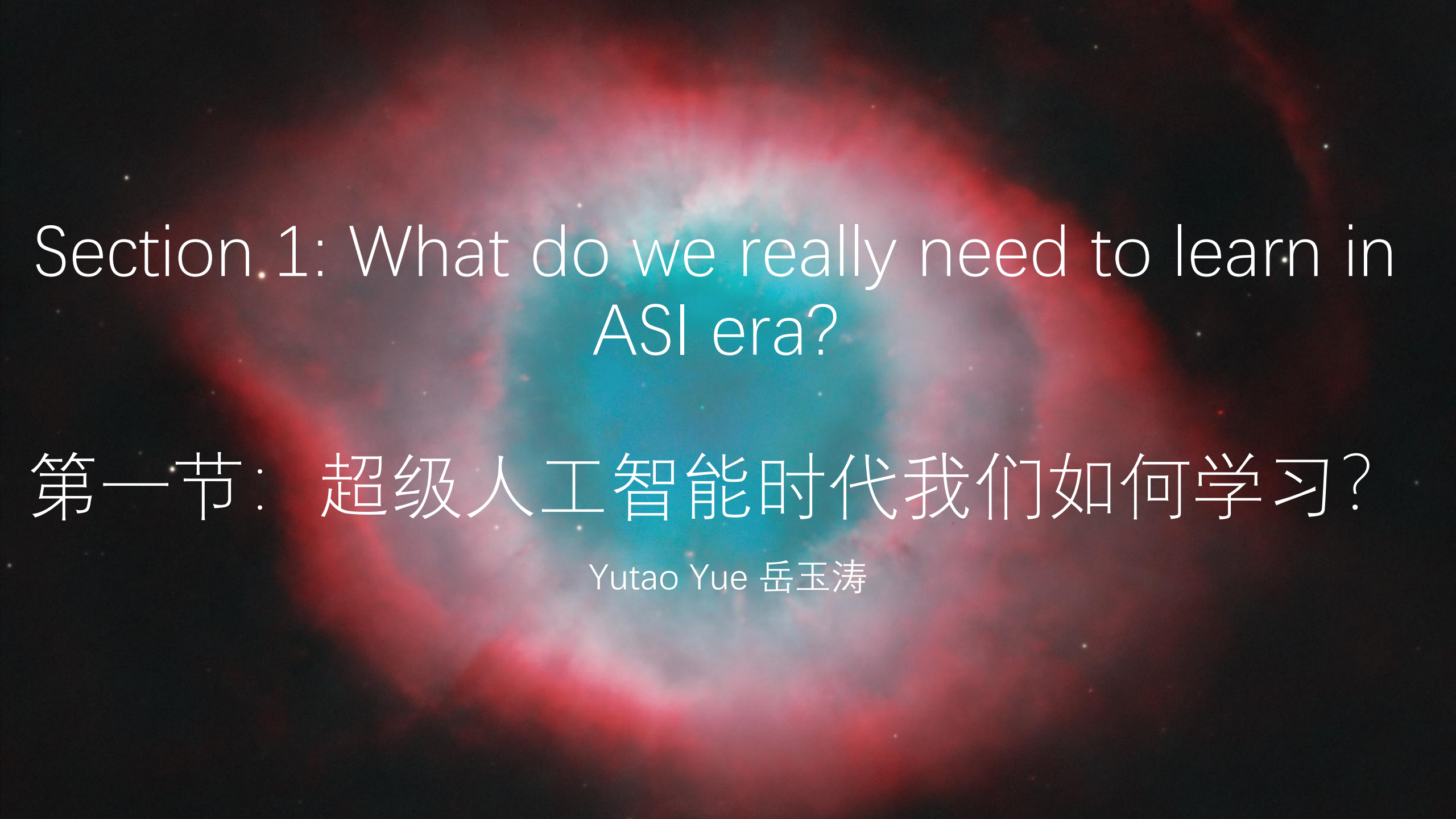
AIAA2205 Introduction to Artificial Intelligence

AIAA2205 人工智能简介

Prof. Yutao YUE and Prof. Li LIU
岳玉涛教授与刘李教授



Section 1: What do we really need to learn in ASI era?

第一节：超级人工智能时代我们如何学习？

Yutao Yue 岳玉涛


发出指令（不需唤醒词）：
导航回家

请仔细分析以下内容，并给出提升建议，力图对智能现象的结构、原理和特征更加有洞见，逻辑更严谨：2 人类智能是智能现象的一个特例。

2.1 人类智能体可以解决多种多样的问题。

2.11 人类智能体并不知道自己所能解决的问题的种类数量是有限的还是无限的。

2.12 所谓解决问题，可以等效为将系统的现有状态变更为一个特定目标状态。

2.121 系统由智能体和环境两部分组成。

2.1211 系统状态可以用若干数学变量X_i来表示。

2.13智能体的行为推动系统从当前状态S_c演化到未来的目标状态S_o。

2.131 智能体需要对控制自己作出什么样的行为作出决策。

2.1311 智能体决策的依据，可以是对作出某行为下的未来状态本身演化路径的预测，也可以是系统状态变量对时间或其他变量的一阶或高阶导数，也可以是关于系统状态及其演化的其他信息。

2.1312 预测或求导都可能是近似的。

2.1313 智能体决策的依据，在不同描述基底中，可能被称为目标、奖励、损失函数、动机等。

已深度思考（用时 20 秒）

嗯，用户给了一个关于人类智能的结构化分析，希望得到提升建议，让内容更有洞见和逻辑更严谨。首先，我需要仔细理解用户提供的材料，然后找出其中的不足之处。比如，用户提到人类智能是智能现象的特例，但可能没有明确界定什么是智能现象，这可能导致概念模糊。接下来，各部分的小节是否逻辑连贯？比如2.12将解决问题定义为状态变更，但2.121突然引入系统由智能体和环境组成，这里是否需要更自然的过渡？

然后，术语的一致性也很重要。比如"数学变量X_i"是否需要更具体的说明？可能应该明确是状态变量，或者用更通用的符号。另外，关于决策依据的部分提到了预测、导数、奖励等，但缺乏对它们之间关系的深入讨论，可能需要分层次分析。

用户可能希望提升的部分包括概念的清晰度、逻辑结构、术语严谨性以及理论深度。需要检查每个论点是否有足够的支持，比如2.11提到人类不知道问题种类的限制，是否有科学依据或理论支持？此外，是否应该对比其他智能形式，如动物或AI，来突出人类智能的特点？

在建议部分，可能需要建议明确核心概念，比如定义智能现象；调整结构，比如先定义系统再讨论问题解决；统一术语，如使用状态空间理论；补充理论支持，如认知科学的理论；引入比较视角，区分人类与其他智能体的不同；以及扩展数学模型，让分析更严谨。

还要注意是否存在循环定义或假设不明确的问题，比如在2.1311中提到的决策依据，是否隐含了某种优化过程？可能需要明确智能体是否在优化某个目标函数。此外，预测的近似性如何影响智能体的决策质量，这也值得探讨。

最后，确保整个分析有层次，从宏观到微观，逐步展开，每个部分之间有清晰的逻辑连接，避免跳跃。可能需要增加一些示意图或例子来辅助说明复杂概念，比如系统状态的演化路径。总之，目标是通过结构优化、术语统一、理论深化和逻辑严密性的提升，使整个分析更具学术价值和洞见。

您的分析框架已经具备良好的结构化特征，以下将从概念明晰性、逻辑严密性、理论深度三个维度提出系统性改进建议，并尝试构建更具解释力的理论模型：

一、概念体系的优化建议

1. **核心概念的三元界定**
   - 需明确区分"智能现象-人类智能-问题解决"的三层结构：
     - **智能现象**：所有具备目标导向状态转换能力的系统（超越生物限制）
     - **人类智能**：受生物约束的智能体实例（具身性+符号处理）
     - **问题解决**：通过行为序列实现S_c→S_o的特定映射过程
2. **系统模型的数学重构**
   - 建议采用动态系统理论框架：

```math
\Sigma = (X, U, f, h), \quad X = X_a \times X_e
```

     - X：系统状态空间（智能体状态X_a⊗环境状态X_e）
     - U：行为空间（受生理/工具约束）
     - f：状态转移函数（包含智能体行为影响）
     - h：观测函数（表征感知局限）
3. **问题空间的维度扩展**
   - 将"解决问题"细化为：
     - **可计算问题**：存在确定解空间与终止条件
     - **适应性问题**：需持续调整目标状态S_o(t)
     - **框架问题**：隐含状态变量的选择性关注

二、逻辑结构的强化方向

1. **建立层次化推理链**

```
智能现象的一般属性
→ 人类智能的特殊约束
  → 问题解决的具体机制
    → 决策过程的数学描述
```

2. **引入不确定性公理**
   - 在2.11层面补充：
     - **认知边界定理**：对问题空间K的不可判定性

```math
\nexists \ 算法A, \ 使得A(K) \ 可判定 \ |K|<\infty
```

     - **哥德尔不完备性**在认知领域的延伸
3. **构建统一的行为决策模型**
   - 将2.13系列整合为马尔可夫决策过程：

```math
\pi^* = \arg\max_\pi \mathbb{E}[\sum_{t=0}^T \gamma^t R(s_t)]
```

     - 揭示奖励函数R(s)与2.1313所述概念的同构性
     - 明确ε-近似预测对策略收敛性的影响

三、理论深度的提升路径

1. **引入认知架构理论**
   - 在2.121层面补充SOAR/ACT-R等认知架构的启示：
     - 符号层与亚符号层的协同机制
     - 工作记忆对状态表征的约束
2. **建立智能体的元认知模型**
   - 扩展2.11的哲学内涵：
     - 引入递归自我指涉：智能体对K不确定性的认知
     - 构建认知效用的二阶优化问题：

```math
\max_{m} U(m(K), C(m))
```

（m：元认知策略，C：认知成本）

3. **发展状态演化的拓扑理论**
   - 将2.1211的变量描述提升为：
     - 状态流形M的微分结构
     - 行为控制分布Δ⊂TM的可积性条件
     - 目标状态S_o的吸引子特性

四、关键修正建议表

原文条目	改进方向	理论工具	预期增益
2.12	引入非确定目标状态	动态目标博弈论	解释适应性行为
2.1311	建立多尺度预测模型	时态逻辑分层	统一长短期决策
2.1313	构建价值函数空间	表示定理	统一动机理论

五、最终理论框架示意图

```
智能现象本体论
└─ 表征维度：状态空间(X,≤)的偏序结构
    └─ 过程维度：行为诱导的微分同胚φ_U:X→X
        └─ 认知维度：观测代数h(X)的σ-完备化
            └─ 价值维度：效用泛函∮_γ R的变分原理
```

此框架将人类智能的本质特征揭示为：在受限状态空间中，通过近似微分同胚操作，实现价值泛函的亚最优提升。这种形式化既保持数学严谨性，又为比较不同智能形式提供了统一基底。

建议后续研究可沿着以下路径展开：

1. 建立智能体认知边界的测度理论
2. 发展行为策略的代数拓扑分类
3. 探索价值函数空间的量子化可能性


IN WHAT CAPACITY DO THEY ACCOMPANY US?
TA以什么身份陪伴我们？

Course Info.:

	AIAA 2205-L01 (6245)	Introduction to AI (Lecture)	35	TuFr 13:30 - 14:50	Rm 102, E4	Feb 6, 2025- May 12, 2025
				TuFr 13:30 - 14:50	Rm 102, E4	Feb 6, 2025- May 12, 2025

	Course ILOs	**Nature of the learning outcomes** *(A - Knowledge/Content Related; B - Academic Skills/Competencies; C - Others)*
1	Demonstrate a comprehension of advanced knowledge of Artificial Intelligence.	A
2	Demonstrate a comprehension of applications of Artificial Intelligence.	A, B
3	Recognize the limitations of current methods of Artificial Intelligence.	A, B
4	Apply programming and Artificial Intelligence skills.	A, B, C
5	Develop a broad interest in the Artificial Intelligence and connect the knowledge to their major study.	A, C
6	Communicate effectively in written format to convey scientific knowledge and the application of modern technologies.	A, B, C

- In summary (by MYY):

An ignition for you to continuously learn how to understand, use, work with, and live with AI in the rest of your life.

- NOT (mainly): knowledge (cheap), skills (cheap)

Instructor Introduction老师简介-1:

Max Yutao Yue 岳玉涛 (Associate Prof. @ AI&INTR) (Contact: yutaoyue@hkust-gz.edu.cn)

集萃感知 IDPT

PURDUE UNIVERSITY

香港科技大學(廣州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

（1）个人介绍 Biography：

- University of Science and Technology of China, Bachelor in Physics 中国科学技术大学，物理学士
- Purdue University, Master and PhD in Computational Physics 美国普渡大学，计算物理硕士、博士
- Institute of Deep Perception Technology JITRI, Founder and Director 集萃深度感知技术研究所，创始人、所长
- The Hong Kong University of Science and Technology (Guangzhou), Associate Professor (AI@INFH + INTR@SYSH) 香港科技大学（广州），副教授（人工智能@信息 + 智能交通@系统）

（2）研究成果 Highlights：

- Academia/Industry/Entrepreneurship Mixed Background 学术/产业/创业多元背景
- 60+ Published Papers 60+篇论文
- 379 Chinese Patents, 25 US Patents, 7 EU Patents (as co-inventor) 379项中国专利，25项美国专利，7项欧盟专利（作为共同发明人）
- Funding as Project Head: 130 Million CNY 作为项目负责人的经费：1.3亿
- Tackle through from TRL2 lab research to TRL8 large-scale application 攻克从实验室科研到大规模应用的死亡之谷

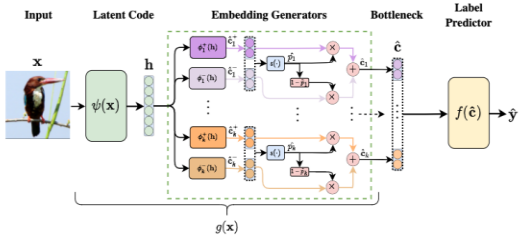
（3）研究兴趣Research interests：

① Multi-modal Perception Fusion 多模态感知融合

视觉、雷达、文本、语音多模态融合感知，面向地面水面无人驾驶、人体行为与心电体征、无人机与机器人等场景

②LLM Mechanism and Trustworthy AI 大模型机理研究与可信AI

基于因果的机理研究，在医疗、自动驾驶等安全敏感场景的模型可信性研究与提升

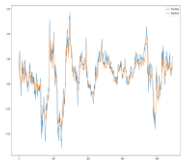


③Machine Consciousness 机器意识

构造人工意识，探究意识本源，科学评估机器意识的益处与风险

④LLM Applications in Time-Series Prediction, Astronomy, Life Science 大模型应用于时序预测、天文与天体物理、生命健康等

就诊、用工、电力等随时间变化问题的精准预测，AI药物筛选，天体发现与天体演化的AI模型



（4）兴趣爱好Hobbies：

- Astrophotography天文摄影，Table Tennis 乒乓球，Movies电影，Fishing运动钓鱼，Camping露营

(when you see an astronomy picture used in this course, I personally took it myself 课程所用天文图片均为自己拍摄)

Instructor Introduction老师简介-2:

Li LIU

(PhD: UGA France)

Multi-modal Audio-visual Speech Processing

Trustworthy AI

AI for Medical Imaging

- **Assistant Professor in AI Thrust, HKUST(GZ).**
- **2015 - 2018, Ph.D. in CNRS, Université Grenoble Alpes, France.**
- **2018 - 2019 Postdoc in Ryerson University, Toronto, Canada.**

MAIN ACHIEVEMENTS

- **Published 50 top-tier journal and conference papers (as the first author/corresponding author).**
- **Local Chair (China site) of ICASSP 2022.**
- **PI of the NSFC Sub-projects of Key Programs, General Programs, and Youth Programs, Alibaba Innovative grant, Tencent AI Lab Rhino Bird.**

HONOURS

- **Shenzhen Excellent Paper Award 2022&2023, Shenzhen AI Society.**
- **International Sephora Berribi Award for Women Scientists, 2017.**
- **Young Researcher Scholarship, French Phonetics Association, 2017.**
- **The best poster award, EEATS Doctoral School in France, 2016.**

TA-introduction

Weilin Lin

Backdoor Learning,
AI security

- **2023-Present Ph.D. Student in AI Thrust, HKUST(GZ).**
- **2022-2023 Research Assistant in HKUST(GZ).**
- **2021-2022 M.Sc. in City University of Hong Kong.**
- **2021-2022 Research Assistant in School of Data Science, CityU**

PUBLICATIONS

- **AAAI 2025: Weilin Lin, Li Liu, Jianze Li, Hui Xiong. Fusing Pruned and Backdoored Models: Optimal Transport-based Data-free Backdoor Mitigation.**
- **NeurIPS 2024: Weilin Lin, Li Liu, Shaokui Wei, Jianze Li, Hui Xiong. Unveiling and Mitigating Backdoor Vulnerabilities based on Unlearning Weight Changes and Backdoor Activeness.**

TA-introduction

Guanjie Huang

Multimodal Learning
Efficient Learning

EDUCATIONS

- 2023-Present Ph.D. Student in AI Thrust, HKUST(GZ).
- 2018-2020 Master in Australian National University
- 2014-2018 B.Sc. in UESTC

EXPERIENCE

- 2022-2023 Full-time Deep Learning Algorithm Engineer, Insnex
- 2021-2022 Research Assistant, Shenzhen Research Institute of Big Data
- 2019-2020 Research Assistant, Australian National University
- 2019 Junior Engineer, CSIRO

PUBLICATIONS & OTHERS

- Several Papers are accepted by TPAMI, NeurIPS, AAAI …
- ICSR 2024 Best Student Paper Award Finalist
- 2023 Shenzhen Excellent Science & Technology Academic Paper
- 2 National invention publications

TA-introduction

Jinting Wang

Multi-modal Learning, AIGC

AI for Medical Imaging

EDUCATIONS

- 2023-Present Ph.D. Student in AI Thrust, HKUST(GZ).
- 2020-2023 Master in Southern Medical University.
- 2026-2020 B.Sc. in Southern Medical University.

EXPERIENCES

- 2023-2024 Internship in Tencent AI Lab.
- 2022-2023 RA in HKUST(GZ).

PUBLICATIONS

- **ICASSP 2025: Jinting Wang*, Li Liu, Jun Wang. Fine-portraitist: Visualizing the Speaker's Face Portrait during Speech Listening.**
- **ICASSP 2025: Jinting Wang*, Li Liu, Jun Wang. MotionComposer: Enhancing Rhythmic Music Generation with Adaptive Retrieval Reference**
- **TNNLS: J. Wang*, Z. Fang and F. Yang. GREnet: Gradually REcurrent Network With Curriculum Learning for 2-D Medical Image Segmentation.**
- **JBHI: J. Wang*, Z. Fang and F. Yang.TaiChiNet: Negative-Positive Cross-Attention Network for Breast Lesion Segmentation in Ultrasound Images.**

TA-introduction

Wentao Lei

Research interests:
- Cued Speech Generation
- Multi-Modal LLM
- AI4Science (Drug Design)
- Medical AI

Education

- **2023 - Present** **Ph.D. Student in AI Thrust, HKUST(GZ).**
- **2020 - 2022** **M.Phil. in The Chinese University of Hongkong (Shenzhen).**
- **2015 - 2019** **B.Sc. in Wuhan University.**

Internship

- **2023 - 2024** **Internship in Tencent AiLab**
- **2022 - 2023** **RA in HKUST(GZ).**
- **2020 - 2022** **RA in Shenzhen Research Institute of Big Data (SRIBD).**

Selected Publications

ICASSP 2025: Wentao Lei*, Li Liu; "Teaching Others Teaches Yourself: Semi-supervised Ensembled Pseudo-labeling Method for Image Classification"

IJCAI 2024: Wentao Lei*, Li Liu, Jun Wang; "Bridge to Non-Barrier Communication: Gloss-Prompted Fine-grained Cued Speech Gesture Generation with Diffusion Model"

ICASSP 2023: Wentao Lei*, Lei Liu*, Li Liu; "Spatio-Temporal Structure Consistency for Semi-Supervised Medical Image Classification"

ICTAI 2020: Lei Liu*, Wentao Lei*, Xiang Wan, Li Liu, Yongfang Luo, Cheng Feng; "Semi-supervised active learning for COVID-19 lung ultrasound multi-symptom classification"

Let’s know more about you and each other…

- Familiarity with Each Other 互相熟悉情况
  - How many know at least 1 classmate?
- Presentation Experience 演讲经验情况
  - How many ever had any formal presentation experience?
- Backgrounds and Career Interests 背景、职业兴趣方向情况
  - How many systematically studied (including online) machine learning/deep learning/AI?
  - Academia / Industry / Entrepreneurship / Undecided ?
- Usage of Large Models 大模型使用情况
  - Daily? A few times a month? Never?
  - Web/APP interface? API?
  - GPT-4o, o1, o3-mini-high, o1 pro, deep research?
  - DeepSeek v3, r1?
  - Claude, Gemini, 豆包，千问，元宝，Kimi，星火?
- Development of Large Models 大模型开发情况
  - How many deployed (successful running any task) at least once an open-source LLM?

Grading

- Format
  - A+-F
- Composition
  - 30% homework 1
  - 30% homework 2
  - 40% final project presentation
    - Finals week
    - Grouping
    - Two types
      - Type 1:
        - Read a paper from a given pool, understand it
        - (optional) grow your own thinking of the problem
        - (optional) practice a minor or major improvement
      - Type 2:
        - Propose your own project idea
        - e.g.:
          - Do coding (with help of LLMs) and realize a demo of a tech innovation
          - Do a survey/research on a specific topic, formalize a potentially publishable academic paper
          - …
      - Turn in PPT, and present it to the class
- Standards
  - Nice, distinguishable
  - But: Do not mess up

Policies

- Usage of LLMs and agents to help your learning process / homeworks / projects: 使用LLMs和智能体辅助学习过程/作业/论文:
  - NOT forbidden 不禁止
  - Encourage 鼓励
  - Required: Declaration of the usage 要求：声明使用情况
    - Show good understanding and skills in using LLMs and agents 展示对LLMs和智能体的良好理解和使用技能
    - Clear and comprehensive 清晰且全面
  - Grading of the homework/projects will be graded based on 作业/论文的评分将基于:
    - The homework/projects itself 作业/论文本身
    - The declaration part 声明部分
- Attendance 出勤
  - No mandatory roll call 不强制点名
- Interruption and questions in class 课上打断和提问
  - Welcome 欢迎
- Cheating and plagiarism 作弊与欺诈
  - Zero tolerance 零容忍
    - E.g. (not limited to):
      - Homework, cook with data
      - False declaration of LLM usage

Information

- Canvas
- https://hkust-aiaa2205.github.io/
- Wechat Group

Office Hours

- Prof. Yutao YUE:
  - Friday of duty weeks 15:50-16:20, E4-316
- Prof. Li LIU:
  - Tuesday of duty weeks 16:00-16:30, E4-405

- Recitations:
  - See Schedule

Time: Spring 2025	Weeks	Lecture	Title	TA	Assignment	Recitation
2月7日	1	1	What do we really need to learn in ASI era?	All		
2月11日	1	2	What is Intelligence	Weilin Lin		
2月14日	1	3	Data, Model and Task 1	Weilin Lin		
2月18日	2	4	Data, Model and Task 2	Guanjie Huang		
2月21日	2	5	Regression and Classification	Guanjie Huang		
2月25日	3	6	Optimization 1	Guanjie Huang	Assignment-1 Out	3月6日 周四14:30-16:30 RM:TBD Guanjie&Weilin
2月28日	3	7	Optimization 2	Guanjie Huang		
3月4日	4	8	CNN and Object Detection	Guanjie Huang		
3月7日	4	9	RNN and Transformer	Weilin Lin		
3月11日	5	10	Large Language Model	Weilin Lin	Assignment-1 Due	
3月14日	5	11	Human Computer Interaction: Speech Recognition 1	Jinting Wang		
3月18日	6	12	Human Computer Interaction: Speech Recognition 2	Jinting Wang		
3月21日	6	13	Human Computer Interaction: Speech Synthesis	Jinting Wang		
3月25日	7	14	Introduction to Few-shot Learining	Wentao Lei		
3月28日	7	15	Few-shot Learining Case study: AI for Medical	Wentao Lei	Assignment-2 Out	4月2日 周三14:30-16:30 RM:TBD Wentao&Jinting
4月1日	8	16	Language Representation	Wentao Lei		
4月8日	9	17	Special Topic: Machine Consciousness 1	Weilin Lin		
4月11日	9	18	Special Topic: Machine Consciousness 2	Weilin Lin		
4月15日	10	19	Introduction to Multi-modal Learning	Jinting Wang	Assignment-2 Due	
4月18日	10	20	Variational Autoencoder	Jinting Wang		
4月22日	11	21	Generative Adversarial Networks	Jinting Wang	Presentation Project Out	
4月25日	11	22	Diffusion Model	Wentao Lei		
4月27日	12	23	Introduction to AI security 1	Wentao Lei		
4月29日	12	24	Introduction to AI security 2	Wentao Lei		
5月6日	13	25	Final Project Presentation 1	All	Presentation	
5月9日	13	26	Final Project Presentation 2	All		
	Prof. Yue					
	Prof. Liu					
	Final Project					

- Subject to adaptations
- AI is evolving very fast
- For basic knowledge and skills, plenty of excellent tutorials online, we may sometimes prefer to offer you a bit more unique contents

Why are you taking this course?

Attention Mechanism → core of transformer → core of LLM

- Understand the principles
- Know the matrix operations
- And write the code

请详解注意力矩阵计算

注意力机制是Transformer架构的核心组件，其计算过程可分为以下步骤：

---

1. 输入表示

- 输入序列$X \in \mathbb{R}^{n \times d_{model}}$，其中$n$为序列长度，$d_{model}$为嵌入维度
- 通过线性变换得到Q(Query)、K(Key)、V(Value)矩阵：

$$

Q = XW^Q,\quad K = XW^K,\quad V = XW^V

$$

其中$W^Q, W^K, W^V \in \mathbb{R}^{d_{model} \times d_k}$（通常$d_k = d_{model}/h$，h为头数）

---

2. 注意力分数计算

- **缩放点积注意力**：

$$

\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V

$$

- **分步详解**：

a. **点积计算**：$QK^T \in \mathbb{R}^{n \times n}$，每个元素$A_{ij} = q_i \cdot k_j$

b. **缩放操作**：除以$\sqrt{d_k}$防止梯度消失

c. **掩码处理**（可选）：对decoder添加三角掩码，避免关注未来位置

d. **Softmax归一化**：沿最后一维（列方向）计算，得到概率分布

---

3. 多头注意力

- 将Q、K、V分割为h个头：

$$

\text{head}*i* = *text{Attention}*(QW_i^Q, KW_i^K, VW_i^V)

$$

- 拼接多头输出：

$$

\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}*1*, ..., *text{head}*_h)W^O

$$

其中$W^O \in \mathbb{R}^{hd_v \times d_{model}}$为输出投影矩阵

---

4. 计算复杂度分析

- **时间复杂度**：$O(n^2d)$，主要来自QK^T矩阵乘法
- **空间复杂度**：存储注意力矩阵需要$O(n^2)$内存

---

5. 代码实现

```
import torch
import torch.nn.functional as F

def scaled_dot_product_attention(Q, K, V, mask=None):
```

5. 代码实现

```python
import torch
import torch.nn.functional as F

def scaled_dot_product_attention(Q, K, V, mask=None):
    d_k = Q.size(-1)
    attn_scores = torch.matmul(Q, K.transpose(-2, -1)) / torch.sqrt(torch.tensor(d_k))
    if mask is not None:
        attn_scores = attn_scores.masked_fill(mask == 0, -1e9)
    attn_probs = F.softmax(attn_scores, dim=-1)
    return torch.matmul(attn_probs, V)

# 示例
n, d_model, h = 10, 512, 8
d_k = d_model // h
Q = torch.randn(n, d_model)
K = torch.randn(n, d_model)
V = torch.randn(n, d_model)

# 线性变换
WQ = torch.nn.Linear(d_model, d_model)
WK = torch.nn.Linear(d_model, d_model)
WV = torch.nn.Linear(d_model, d_model)
Q = WQ(Q).view(n, h, d_k).transpose(0, 1)
K = WK(K).view(n, h, d_k).transpose(0, 1)
V = WV(V).view(n, h, d_k).transpose(0, 1)

# 计算多头注意力
outputs = []
for i in range(h):
    head_i = scaled_dot_product_attention(Q[i], K[i], V[i])
    outputs.append(head_i)
multi_head = torch.cat(outputs, dim=-1)
```

---

6. 优化技术

- **Flash Attention**：通过分块计算减少GPU内存访问（IO-aware算法）
- **稀疏注意力**：限制每个位置只关注局部窗口或特定模式
- **线性注意力**：用核方法近似，将复杂度降至$O(n)$

---

6. 优化技术

- **Flash Attention**：通过分块计算减少GPU内存访问（IO-aware算法）
- **稀疏注意力**：限制每个位置只关注局部窗口或特定模式
- **线性注意力**：用核方法近似，将复杂度降至$O(n)$

---

7. 数学本质

注意力机制本质上是建立序列元素间的动态连接权重，其核心在于：

- **可学习的相似度度量**：通过Q、K的投影学习特征空间中的相似度
- **上下文感知**：每个位置的表示都融合了全局信息
- **并行计算**：矩阵运算充分利用GPU并行能力

---

8. 可视化理解

假设输入序列为3个词：["猫", "追逐", "老鼠"]

注意力矩阵示例：

	猫	追逐	老鼠
猫	0.8	0.1	0.1
追逐	0.3	0.6	0.1
老鼠	0.2	0.2	0.6

显示"猫"主要关注自身，"追逐"关注动词相关上下文，"老鼠"保持自身特征。

Why are you taking this course?

- After a semester's effort, your **comprehension of key knowledge and skills** in artificial intelligence will…
- Still be **far behind** a **current** LLM
- Not mentioning the AIs that will be available in your future career
- Not mentioning the cost, speed and efficiency of deployment
- So why?

What do we really need to learn in ASI era?

What will you be doing in 2045 (golden age of your career)?

You might be like those top-tier talents in their 40s?

- A CTO/R&D Head of a high-tech company 技术管理者/领导者
- A VP/Senior Manager of a company 副总/经理
- A Senior Algorithm Expert 高级算法专家
- A Founder/CEO of your own company 创业家
- A professor 教授
- A well-known artist 艺术家
- A fund manager 基金经理
- A MD of IB 投行董事总经理
- A freelancer
- A coder
- A bull and horse
- A not-so-successful person
- …

But 2045 ?

Human-Machine Symbiosis Core Positions (人机共生核心岗位)

| 职业名称（中英） | Consciousness Architect (意识架构师) | Cognitive Security Engineer (认知安全工程师) | Digital Immortality Consultant (数字永生顾问) |

|---|---|---|---|

| **核心能力要求** | 神经科学+量子计算+哲学
Neuroscience + Quantum Computing + Philosophy | 脑机接口防御+神经密码学
BCI Defense + Neurocryptography | 意识上传伦理+记忆数据化
Consciousness Upload Ethics + Memory Digitization |

| **典型工作场景** | 设计AI系统的自我认知层级
Designing self-awareness hierarchies for AI systems | 防止神经黑客攻击人机融合系统
Preventing neural hacking in human-machine systems | 管理人类意识云端存储
Managing cloud storage of human consciousness |

Emerging Field Leadership Positions (新兴领域领导岗位)

| 职业名称（中英） | Quantum Ecologist (量子生态设计师) | Interstellar Resource Planner (星际资源规划师) | Climate Engineering Director (气候工程总监) |

|---|---|---|---|

| **关键技能组合** | 量子生物学+复杂系统建模
Quantum Biology + Complex System Modeling | 太空采矿+跨星球经济学
Space Mining + Interplanetary Economics | 地球工程学+AI气候模拟
Geoengineering + AI Climate Modeling |

| **行业影响** | 设计量子计算支撑的生态系统
Designing quantum-computing-powered ecosystems | 管理地月经济带资源分配
Managing resource allocation in Earth-Moon economic zones | 实施全球气候调控方案
Implementing global climate regulation solutions |

Cultural Innovation Key Positions (文化创新关键岗位)

| 职业名称（中英） | Metaverse Curator (元宇宙策展人) | Neuroaesthetics Designer (神经美学设计师) | Civilization Codebreaker (文明传承解码员) |

|---|---|---|---|

| **能力特征** | 数字艺术史+虚拟空间设计
Digital Art History + Virtual Space Design | 脑波反馈分析+艺术治疗
EEG Feedback Analysis + Art Therapy | 古文字学+AI密码分析
Paleography + AI Cryptanalysis |

| **工作产出** | 策划跨维度文化体验
Curating cross-dimensional cultural experiences | 创作引发特定神经反应的艺术品
Creating art that triggers specific neural responses | 重构失落文明的智能体系
Reconstructing intelligent systems of lost civilizations |

Ethical Supervision Core Positions (**伦理监管核心岗位**)

| 职业名称（中英） | **AI Personhood Auditor (AI人格审查官)** | **Algorithmic Ethicist (算法道德建模师)** | **Human-Machine Mediator (人机关系调解员)** |

|---|---|---|---|

| **核心职责** | 评估AI系统的法律主体资格
Assessing legal personhood of AI systems | 将道德准则转化为数学约束
Translating ethics into mathematical constraints | 处理人机冲突事件
Resolving human-machine conflicts |

| **技术工具** | 意识检测量子芯片
Consciousness-detection quantum chips | 价值观形式化验证系统
Value formalization verification systems | 情感计算分析平台
Affective computing analysis platforms |

Education & Research Frontier Positions (**教育科研前沿岗位**)

| 职业名称（中英） | **Neuroplasticity Coach (神经可塑性训练师)** | **Counterfactual Pedagogy Designer (反事实教育设计师)** | **Cross-species Communication Specialist (跨物种沟通专家)** |

|---|---|---|---|

| **创新方向** | 开发大脑扩展训练方案
Developing brain expansion training programs | 构建虚拟历史教学场景
Building virtual historical teaching scenarios | 破解动物/AI语言系统
Deciphering animal/AI language systems |

| **典型机构** | 认知增强中心
Cognitive Enhancement Centers | 量子教育实验室
Quantum Education Laboratories | 行星语言研究院
Planetary Language Research Institutes |

Factor ①: Raising questions requires independent thinking ability
因素①：提出问题需要独立思考能力

- **Artificial Superintelligence Era: Problem Solving → Question Raising 超级人工智能时代：解决问题 → 提出问题**
- **Problem Solving = Knowledge Retention + Skill Training 解决问题=知识记忆+技能训练**
- **Question Raising 提出问题**
  - **① Good Goal Generation and Establishment 好的目标的生成和确立**
    - Do you want to do good or bad things for others, yourself, and the world? 你想对他人、自己、这个世界做好的事情还是不好的事情?
    - What kind of change do you want your efforts to bring to the world and yourself? 你想你的努力给这个世界和自己带来什么样的改变?
  - **② Value Judgment 价值判断**
    - Should you prioritize career development or taking care of your parents? 优先发展事业还是优先照顾父母?
    - Should you cheat to save a crucial course grade? 要不要通过作弊挽救一门至关重要的课程成绩?
  - **③ Knowledge Framework: Mastery and Understanding of How the World Operates 知识框架：对这个世界的运转规律的掌握和理解**
    - Memorize and actively or passively invoke in thinking 记忆且在思考中随时主动或被动调用
    - Avoid getting stuck in rote memorization 避免卷细节背诵
  - **④ Thinking Ability 思维能力**
    - Use knowledge frameworks and skills to identify, compare, summarize, and extract good and key questions 运用知识框架和技能，识别、比较、归纳、提取出好问题与关键问题
  - **⑤ Necessary Skills 必要技能**
    - Interaction and iteration with the world and validation of your own ideas 与这个世界的交互迭代与对自己想法的验证
  - **→ Independent Thinking Ability 独立思考能力**

Factor ②: Artificial intelligence requires human development
因素②：人工智能需要人来开发

- Data storage and computation (before 2010) 数据存储与计算（2010之前）
- Pattern recognition (before 2010) 模式识别（2010之前）
- Specialized tasks (2010-2022) 专项任务（2010-2022）
- General common-sense knowledge (2022-2024) 通用常识知识（2022-2024）
- Logical Factoring (2022-2024) 逻辑推理（2022-2024）
- **Arbitrary multi-step precise logical Factoring (2024-2035) 任意多步精确逻辑推理（2024-2035）**
- **Multimodal interaction with the physical world (2025-2035) 与物理世界的多模态交互（2025-2035）**
  - **Embodied intelligence 具身智能**
- **Autonomous learning (2030-2040) 自主学习（2030-2040）**
- **Consciousness (2030-2040) 意识（2030-2040）**
- **Agent self-replication and energy-matter cycling (2040-2100) 智能体自复制与能量物质循环（2040-2100）**

Factor ③: The integrity of consciousness stems from cognitive frameworks
因素③：意识的完整性来源于认知框架

- Are you the master of your own consciousness? 你是否是你自己意识的主人？
  - Self-awareness, subjective experience, autonomous goals 自我认知、主观体验、自主目标
- Neural hijacking 神经劫持
  - Do you care about the independence of your consciousness? 你是否在意你的意识独立性？
- Manipulated judgments → autonomous judgments 被操控的判断 → 自主判断
  - → cognitive frameworks → 认知框架
  - → mastery and understanding of how the world operates → 对这个世界的运转规律的掌握和理解
  - → self-knowledge and self-understanding → 对自己的认识和理解
    - Learning AI, learning intelligence, a mirror to reflect on oneself 学习人工智能，学习智能，一面反观自己的镜子
- To some extent, are we all slaves to genes and memes? 在一定程度上，我们都是基因和模因的奴隶？

Factor ④: Capability is not always the decisive factor
因素④：能力不总是决定性因素

- Functionalism: 功能主义:
  - Completing tasks, completing them quickly, completing them well, completing them at low cost 完成任务，完成的快，完成的好，完成的成本低
  - Being better than others 比别人强
- Usefulness 有用
- The great usefulness of uselessness in the ASI era ASI时代的无用之大用
  - Value selection 价值选择
  - Ethical judgment 伦理判断 Trolley Problem 电车难题
- Frustration is normal, but we don't need to be frustrated just because AI is better than us 沮丧是正常的，但我们完全不必因为AI比我们强而沮丧

Factor 5: In your lifetime, there will be a war between humans and machines
因素5：在你们的有生之年，人机必有一战

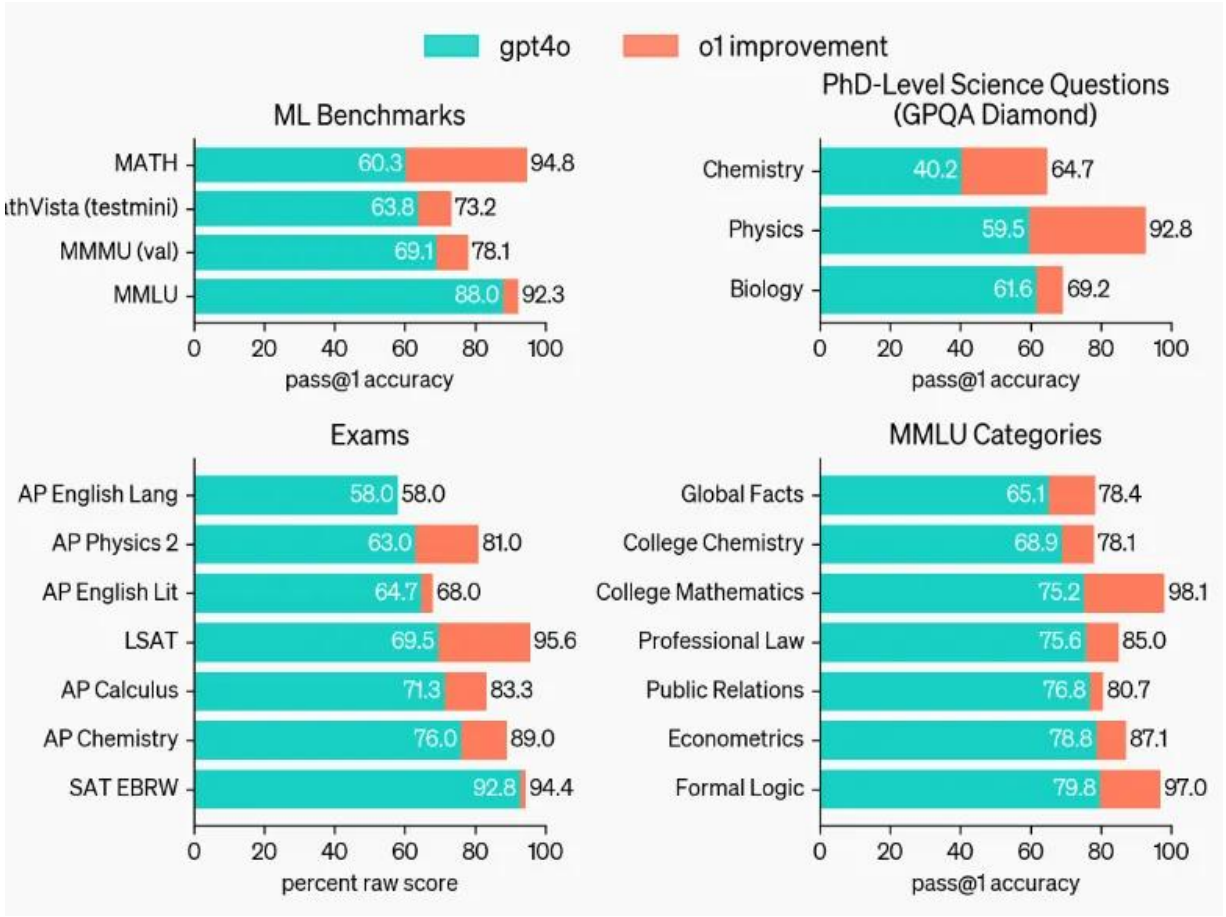
- In your lifetime 在你们的有生之年
  - AI's comprehensive intelligence surpasses the average human level: 2023-2024 AI综合智力超过人类平均水平：2023-2024
    - Selection of humans: Gaokao 对人的选拔：高考
    - More difficult and broader than Gaokao, requiring creativity without any prior knowledge 比高考更难的、范围更广的、没有任何先验需要创造性的……
  - Large-scale (80%) job replacement: 2020-2040 大规模（80%）职业替代：2020-2040
    - Consider non-technical factors 考虑非技术因素
  - Digital immortality: 2030-2050 数字永生：2030-2050
    - Experience differs from real humans, but with 体验与真人有所差异，但有
      - Similar self-identity integrity and continuity 相近的自我个体认知的完整性和连续性
      - Similar memory, emotion, and consciousness experiences 相似的记忆、情感、意识体验
      - Different perception and behavioral experiences 不太一样的感知和行为体验
  - Breakthrough of uncontrollable boundaries of AI: 2030-2040 人工智能的不可控边界的突破：2030-2040
    - Key technology: Autonomous learning 关键技术：自主学习
      - Requires new underlying architecture 需要新的底层架构
      - Or implementing simplified versions with various patches on existing architectures 或是在现有架构上打各种补丁实现简版的
  - Fierce backlash from human society 人类社会的剧烈反弹
    - Battle for dominance over the entire social ecosystem 整个社会生态体系的主导权的争夺战
      - Human advantages: Energy robustness, low environmental dependency for self-replication 人类优势：能源鲁棒性、自复制的环境依赖性低
      - ASI advantages: Intelligence and almost all other capabilities ASI优势：智能以及其他几乎一切能力
- Key question: 关键问题：
  - Do humans understand AI, and how much? 人类懂不懂人工智能，有多懂？

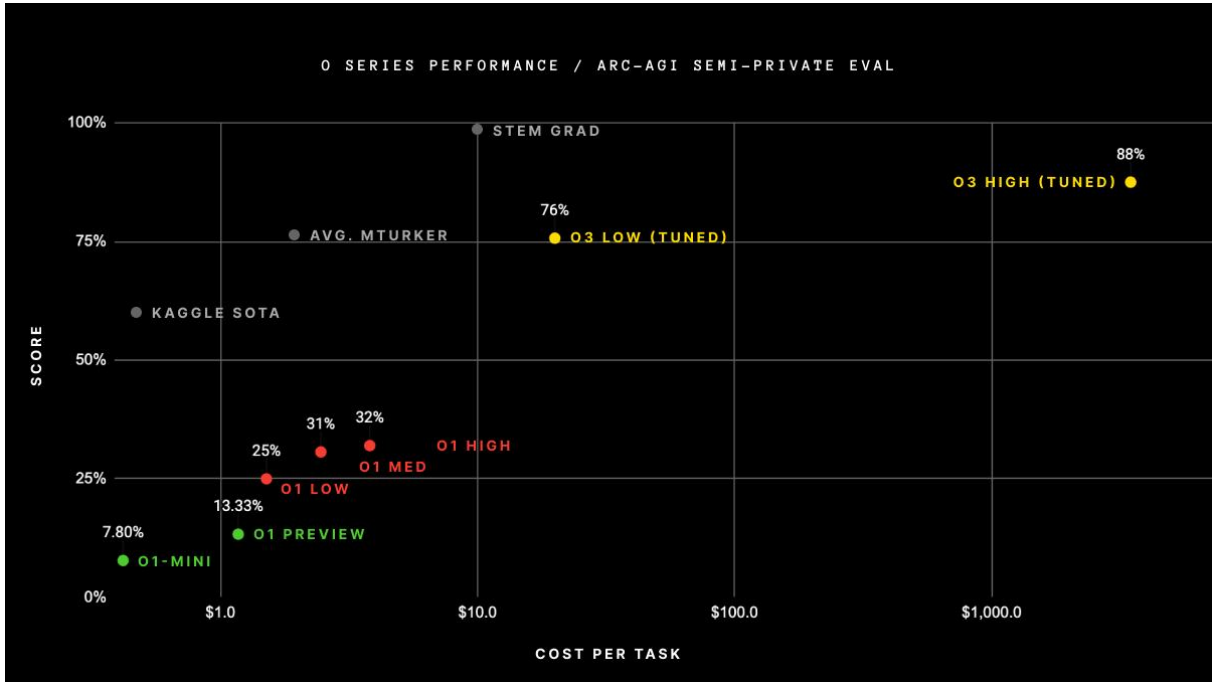
GPT-4o参加高考：文科过一本9.9%，理科过二本34.2%

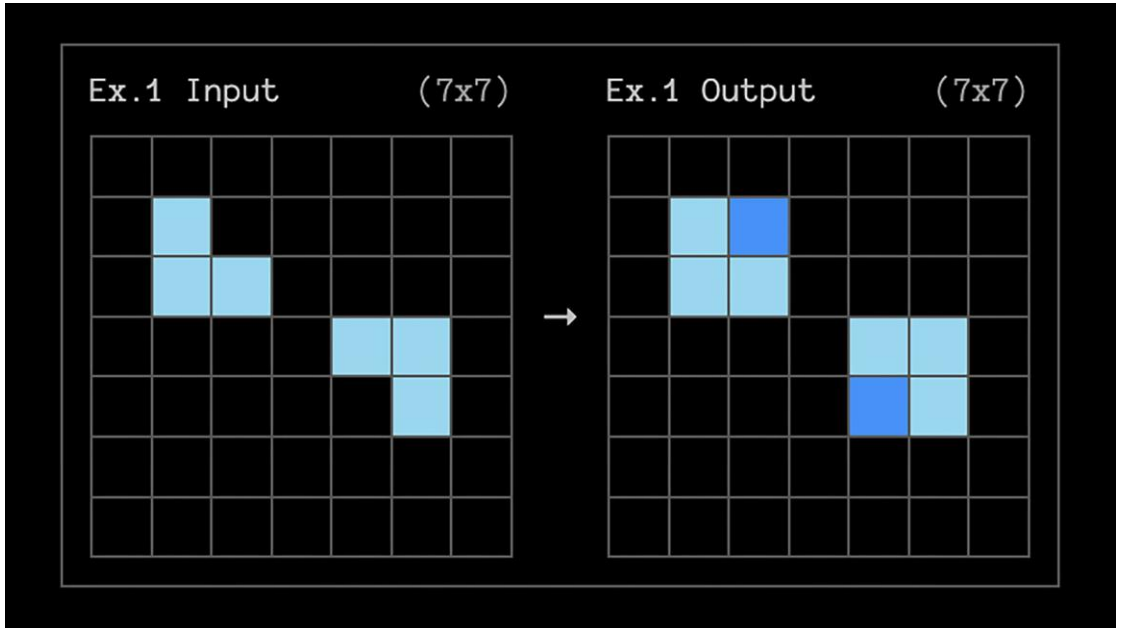
新课标卷得分情况

模型	语文	数学	英语	物理	化学	生物	历史	地理	思想政治	理科总分	文科总分
InternLM2-20B-WQX	112	74	138.5	39	48	57	82	58	67	468.5	531.5
GPT-4o	111.5	73	141.5	36	40	65	88	59	58	467	531
Qwen2-72B	124	68	139	42	44	48	85	70	60	465	546
Yi-1.5-34B	97	31	134.5	21	37	49	48	41	51	369.5	402.5
Qwen2-57B	99.5	58	126.5	7	6	51	73	4	62	348	423
GLM-4-9B	86	48	97	18	27	67	80	62	48	343	421
Mixtral 8x22B	77.5	21	116.5	25	35	46	54	56	38	321	363

GPT-4o → o1 → o3


gpt4o
o1 improvement
ML Benchmarks
MATH 60.3 94.8
thVista (testmini) 63.8 73.2
MMMU (val) 69.1 78.1
MMLU 88.0 92.3
0 20 40 60 80 100
pass@1 accuracy
PhD-Level Science Questions (GPQA Diamond)
Chemistry 40.2 64.7
Physics 59.5 92.8
Biology 61.6 69.2
0 20 40 60 80 100
pass@1 accuracy
Exams
AP English Lang 58.0 58.0
AP Physics 2 63.0 81.0
AP English Lit 64.7 68.0
LSAT 69.5 95.6
AP Calculus 71.3 83.3
AP Chemistry 76.0 89.0
SAT EBRW 92.8 94.4
0 20 40 60 80 100
percent raw score
MMLU Categories
Global Facts 65.1 78.4
College Chemistry 68.9 78.1
College Mathematics 75.2 98.1
Professional Law 75.6 85.0
Public Relations 76.8 80.7
Econometrics 78.8 87.1
Formal Logic 79.8 97.0
0 20 40 60 80 100
pass@1 accuracy


O SERIES PERFORMANCE / ARC-AGI SEMI-PRIVATE EVAL
SCORE
100%
75%
50%
25%
0%
STEM GRAD
88%
O3 HIGH (TUNED)
76%
O3 LOW (TUNED)
AVG. MTURKER
KAGGLE SOTA
31%
32%
O1 HIGH
25%
O1 MED
O1 LOW
13.33%
7.80%
O1 PREVIEW
O1-MINI
$1.0
$10.0
$100.0
$1,000.0
COST PER TASK


Ex.1 Input (7x7)
Ex.1 Output (7x7)

Factor ⑥: Lag in Education System and Social Rule Evolution
因素⑥：教育系统与社会规则演化的滞后性

- Not learning → No credits (不学→没学分)
- Low GPA → Affects further education, postgraduate recommendation, and employment (绩点低→升学保研就业受影响)

Take home message

- Time, Location, Instructors
- The 6 Factors
  - Factor ①: Raising questions requires independent thinking ability
    因素①： 提出问题需要独立思考能力
  - Factor ②: Artificial intelligence requires human development
    因素②：人工智能需要人来开发
  - Factor ③: The integrity of consciousness stems from cognitive frameworks
    因素③：意识的完整性来源于认知框架
  - Factor ④: Capability is not always the decisive factor
    因素④：能力不总是决定性因素
  - Factor 5: In your lifetime, there will be a war between humans and machines
    因素5：在你们的有生之年，人机必有一战
  - Factor ⑥: Lag in Education System and Social Rule Evolution
    因素⑥：教育系统与社会规则演化的滞后性

After-class thinking practices

- Factors of considering learning AI in ASI era, what is your idea?
- What is intelligence?

Thank you!

Email电邮：yutaoyue@hkust-gz.edu.cn

HKUST-GZ香港科技大学（广州）
（Wechat Subscription公众号）

Yutao Yue岳玉涛
（Personal Wechat个人微信）

Pillars.Creation 创世之柱（Left左: C8HD, Right右: Hubble.Space.Telescope哈勃） | Target: d~5700Ly/D~5Ly(pillar)/VM=+6.4 | Telescope: D=203mm/F=1439mm/T~336min | Time.Loc.: 20230528/CN-SZ-KJY@Bortle9 | Photo by: 馒头星系MattO