection in english tweets”. In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pp. 39–50.
  - [176] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
  - [177] Oriol Vinyals et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *nature* 575.7782 (2019), pp. 350–354.
  - [178] Tu Vu, Harsh Jhamtani, and Eduard Hovy. “SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer”. In: *arXiv preprint arXiv:2110.07904* (2021).
  - [179] Alex Wang et al. “GLUE: A multi-task benchmark and analysis platform for natural language understanding”. In: *arXiv preprint arXiv:1804.07461* (2018).
  - [180] Hui Wang et al. “Seeing ChatGPT Through Universities’ Policies, Resources and Guidelines”. In: *arXiv preprint arXiv:2312.05235* (2023).
  - [181] Jiaqi Wang et al. “A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks”. In: *arXiv preprint arXiv:2408.01319* (2024).
  - [182] Jiaqi Wang et al. “Large language models for robotics: Opportunities, challenges, and perspectives”. In: *arXiv preprint arXiv:2401.04334* (2024).
  - [183] Jiaqi Wang et al. “Prompt engineering for healthcare: Methodologies and applications”. In: *arXiv preprint arXiv:2304.14670* (2023).
  - [184] Lei Wang et al. “A survey on large language model based autonomous agents”. In: *Frontiers of Computer Science* 18.6 (2024), p. 186345.
  - [185] Tingwu Wang et al. “Benchmarking model-based reinforcement learning”. In: *arXiv preprint arXiv:1907.02057* (2019).

- 
- [186] Xuezhi Wang et al. “Self-consistency improves chain of thought reasoning in language models”. In: *arXiv preprint arXiv:2203.11171* (2022).
- [187] Yizhong Wang et al. “Efficient Multimodal Large Language Models with Pruning-Based Compression”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 10112–10125.
- [188] Yousef Wardat et al. “ChatGPT: A revolutionary tool for teaching and learning mathematics”. In: *Eurasia Journal of Mathematics, Science and Technology Education* 19.7 (2023). Publisher: Modestum Ltd. ISSN: 1305-8215, 1305-8223. DOI: [10.29333/ejmste/13272](https://doi.org/10.29333/ejmste/13272). URL: <https://www.ejmste.com/article/chatgpt-a-revolutionary-tool-for-teaching-and-learning-mathematics-13272> (visited on 09/17/2024).
- [189] Taylor Webb, Keith J Holyoak, and Hongjing Lu. “Emergent analogical reasoning in large language models”. In: *Nature Human Behaviour* 7.9 (2023), pp. 1526–1541.
- [190] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: [2201.11903 \[cs.CL\]](https://arxiv.org/abs/2201.11903). URL: <https://arxiv.org/abs/2201.11903>.
- [191] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.
- [192] Yaonai Wei et al. “Chat2Brain: A Method for Mapping Open-Ended Semantic Queries to Brain Activation Maps”. In: *arXiv preprint arXiv:2309.05021* (2023).
- [193] Johannes Welbl, Nelson F. Liu, and Matt Gardner. “Crowdsourcing multiple choice science questions”. In: *arXiv preprint arXiv:1707.06209* (2017). URL: <https://arxiv.org/abs/1707.06209>.
- [194] Adina Williams, Nikita Nangia, and Samuel R Bowman. “A broad-coverage challenge corpus for sentence understanding through inference”. In: *arXiv preprint arXiv:1704.05426* (2017).
- [195] Zihao Wu et al. “Exploring the Trade-Offs: Unified Large Language Models vs Local Fine-Tuned Models for Highly-Specific Radiology NLI Task”. In: *arXiv preprint arXiv:2304.09138* (2023).
- [196] Shaochen Xu et al. “Reasoning before comparison: LLM-enhanced semantic similarity metrics for domain specialized text analysis”. In: *arXiv preprint arXiv:2402.11398* (2024).
- [197] Yixuan Yang et al. *LLplace: The 3D Indoor Scene Layout Generation and Editing via Large Language Model*. 2024. arXiv: [2406.03866 \[cs.CV\]](https://arxiv.org/abs/2406.03866). URL: <https://arxiv.org/abs/2406.03866>.
- [198] Zhenyu Yang et al. “Multimodal Reasoning Benchmark: Integrating Vision and Language for Complex Problem Solving”. In: *arXiv preprint arXiv:2402.12425* (2024).
- [199] Shunyu Yao et al. “ReAct: Synergizing reasoning and acting in language models”. In: *arXiv preprint arXiv:2210.03629* (2022).
- [200] Shunyu Yao et al. “Tree of thoughts: Deliberate problem solving with large language models”. In: *arXiv preprint arXiv:2305.10601* (2023).
- [201] Michael Ku Yu et al. “Translation of genotype to phenotype by a hierarchy of cell subsystems”. In: *Cell systems* 2.2 (2016), pp. 77–88.
- [202] Joy He-Yueya et al. *Solving Math Word Problems by Combining Language Models With Symbolic Solvers*. 2023. arXiv: [2304.09102 \[cs.CL\]](https://arxiv.org/abs/2304.09102). URL: <https://arxiv.org/abs/2304.09102>.
- [203] yymYM. *Stock Trading QA: A Question Answering Dataset for Finance and Trading*. [https://huggingface.co/datasets/yymYM/stock\\_trading\\_QA](https://huggingface.co/datasets/yymYM/stock_trading_QA). Accessed: 2024-09-17. 2023.
- [204] Marcos Zampieri et al. “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Ed. by Jonathan May et al. Minneapolis, Minnesota, USA: Association for

- 
- Computational Linguistics, June 2019, pp. 75–86. DOI: [10.18653/v1/S19-2010](https://doi.org/10.18653/v1/S19-2010). URL: <https://aclanthology.org/S19-2010>.
- [205] Eric Zelikman et al. “Quiet-star: Language models can teach themselves to think before speaking”. In: *arXiv preprint arXiv:2403.09629* (2024).
  - [206] Kai Zhang et al. “A generalist vision–language foundation model for diverse biomedical tasks”. In: *Nature Medicine* (2024), pp. 1–13.
  - [207] Kai Zhang et al. “BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks”. In: *arXiv preprint arXiv:2305.17100* (2023).
  - [208] Qiang Zhang et al. “mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality”. In: *arXiv preprint arXiv:2304.14178* (2023).
  - [209] Shengyu Zhang et al. “Instruction tuning for large language models: A survey”. In: *arXiv preprint arXiv:2308.10792* (2023).
  - [210] Tiezheng Zhang et al. “Leveraging AI Predicted and Expert Revised Annotations in Interactive Segmentation: Continual Tuning or Full Training?” In: *arXiv preprint arXiv:2402.19423* (2024).
  - [211] Yutong Zhang et al. *Potential of Multimodal Large Language Models for Data Mining of Medical Images and Free-text Reports*. 2024. arXiv: [2407.05758 \[eess.IV\]](https://arxiv.org/abs/2407.05758). URL: <https://arxiv.org/abs/2407.05758>.
  - [212] Zhuosheng Zhang et al. “Automatic Chain of Thought Prompting in Large Language Models”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=5NTt8GFjUHkr>.
  - [213] Huan Zhao et al. “Ophtha-llama2: A large language model for ophthalmology”. In: *arXiv preprint arXiv:2312.04906* (2023).
  - [214] Huaqin Zhao et al. “Revolutionizing finance with llms: An overview of applications and insights”. In: *arXiv preprint arXiv:2401.11641* (2024).
  - [215] Lin Zhao et al. “When brain-inspired ai meets agi”. In: *arXiv preprint arXiv:2303.15935* (2023).
  - [216] Tianyang Zhong et al. *ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-institution and Multi-system Data*. 2023. arXiv: [2310.05242 \[cs.CL\]](https://arxiv.org/abs/2310.05242). URL: <https://arxiv.org/abs/2310.05242>.
  - [217] Ce Zhou et al. “A comprehensive survey on pretrained foundation models: A history from bert to chatgpt”. In: *arXiv preprint arXiv:2302.09419* (2023).
  - [218] Daniel M Ziegler et al. “Fine-tuning language models from human preferences”. In: *arXiv preprint arXiv:1909.08593* (2019).