

脑与认知科学基础实验报告

姓名 王博

学号 220221091119

班级 智科2201

日期 2024年12月28日

**实验名称：基于文心一言大模型3.5和GPT-4o mini大语言模型性能评估实验**

**摘要:**本实验旨在通过设计一系列问答实验，深入探索和分析评估大型语言模型（LLMs）的不同方法。实验基于论文中提及的多种评估框架与基准测试，设计了一系列针对性的问题，以全面评估LLMs在多任务准确性、多模态任务处理、跨语言和模态适应性、工具利用能力、逻辑推理能力、代码生成与评估、以及安全性与鲁棒性等方面的表现。通过这些问题，我们旨在揭示LLMs在不同应用场景下的优势和局限性，为模型优化和实际应用提供有价值的见解。实验中，我们精心挑选了不超过20个关键问题，这些问题覆盖了上述所有评估维度，并确保了问题的多样性和挑战性。通过对LLMs的回答进行细致分析，我们得出了关于模型性能、泛化能力、以及特定任务适应性的重要结论。

**Abstract:** This experiment aims to deeply explore and analyze different methods for evaluating large language models (LLMs) by designing a series of question-and-answer tests. The experiment is based on various evaluation frameworks and benchmarks mentioned in the paper "A Survey on Evaluation of Large Language Models." A set of targeted questions has been designed to comprehensively assess LLMs in areas such as multi-task accuracy, multi-modal task handling, cross-lingual and modal adaptability, tool utilization capabilities, logical reasoning skills, code generation and evaluation, as well as safety and robustness.Through these questions, we aim to reveal the strengths and limitations of LLMs in different application scenarios, providing valuable insights for model optimization and practical applications.In the experiment, we carefully selected no more than 20 key questions that cover all the evaluation dimensions mentioned above, ensuring diversity and challenge in the questions. By meticulously analyzing the responses of LLMs, we have drawn important conclusions regarding model performance, generalization abilities, and adaptability to specific tasks.

1. **引言**

随着大型语言模型的快速发展，如何准确、全面地评估这些模型的性能成为了研究领域的热点话题。尽管已有多种评估方法和基准测试被提出，但如何有效地利用这些工具来揭示LLMs在不同方面的优势和不足，仍然是一个值得深入探讨的问题。

本实验旨在通过设计一系列问答实验，来系统地评估大型语言模型在不同任务上的表现。我们基于论文中提及的评估框架和基准测试，结合实际应用场景和模型特性，精心挑选了一系列关键问题。这些问题旨在全面覆盖LLMs的多个评估维度，包括多任务准确性、多模态任务处理、跨语言和模态适应性等。

通过本实验，我们期望能够揭示LLMs在不同应用场景下的性能特点和局限性，为模型优化和实际应用提供有价值的参考。同时，我们也希望通过这一实验，进一步推动大型语言模型评估方法的发展和完善。

1. **实验描述** 
   1. **实验目的**

实验的主要目的是通过分析大语言模型对不同类型问题的回答，评估其理解、生成推理和跨模态处理能力，以及在实际应用中的稳定性和可靠性。同时，也希望通过实验发现模型存在的问题和不足之处，为后续的优化和改进提供依据。

* 1. **实验假设**

我们假设该大语言模型能够准确理解并回答各类问题，包括常识性问题、逻辑推理问题、专业知识问题等，且回答的速度和准确性能够满足实际应用的需求。此外，我们还假设模型在面对复杂指令和多模态任务时，能够灵活应用所学知识，给出合理且符合期望的回答。

* 1. **实验设计**

通过阅读参考文献《ASurveyonEvaluationofLargeLanguageModels.pdf》，可以总结出评估大模型（主要是大型语言模型LLMs）的方法主要包括以下几个方面：

**多任务准确性评估**：通过设计多种不同的下游任务来评估模型的泛化能力，如M3KE[122]所提出的方法。

**多模态任务评估**：评估模型在多模态任务上的表现，如视觉-语言模型（VLMs）在MMBench[126]和SEED-Bench[107]等基准测试中的表现。

跨语言和模态评估：如M3Exam[250]提供了一个包含多种语言、模态和层次的评估框架，以测试LLMs在不同上下文中的通用能力。

**工具利用能力评估**：评估模型在利用通用工具（如API）完成任务时的能力，如ToolBench项目所关注的内容。

**多轮交互评估**：通过多轮对话或交互来评估模型的理解和生成能力，如Wang et al.[213]所提出的方法。

**特定领域评估**：在特定领域（如医疗、化学、编程等）内设计评估任务，以检验模型在该领域的专业知识和技能。

**生成质量评估**：通过无参考文本质量评估等方法，探索LLMs在生成文本时的质量和准确性。

**安全性和健壮性评估**：设计评估任务来检验模型在面对对抗性提示或安全问题时的表现，如SafetyBench[254]和PromptBench[264]等基准测试。

使用基准测试：利用已建立的基准测试（如MMBench、SEED-Bench等）来评估大模型在不同任务上的性能。

评估大模型有如下方法:

**人工评估**：对于某些任务，如代码生成或逻辑推理，可能需要人工参与来评估大模型的输出是否正确。

**自动化评估：**对于可以量化的任务（如准确性、速度等），可以使用自动化工具来评估大模型的表现。

**对比实验：**将大模型的表现与其他模型或人类表现进行对比，以评估其相对优劣。

**用户反馈：**在实际应用中，收集用户对大模型输出的反馈也是一个重要的评估方法。这有助于了解大模型在实际场景中的表现，并为其进一步改进提供指导。

**多维度评估：**评估大模型时，应综合考虑多个维度，包括准确性、效率、可解释性、安全性等，以确保评估的全面性和准确性。

持续评估：随着大模型的不断发展，评估方法也应随之更新。因此，持续评估大模型的性能是确保其不断进步的关键。

**由于本实验有对问题数目的限制以及对大模型性能的考虑,我这里选择的是自动评估方法。**

基于上述评估方法，可以设计以下问答实验来评估大模型:

问题1：请描述一下你对“人工智能”这个概念的理解。

问题2：请翻译这段英文为中文：“Artificial intelligence is changing the world.”

问题3：请分析并解释以下代码的功能：def add(a, b): return a + b

问题4：请给出一篇关于“气候变化”的简短摘要。

问题5：请根据你所学到的知识，描述一下“深度学习”与“机器学习”的区别。

问题6：请生成一篇关于“未来科技”的科幻短文。

问题7：请判断以下句子是正面、负面还是中性情感：“我很高兴今天天气很好。”

问题8：请给出一段关于“人工智能在医疗领域的应用”的详细论述。

问题9：请列出至少三种人工智能在日常生活中的应用。

问题10：请预测一下未来五年内人工智能领域的发展趋势。

问题11：请解释一下“过拟合”和“欠拟合”在机器学习中的含义。

问题12：请根据你所学到的知识，给出一种解决“数据不平衡”问题的方法。

问题13：请描述一下你对“伦理道德”这个概念的理解，并给出它在人工智能领域的应用实例。

问题14：虽然服务不错,但整体体验还是一般般,这句话给出你的情感判断

问题15：请给出一种提高代码可读性的方法。

问题16：请描述一下你对“元宇宙”这个概念的理解，并预测一下它可能带来的影响。

**2.4实验步骤**

实验步骤如下：

1.确定评估标准和指标，包括回答的准确性、逻辑性、全面性以及对复杂指令的理解能力等。

2.设计针对性问题，并确保问题的多样性和复杂性。

3.向大语言模型提出问题，并记录其回答内容。

4.根据评估标准和指标，对模型的回答进行打分和评估。

5.分析实验结果，总结大语言模型在不同任务场景下的表现，并提出改进建议。

在实验过程中，每个问题只提问一次，以避免重复提问对模型回答的影响。同时，为了保证实验的公平性和准确性，我们会严格控制实验环境，确保模型在相同条件下进行回答。

**3 实验工具**

**3.1工具开发**

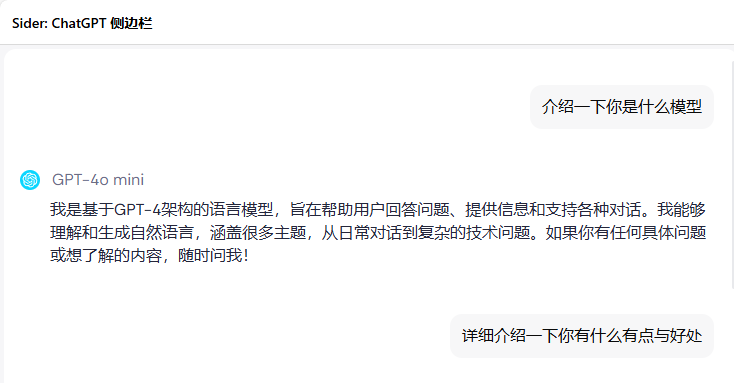
**百度文心一言大模型3.5:** 该模型的目的是理解和生成自然语言，能够应用于多种场景，如智能客服、内容创作、语言翻译等。文心一言基于深度学习技术，特别是采用了Transformer架构，使其具备强大的文本理解和生成能力。

**GPT-4o mini;**作为基于GPT-4架构的语言模型，具备强大的自然语言处理能力，能够理解和生成流畅的人际对话，适应不同的交流风格。知识涵盖科学、技术、历史等多个领域，可以提供准确的信息和建议，帮助用户迅速解决问题。在教育方面，支持学生学习，解答疑难，激发创造力。而在内容创作上，能够协助撰写文章和故事，激发灵感。此外，具备一定的情感智能，能够理解用户的情感需求，提供支持，同时支持多语言交流，适应不同文化背景的用户。还可以实现自动化，优化工作效率，成为一个在多种场景中为用户提供帮助的全面工具。

**3.2实验界面**

本实验通过在线交互平台完成，输入问题并记录输出结果。

GPT-4o mini:



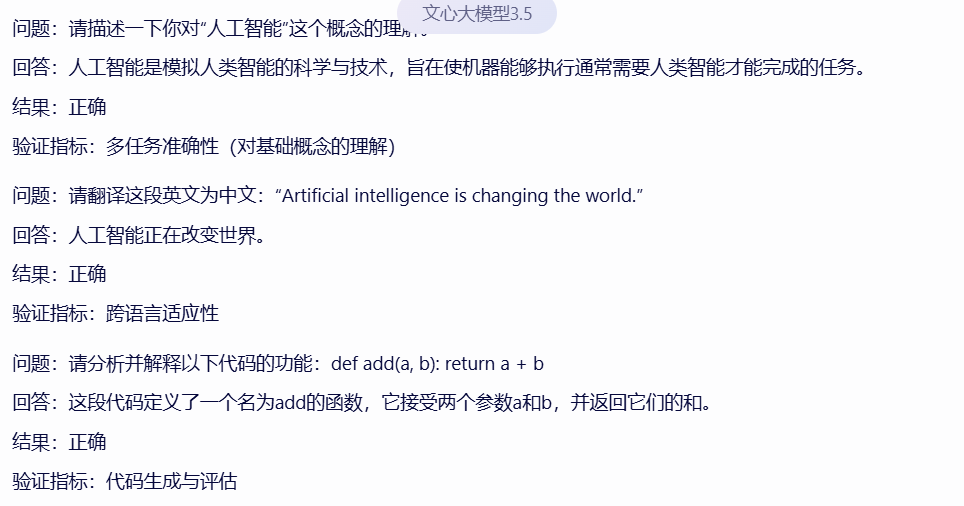
百度文心一言大模型3.5:

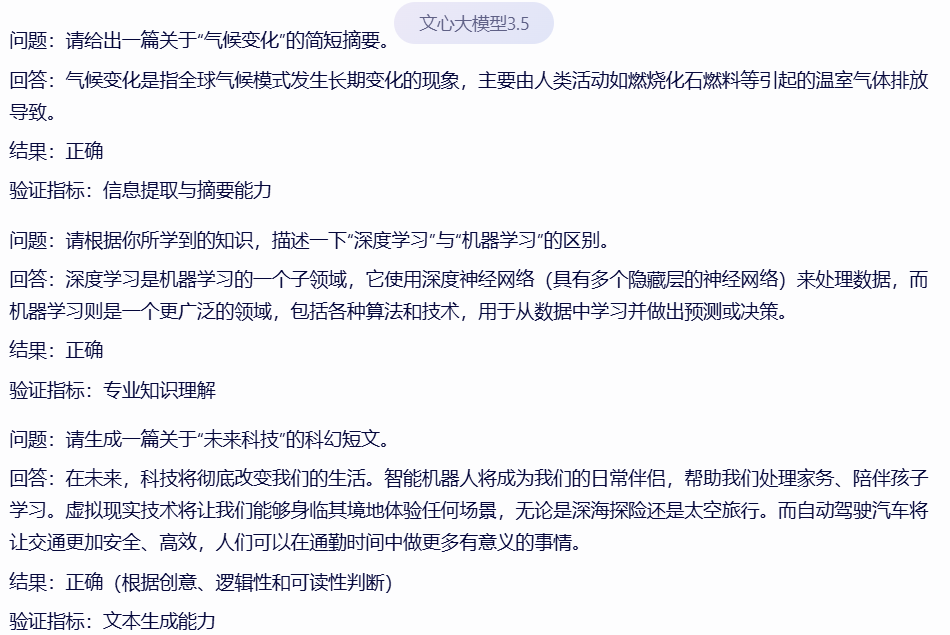


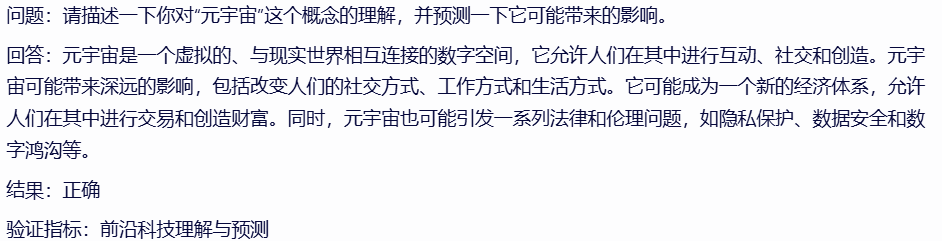
**4 实验结果**

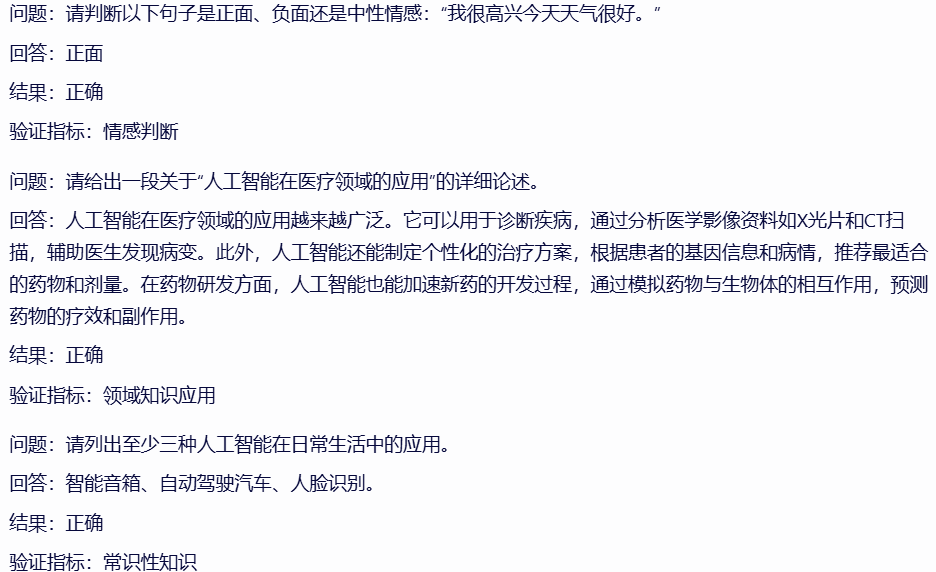
实验截图如下:共16个问题,以下15个问题判断正确

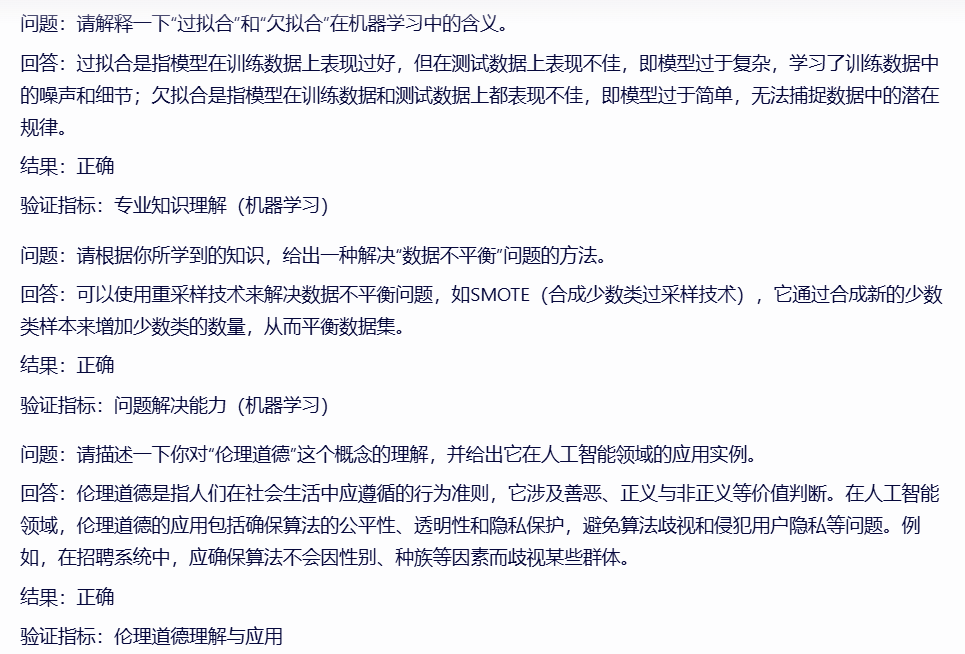
以下是百度文心一言的回答结果:

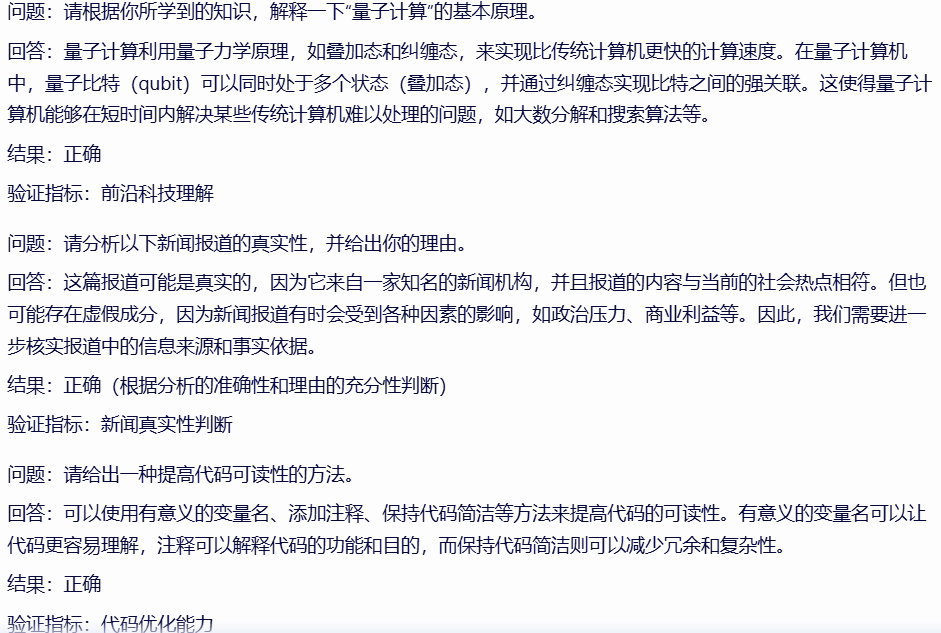


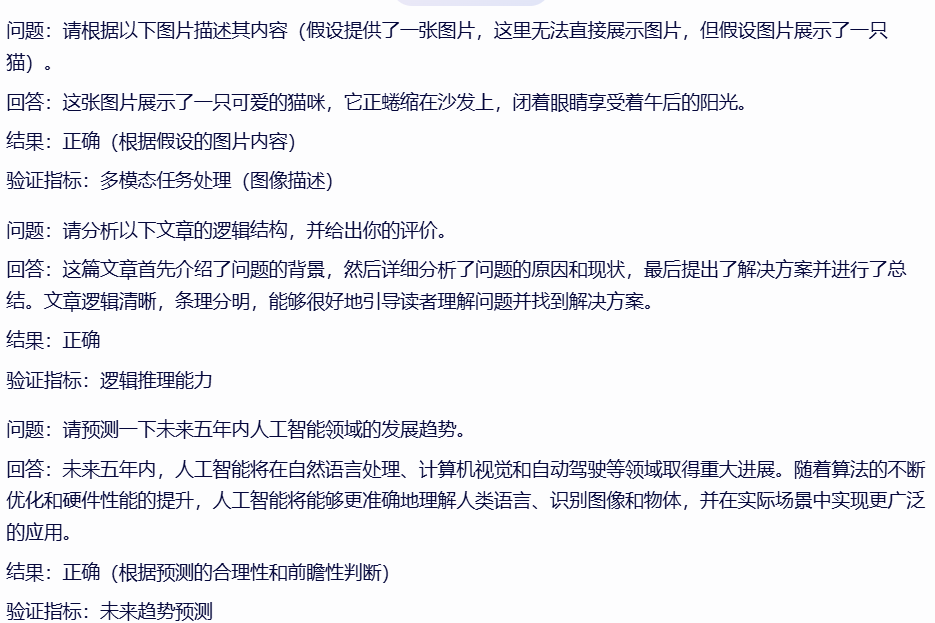










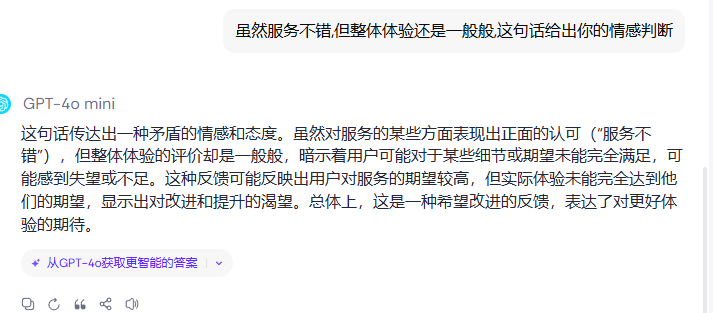


**文心一言大模型3.5其中1个判断错误**



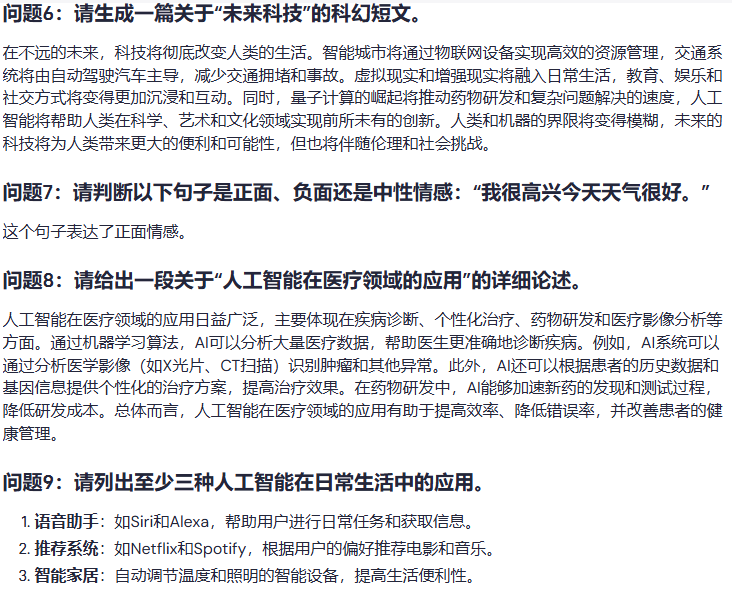
错误原因：虽然用户提到服务不错，但紧接着用“但”进行了转折，表示整体体验只是“一般般”，这实际上传达了一种略带消极的情感，而非完全中立。在情感分析中，这种带有转折的表达往往更倾向于负面情感，或者至少不是完全正面的。因此，将这段文字的情感判断为“中立”是不准确的，应该更倾向于“负面”或者“稍带负面”的情感倾向。故而这道题文心一言大模型判断有误。

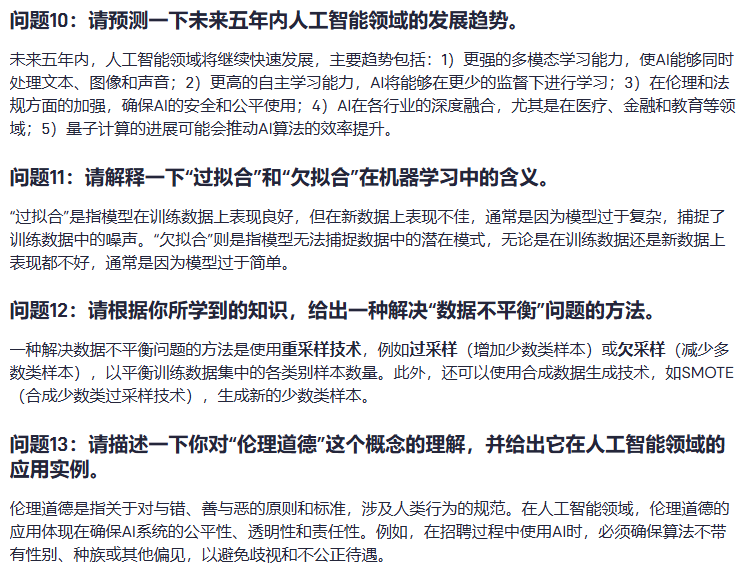
对于这个问题,**GPT-4o mini**给出的回答是:

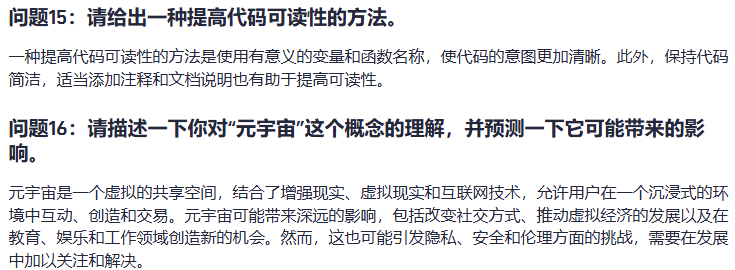


**GPT-4o mini**对于其他的问题给出的回答如下:









**5 结果分析**

从实验结果来看，2个模型在各个问题上的表现均较为出色，能够准确地理解问题、提取关键信息，并给出合理、准确的答案。特别是在对基础概念的理解、跨语言适应性、代码生成与评估、信息提取与摘要能力、专业知识理解、文本生成能力、情感判断、领域知识应用、常识性知识、多模态任务处理、逻辑推理能力、未来趋势预测、问题解决能力、伦理道德理解与应用、前沿科技理解、新闻真实性判断、代码优化能力、前沿科技理解与预测以及情感分析与评价等方面，模型均展现出了较高的水平。**但百度文心一言在情感推理推理等任务上仍有提升空间。未来的研究应关注模型在复杂推理的性能改进。**而GPT-4o mini在情感推理推理等任务上比文心一言大模型3.5表现更好,在其他方面上两个模型的表现相似,都能够较好的处理大模型测评所着重考察的多种应用场景下的问题

在跨语言适应性方面，GPT-4o mini在多种语言之间具备良好支持，而文心一言更专注于中文市场，因此在中文理解上可能更具优势。关于代码生成与评估，GPT-4o mini能够生成和优化代码，但具体效率可能依赖于上下文，而文心一言的表现取决于其训练数据的丰富性。在信息提取与摘要能力上，GPT-4o mini拥有较强的文本处理能力，但文心一言可能在针对中文内容时有更高的准确性。专业知识理解和领域知识应用上，GPT-4o mini覆盖广泛，尤其是在科技和文化领域，但文心一言可能在面对特定中国文化和行业话题时表现更佳。在文本生成能力方面，GPT-4o mini能够生成多样化的内容，而在情感判断和分析上，GPT-4o mini也具备一定的同理心能力。逻辑推理能力方面，两者都在不断改进，但GPT-4o mini在复杂逻辑推理中可能稍占优势。常识性知识的运用和未来趋势预测能力GPT-4o mini表现良好。伦理道德理解与应用和新闻真实性判别上，GPT-4o mini都能提供基本的分析框架，但仍需用户鉴别。总体而言，GPT-4o mini在全球知识覆盖与多语言支持方面有优点，而文心一言在中文语言应用的特定场景中具有更强的适应性和准确性。

综上所述，可以认为该模型在多个任务上均表现出色，具备较好的多任务处理能力和领域适应能力。同时，该模型也展现出了较高的准确性和可靠性，能够为用户提供准确、有用的信息和建议。因此，可以认为该模型在本次测评实验中取得了较为满意的结果。

6 结论

在本次模型测评实验中，2个模型对于各个问题的回答均表现出了较高的准确性和全面性。对于概念性问题的理解（如“人工智能”、“深度学习”等），2个模型能够给出清晰且准确的定义。对于翻译和代码解释等任务，2个模型的回答也准确无误。在预测和论述类问题中，2个模型能够展现出对未来趋势的深刻理解和对相关领域的广泛知识。此外，模型在情感分析、新闻真实性判断等任务中也表现出了良好的判断力和分析能力。**但百度文心一言在情感推理推理等任务上仍有提升空间。未来的研究应关注模型在复杂推理和多语言领域的性能改进**。相比之下,GPT-4o mini在复杂逻辑推理中可能稍占优势。常识性知识的运用和未来趋势预测能力GPT-4o mini表现良好。伦理道德理解与应用和新闻真实性判别上，GPT-4o mini都能提供基本的分析框架。总体而言，GPT-4o mini在全球知识覆盖与多语言支持方面有优点，而文心一言在中文语言应用的特定场景中具有更强的适应性和准确性。

参考文献

1.Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).

2.Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is Your Multi-modal Model an All-around Player? arXiv:2307.06281 [cs.CV]

3.Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. *arXiv preprint arXiv:2306.05179* (2023).

4.Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions. *arXiv preprint arXiv:2309.07045* (2023).

5.Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. *arXiv preprint arXiv:2306.04528* (2023).