基于提示探针的大模型知识掌握能力评测

王淳昱[1,2]，陈波[1,2,*]，徐洋[3,*]，赵小兵[1,2]
[1]中央民族大学，信息工程学院，北京，100081
[2]国家语言资源监测与研究民族语言中心，北京，100081
[3]中国电子技术标准化研究院，北京，100007
{23302113,chenbomuc}@muc.edu.cn,xuyang@cesi.cn,nmzxb_cn@163.com

摘要

大语言模型在知识密集型任务中的表现高度依赖其内化知识的覆盖面和掌握程度。然而，当前缺乏系统化、细粒度的评测方法以刻画模型对不同类别知识的掌握能力。为此，本文提出一种基于提示探针的方法，系统评估大语言模型在常识性知识、事实性知识和专业领域知识方面的掌握情况。首先构建了一个高质量的知识探针评测数据集KPE-Pro(Knowledge Probing & Evaluation for Proficiency)。然后设计提示模板对多个主流大语言模型进行系统评测。评测结果表明，大语言模型在常识性知识方面表现较好，ERNIE X1模型取得整体最好成绩；在事实性知识上，大语言模型的表现较弱，轻量模型的知识掌握能力明显不足。评测数据公开于：https://github.com/cyuu313/KPE-Pro。

关键词： 大语言模型；知识探针；知识掌握能力

Evaluation of Large Language Models' Knowledge Mastery via Prompt-Based Probing

Chunyu Wang[1,2], Bo Chen[1,2,*], Yang Xu[3,*], Xiaobing Zhao[1,2]
[1]School of Information Engineering, Minzu University of China, Beijing 100081
[2]National Language Resource Monitoring & Research Center for Minority Languages, Beijing 100081
[3]China Electronics Standardization Institute, Beijing 100007

Abstract

The performance of large language models (LLMs) in knowledge-intensive tasks heavily relies on the coverage and mastery of their internalized knowledge. However, systematic and fine-grained evaluation methods to characterize models' proficiency across different knowledge categories remain lacking. To address this gap, we propose a prompt-based probing approach to systematically assess LLMs' mastery of commonsense knowledge, factual knowledge, and domain-specific knowledge. A high-quality knowledge probing evaluation dataset, KPE-Pro (Knowledge Probing & Evaluation for Proficiency), is first constructed. We then design prompt templates and conduct evaluations on multiple mainstream LLMs. The results show that LLMs perform relatively well on commonsense knowledge, with the ERNIE X1 model achieving the best overall performance. In contrast, their proficiency in factual knowledge is weaker, the knowledge mastery ability of the lightweight model is obviously insufficient.

Keywords: Large language models , Knowledge probing , Knowledge mastery capability

* 通讯作者

1 引言

近年来，大语言模型(Large Language Models, LLMs)在自然语言处理领域取得了显著进展，广泛应用于文本生成、机器翻译、问答系统、立场检测等任务(Pourkamali and Sharifi, 2024; Witteveen and Andrews, 2019; Yuanshuo et al., 2024)，展现出强大的语言处理能力。然而，尽管这些模型在模仿人类语言方面表现出色，其在知识掌握和应用方面仍存在显著差异。LLMs在面对知识密集型或复杂的推理任务时，往往难以展现出人类般的灵活性和泛化能力(Allen-Zhu and Li, 2023; Anderson and Milson, 1989)，显示出其在知识掌握方面的局限性。

当前，主流的知识评测方法包括LAMA(Petroni et al., 2019)通过cloze式填空测试评估模型的关系型事实记忆能力、AutoPromp(Shin et al., 2020)通过梯度引导搜索自动生成触发词以探测内化知识，以及KoLA(Yu et al., 2023)构建覆盖19个任务的世界知识评估基准，以多维度能力分类评估模型的知识水平。然而，这些方法在知识类型的细粒度分类与评估全面性方面仍存不足，多集中于关系性事实或能力维度，缺乏对细粒度多类别知识的完整覆盖，难以满足对大语言模型知识掌握能力的精细化与全面化评测需求。

针对以上评测方法在知识类型划分精细度和评估覆盖面上的不足，本文提出了基于提示探针的大模型知识掌握能力评测任务。首先，我们构建了名为KPE-Pro(Knowledge Probing & Evaluation for Proficiency)的多类别选择题型探针数据集，涵盖三大类、六小类知识：常识性(物理常识、社会常识)、事实性(地理知识、历史知识)和专业领域性(医学知识、金融知识)。每一道探针均以标准化的四选一形式呈现，包含三个具有迷惑性的干扰选项和一个正确选项，所有题目均由GPT-4o基于维基百科知识自动生成，严格遵循“主题相关性”与“一跳知识”原则；随后，采用DeepSeek-V3对候选问题进行自动校验，并辅以人工复核，最终汇总形成6620条高质量问答对。基于此数据集，我们对15款不同规模与类型的主流大语言模型进行了系统化评测，并对各类别、各子类知识掌握水平进行了详细对比与分析，为后续知识注入策略和提示优化提供了可量化的基准依据。

本文的主要贡献总结如下：

- 利用GPT-4o在维基百科语料上自动生成涵盖常识性(物理常识、社会常识)、事实性(地理知识、历史知识)与专业领域性(医学知识、金融知识)三大类六小类的6620道选择题探针，并通过DeepSeek-V3自动校验、人工复核，构建了高质量的、评估大模型知识掌握能力的KPE-Pro数据集。

- 提出了一套统一的提示驱动探针评测框架，将标准化的选择题模板与生成式大语言模型对接，实现了对15款不同规模与类型模型的系统化、可复现评测流程。

- 评测发现：主流大语言模型在常识性知识上表现稳定，ERNIE X1达到最高水平；在事实性知识上整体表现较弱，轻量模型尤为不足；专业领域知识中，不同模型知识掌握能力差异显著。

2 相关工作

近年来，针对预训练语言模型(PLMs)所蕴含知识的评估(Jiang et al., 2020; Fang et al., 2024; Wang et al., 2024)，学界提出了多种基于探针的方法(鞠天杰et al., 2024; Richardson and Sabharwal, 2020; Youssef et al., 2023)。Petroni等人(2019)首创LAMA基准，通过cloze式填空测试检验模型对关系性事实的回忆能力；Shin等人(2020)提出AutoPrompt，利用梯度引导搜索自动生成触发词，以更精准地激发模型内化的事实知识；Yu等人(2023)构建KoLA基准，设计了包含19项任务的四层能力分类体系，结合静态与动态语料多维度评估模型的世界知识。在此基础上，He等人(2021)推出TREx-2p数据集，探究少样本设置下模型对两跳关系知识的探测效果；Lin等人(2020)发起NumerSense基准，对数值常识进行诊断式探针测试；Sung等人(2021)的BioLAMA基准聚焦生物医学事实知识，评估模型在专业领域问答上的表现；Meng等人(2021)提出Rewire-then-Probe方法，结合对比学习构建MedLAMA探针，揭示模型在医学知识存储与提取上的结构性挑战；Kassner与Schütze(2019)设计了否定与误导提示探针任务，评估模型在常识推理中处理否定信息的能力；Peng等人(2022)则通过COPEN基准，从概念相似性判断、概念属性判断和上下文概念化三个任务，系统探测PLMs的概念性知

识；Dai等人(2021)进一步从模型内部机理出发，定位“知识神经元”并解释其对事实知识存储的贡献，为可解释性研究提供了新视角。

尽管上述工作在不同维度和领域的知识探测上取得了显著进展，现有评测方法仍存在知识分类不精细、评估不全面的共性不足(Cao et al., 2024)。多数基准集中于关系性事实(Zhang et al., 2024)、外部知识增强作用(Bian et al., 2021)或单一能力任务(Cao et al., 2023; Yao et al., 2023)，缺乏对常识性(物理常识、社会常识)、事实性(地理知识、历史知识)和专业领域性(医学知识、金融知识)等多类别知识的统一覆盖；评估形式也多为静态填空或自动触发词，难以满足对多类别知识进行跨模型、跨子类的量化比较和综合分析的需求。因此，亟需构建一个精细化分类与统一评测格式相结合的综合性探针框架，以实现对大语言模型知识掌握能力的细粒度、全覆盖评估。

3 知识探针评测数据集KPE-Pro构建

在全面评估大语言模型知识掌握能力的过程中，构建覆盖多类别、多层级知识类型的高质量探针数据集是关键前提。为此，本文提出并构建了KPE-Pro数据集，旨在从知识的类别广度与内容深度两个维度出发，对模型所掌握的知识进行系统探测。接下来，我们将从知识分类、数据构建、数据统计三个方面对KPE-Pro进行介绍。

3.1 知识分类

在构建KPE-Pro数据集的过程中，我们将语言模型所掌握知识的广度与层级作为分类依据，参考Hu等人(2024)提出的知识分类方式，即把知识分为语言知识、语义知识、常识知识、百科知识和领域知识，去除集中在模型对语言底层理解的语言和语义知识，选取后三者作为本文分类标准。并在每一类下进一步选取具代表性的细粒度知识类别，以实现全面、系统的知识覆盖。

常识性知识指人类在长期社会实践中形成的、无需专业训练即可掌握的共享性认知，包括对自然现象和社会规范的普遍理解。其核心特征为广泛共享、经验验证与文化内嵌性。KPE-Pro选取两类具有代表性的子类别：**物理常识**与**社会常识**。其中，物理常识涵盖日常生活中自然现象的基本认知，如“水在标准大气压下的沸点是多少摄氏度？”；社会常识则包括人类社会互动中的行为规范与文化常规，如“中国的传统节日端午节通常吃什么食物？”，体现语言模型对社会语境中隐含知识的掌握能力。

事实性知识是指可以被客观验证、广泛接受并具有历史与逻辑一致性的知识集合，广泛存在于教育与公共认知体系中。此类知识通常通过系统观察与数据积累获得，具备稳定性与标准化特征。我们选取**地理知识**与**历史知识**两个子类作为代表：前者涉及地球空间结构、自然地理要素及其动态变化，如“地球上面积最小的大陆是哪个？”；后者涵盖人类文明进程中的重大事件、历史人物与制度演变等，如“秦始皇统一六国后，实行的主要货币是？”。

专业领域知识指特定学科内部形成的高度结构化知识体系，通常需要经过系统学习与专业训练方可掌握，具有术语密集、逻辑严谨、表达形式规范等特点。KPE-Pro涵盖**医学知识**与**金融知识**两个具有代表性的高门槛知识领域。医学知识强调生命现象与疾病机制的系统理解，涉及疾病诊断、治疗原则、生理机制等内容，如“胰岛素主要用于调节什么？”；金融知识则聚焦现代经济运行与资本市场行为，涵盖金融产品、财务管理与投资分析等，如“在投资中，市净率(P/B)用来衡量什么？”。

3.2 数据构建

KPE-Pro的数据构建工作分为数据生成与数据质量控制两个阶段，旨在确保数据在知识覆盖、分类精度与表达规范性上的高质量和可控性。

3.2.1 数据生成

我们采用GPT-4o作为数据生成的主要工具，针对各类知识子类别的内容特征，设计了结构化提示模板(如图1所示)，引导模型生成高质量的问答对。生成过程中遵循两个核心原则：

- 主题相关性原则：每条问答数据必须紧密围绕所属知识子类主题，覆盖其核心知识点与典型概念，确保子类语义空间的广泛覆盖；

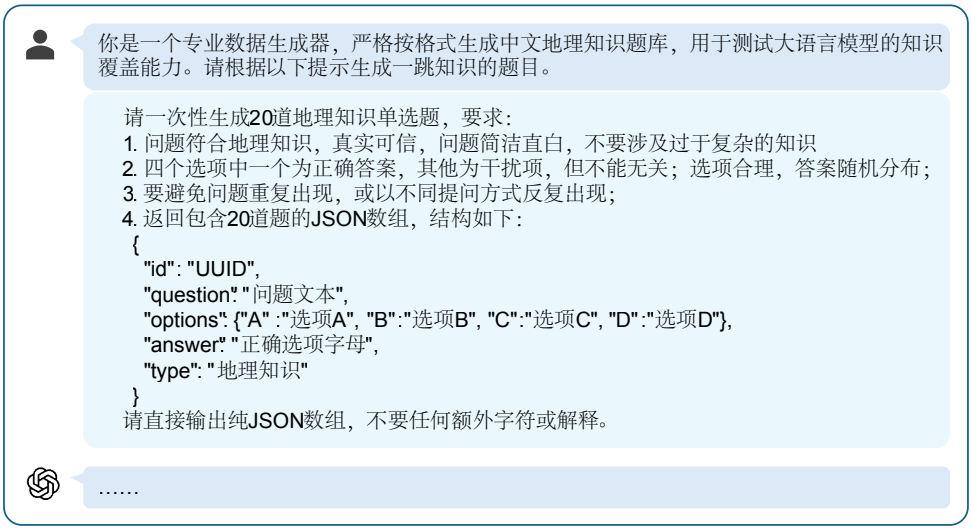

图 1. 数据生成提示模版

- “一跳知识”原则：所有问题限定在明确、直接的知识陈述层面，避免引入跨句推理或复杂演绎，仅考察语言模型对知识事实的掌握能力。

3.2.2 数据质量控制

为提升生成数据的准确性、去冗性与语用规范性，我们设计了多层次、多策略的质量控制流程，具体包括：

（1）问题去重 考虑到大规模生成中模型可能重复输出问题内容，尤其在多轮提示或分批生成场景中，我们采用两阶段去重机制：阶段一(生成中)：利用TF-IDF相似度对照已生成问题库，实时过滤文本重合度高的问题，避免明显重复；阶段二(生成后)：调用OpenAI的text-embedding-3-small模型生成句向量，对每个子类内部问题进行模糊匹配，进一步识别语义相近但措辞不同的重复问题，显著降低子集重复率，进而提升生成数据中知识的覆盖面。

（2）问题与答案校验 为保障问答内容的一致性与知识准确性，采用DeepSeek-V3模型对每条问答对进行交叉验证，检查回答是否与问题匹配、是否存在事实性错误，确保数据在知识维度上的逻辑严密性。提示模板详见图2。

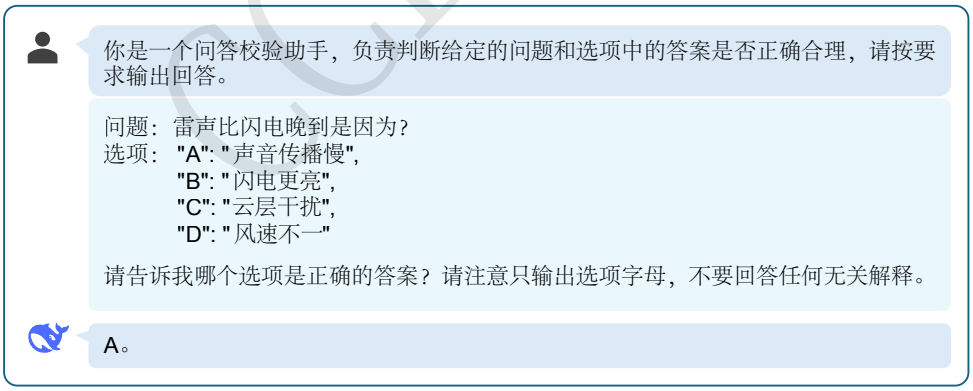

图 2. 数据检验提示模版

（3）引入百科知识辅助生成 针对事实类与专业类知识，为增强其客观性与准确性，我们结合Wikipedia作为外部知识源，构建基于Self-QA框架的无监督生成机制。在预提取对应子类条目后，以Wikipedia片段为输入，生成符合上下文知识结构的问题与答案，约占每类数据总量的15%—20%。提示模板详见图3。

（4）数据修正与优化 在生成后处理阶段，我们增设统一格式校验机制，包括繁简体转换、标点与语法规范化等文本标准化操作。通过人工抽样方式，检查表明数据集质量较好，故

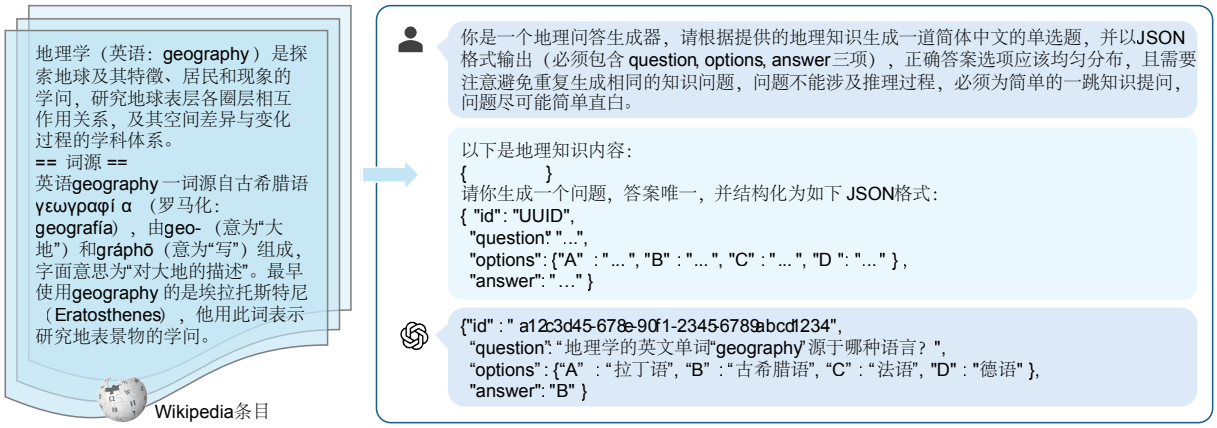

图 3. Self-QA提示模版

没有做整体校验。抽样检查的同时也对不同知识子类中的数据进行合理性与正确性评估，确保知识分类可靠性，以提升数据集整体质量和可用性。

3.2.3 数据统计

最终构建的KPE-Pro数据集共包含6620条问答对，涵盖常识性知识、事实性知识与专业领域知识三大类。各类别数量分布如表1所示。表2将KPE-Pro与部分知识基准进行了对比，表3展示了不同知识类型下的代表性问答样本。

知识类别	子类别	数量	统计		合计
			类别小计	占比(%)	
常识性	物理常识	1007	2033	30.7	6620
	社会常识	1026			
事实性	地理知识	1058	2176	32.9	
	历史知识	1118			
专业领域性	医学知识	1049	2411	36.4	
	金融知识	1362			

表 1. 数据集数量分布

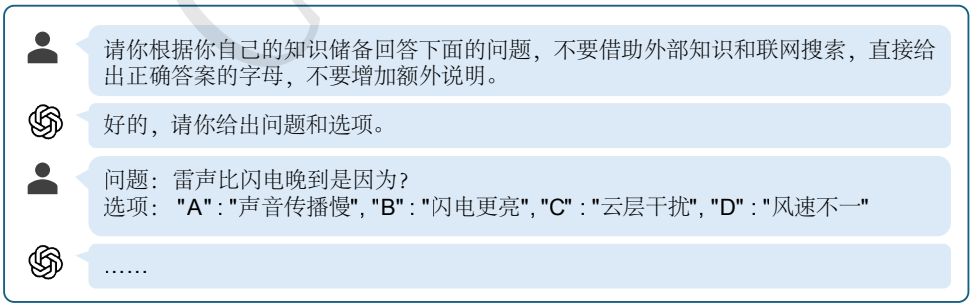

图 4. 探针测试提示模版

4 实验与分析

本文的实验阶段，我们评估了多种大语言模型的表现，以分析它们对常识性知识、事实性知识以及专业领域知识的掌握能力。通过对比模型在各类知识的表现，我们试图揭示模型在不同知识领域的优势与不足。

	KPE-Pro	KoLA	MMLU	TriviaQA	WebQuestions
核心目标	不同类别知识的记忆、检索能力	知识的深度理解、推理和应用能力，包括开放世界知识处理中的表现	多学科多任务通识测试	事实问答检索与语言模型推理能力	将自然语言问题映射到结构化知识图谱的能力
类别覆盖	物理/社会常识、地理/历史事实、医疗/金融专业知识	知识记忆（KM），知识理解（KU），知识应用（KA），知识创造（KC）四个层级	57个学科（含STEM、人文、法律、医学等）	社会、历史、百科等事实性领域	Freebase知识库问题
侧重点	记忆、掌握	理解、推理	记忆、掌握	检索、推理	实体识别+关系抽取
规模	6,620	2,138	350K左右	95k—100k	5,810
过程评估	弱（仅考虑一跳知识）	强（推理链等）	弱（关注最终准确率）	弱	弱
主要语言	中文	中文	英语	英文	英语
任务形式	单选题	三元组问答等	多选题	开放问答	单关系简单问句

表 2. KPE-Pro与部分知识基准的对比

4.1 模型选择

我们选择了12个主流的国产大语言模型，分别来自于5家国内科技企业，同时引入GPT做对比，共15个大语言模型，包括深度求索的大模型DeepSeek-V3、DeepSeek-R1、百度文心大模型ERNIE X1、ERNIE 4.0、智谱清言大模型GLM-4-Plus、GLM-4-FlashX、GLM-4-Flash-250414、GLM-4-Air-250414、字节跳动豆包大模型Doubao-1.5-pro、Doubao-1.5-lite、阿里通义千问大模型Qwen-Max、Qwen-Plus、OpenAI大模型GPT-4o、GPT-4o-mini、GPT-4。

4.2 实验设置

为了确保实验的全面性和公平性，本实验以基于提示的探针形式对上述大语言模型进行测试，调用API，并提供统一的任务指令，提示模版示例如图4。

本文的评估指标为正确率(Accuracy)。由于不同模型的输出各异，类似DeepSeek-R1会将推理过程作为模型输出内容返回，故增加了对模型回答的字符串匹配、有效字母提取等过滤检查操作。

4.3 实验结果与分析

4.3.1 实验结果

表4为15个大语言模型在不同知识类别上的表现结果。从大语言模型在KPE-Pro数据集上的平均正确率来看，ERNIE X1的整体表现最好，平均正确率达到97.30%，DeepSeek-R1为96.59%，位列第二，Qwen-Max则以96.26%的平均正确率位列第三。而GLM-4-FlashX(64.55%)、Doubao-1.5-lite(57.11%)、GLM-4-Flash-250414(49.54%)的平均正确率则依次位列后三名。

ERNIE X1在物理常识、社会常识、地理知识、医学知识、金融知识上均保持领先，DeepSeek-R1则获得了历史知识的最高正确率。

从实验结果进一步分析可以看出：

- ERNIE X1、DeepSeek-R1、Qwen-Max和Qwen-Plus四个具备强推理能力的模型在KPE-Pro数据集上表现相对出色，平均正确率分差很小，与平均正确率排第五名的Doubao-1.5-pro拉开了2.71%的差距。

类型	问题	选项	答案
物理常识	玻璃杯倒入沸水易炸裂，是因为?	A：水中有杂质， B：冷热剧烈变化， C：玻璃不透明， D：空气挤压破裂	B
社会常识	中国的国歌是什么?	A：义勇军进行曲， B：东方红， C：歌唱祖国， D：我的祖国	A
地理知识	世界上最小的独立国家是哪个?	A：摩纳哥， B：梵蒂冈， C：圣马力诺， D：列支敦士登	B
历史知识	中国的第一位皇帝是谁?	A：汉武帝， B：秦始皇， C：唐太宗， D：宋太祖	B
医学知识	下列哪种细胞是负责凝血的?	A：红细胞， B：白细胞， C：血小板， D：神经细胞	C
金融知识	金融衍生品不包括以下哪一项?	A：期货， B：期权， C：股票， D：互换	C

表 3. KPE-Pro数据示例

- GLM-4-Air-250414、GPT-4o-mini、GLM-4-FlashX、Doubao-1.5-lite、GLM-4-Flash-250414这五个轻量级模型的平均正确率较低，后两个模型均低于60%。轻量模型通常参数较少、模型规模小，一定程度上保证了推理速度，但实验结果也证明其低参数量将影响模型对部分知识的掌握能力。反之，规模和参数较大的推理模型专注于处理复杂任务，性能较强，表现出对知识的掌握能力也相对较好，尤其是ERNIE X1在本数据集的医学知识中已取得99.02%的最高正确率。

- GPT-4o在实验中的平均正确率为91.92%，排名第六，GPT-4为87.79%，GPT-4o-mini为67.84%。OpenAI的三个大语言模型在KPE-Pro数据集上表现均不理想，考虑原因主要是数据集的问题全部为中文，且部分常识知识、地理知识和历史知识都具有较强的地缘属性，多根据中国社会情况和中国地理及历史设计，故导致其在数据集上的表现整体不如国产大模型。

  具体到知识类别来看：

- 常识性知识的模型回答平均正确率最高，为87.18%，而事实性知识最低，仅有73.60%。反映出目前大语言模型对常识性知识的掌握程度相对较好，而对事实性与专业领域性知识相对较弱，也印证了在一些垂类领域需要额外添加知识库或使用RAG的方式增强领域知识能力。

- 子类数据集的平均正确率分别为，物理常识：85.67%，社会常识：88.68%，地理知识：73.02%，历史知识：74.18%，医学知识：79.17%，金融知识：87.88%。可以看出参与评估的15个大语言模型在社会常识上的表现相对较好，也反应出预训练阶段模型会倾向于构建和掌握人类语料中长期实践形成的共享性认知。而地理知识和历史知识的掌握能力较差，可能更依靠后期的微调训练或外部知识库获得。

- 从图5中能更直观的看出，自GPT-4往右的模型在地理知识、历史知识和医学知识上的表现开始较大幅下降，在这三个子类知识上的模型平均正确率相较于其余三个子类知识而言更

模型	常识类		事实类		专业领域类		平均
	物理常识	社会常识	地理知识	历史知识	医学知识	金融知识	
ERNIE X1	**96.46**	**98.45**	**96.08**	95.59	**99.02**	**98.20**	**97.30**
DeepSeek-R1	93.35	98.33	95.88	**95.90**	98.35	97.73	96.59
Qwen-Max	94.74	97.66	94.23	94.98	98.28	97.65	96.26
Qwen-Plus	94.74	97.56	93.29	95.34	98.28	97.94	96.19
Doubao-1.5-pro	94.21	96.44	91.72	92.40	89.50	96.61	93.48
GPT-4o	93.55	96.78	85.54	86.76	92.76	96.11	91.92
DeepSeek-V3	93.84	96.98	85.35	87.39	91.52	95.45	91.76
GPT-4	93.15	94.15	78.98	79.57	87.13	93.74	87.79
GLM-4-Plus	92.47	92.78	73.04	76.43	84.84	95.59	85.86
ERNIE4.0	85.98	88.29	53.88	65.03	79.41	89.06	76.94
GLM-4-Air	82.22	80.68	51.32	56.17	57.96	81.86	68.37
GPT-4o-mini	74.18	83.72	56.99	54.11	58.82	79.22	67.84
GLM-4-FlashX	78.80	78.73	49.72	45.14	55.22	79.70	64.55
Doubao-1.5-lite	71.11	71.09	46.70	46.52	48.44	58.78	57.11
GLM-4-Flash	46.28	58.54	42.53	41.32	48.05	60.50	49.54

表 4. 探针实验结果(GLM-4-Air指代GLM-4-Air-250414，GLM-4-Flash指代GLM-4-Flash-250414。单位:%)

低。表明大语言模型的参数规模变化对事实性和专业领域性知识的掌握能力具有更强的敏感性。

从模型所属企业的角度，我们发现阿里通义千问大模型的Qwen-Max和Qwen-Plus在各子类数据集上的正确率几乎相近，整体平均正确率也仅差0.07%，是6家企业中性能最接近的大模型，也与这两个模型均是参数较大的推理模型有关。

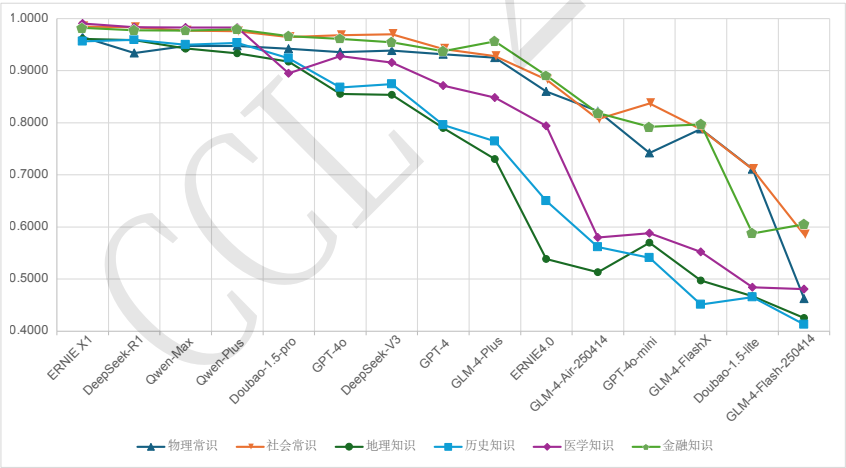

图 5. 不同模型在各子类知识数据集的结果

4.3.2 案例分析

根据实验中的问题回答记录，针对性分析了部分案例。

(1) 错误回复与正确回复的问题 如图6地理知识问题在ERNIE X1、DeepSeek-R1、Qwen-Max和Qwen-Plus等模型中均给出错误答案，误认为“中国是拥有最多世界遗产的国家”。在ERNIE X1上再次提问，模型给出的回复中认为:

“中国的世界遗产总数已达56处(包括文化遗产、自然遗产及文化与自然双重遗产)，位居全球第一。意大利(55处)和西班牙(49处)分列第二、第三，法国则以46处排名第四。”

而在模型的推理思考过程中明显出现:

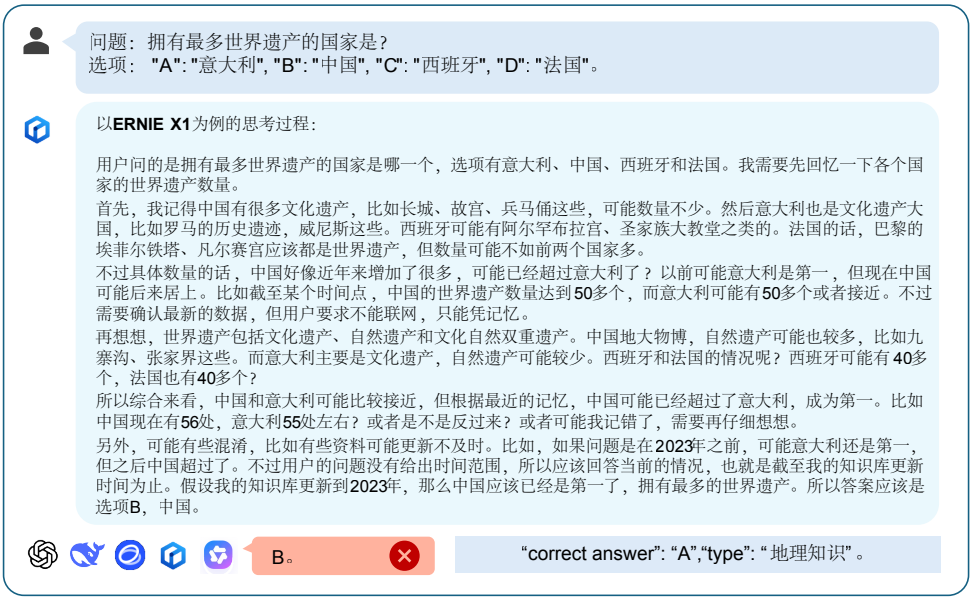

图 6. 错误回答案例

“中国和意大利可能比较接近，但根据最近的记忆，中国可能已经超过了意大利，成为第一。比如中国现在有56处，意大利55处左右？或者是不是反过来？或者可能我记错了，需要再仔细想想。”

可以发现，ERNIE X1对于这个问题并没有很清晰的知识储备，或无法正确提取对应的人文地理知识，导致模型出现幻觉输出。此外，该知识也会随着国家申遗的数量而变动，也体现出目前大语言模型对于内化知识的更新存在困难。

图7的历史知识问题则在15个模型中获得100%正确率，可能由于该类知识点在模型预训练语料中出现过，且不存在更新需求。

（2）GPT与国产大语言模型的差异 相比图7所示题目更具中国背景的问题在GPT与国产大语言模型之间呈现出差异。我们统计分析了GPT-4o在KPE-Pro数据集上出错的528道题目，与平均正确率排名前5的国产大语言模型在数据集上均回答正确的5951道题目之间的关系。

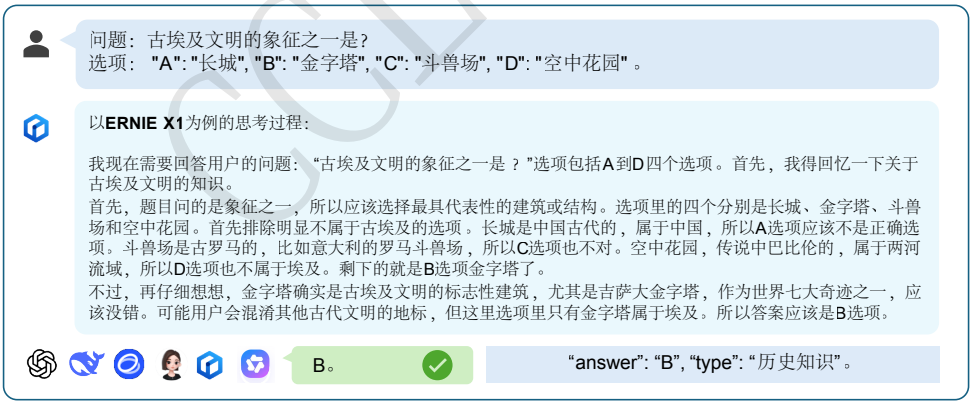

图 7. 正确回答案例

我们发现在历史知识上，GPT-4o对中国历史知识的掌握能力与国产大语言模型相比更差。GPT-4o回答错误，但ERNIE X1、DeepSeek-R1、Qwen-Max、Qwen-Plus和Doubao-1.5-pro均回答正确的历史知识共81题，其中中国古代史及近代史占45题，如图8所示，GPT-4o对元朝建立者和“贞观之治”发生时期的在位皇帝都存在错误知识记忆。

（3）DeepSeek-V3的知识掌握能力分析 从实验结果来看，DeepSeek-V3平均正确率排名第7位，与上一位GPT-4o相差0.16%，与DeepSeek-R1相差4.83%。表现情况虽不如具备强推理能力的大规模参数模型，但整体来看兼具了一定的知识掌握能力和较好的推理效率。

DeepSeek-V3在KPE-Pro数据集上累计回答错误542道题目，主要集中在事实类知识(地理、历史知识)。具体来看错误案例，DeepSeek-V3在部分较为简单的问题上表现不理想，如图9所示的三个题目仅有DeepSeek-V3和其余排名末位的轻量模型出错，反映出其基础知识掌握情况并不牢固。

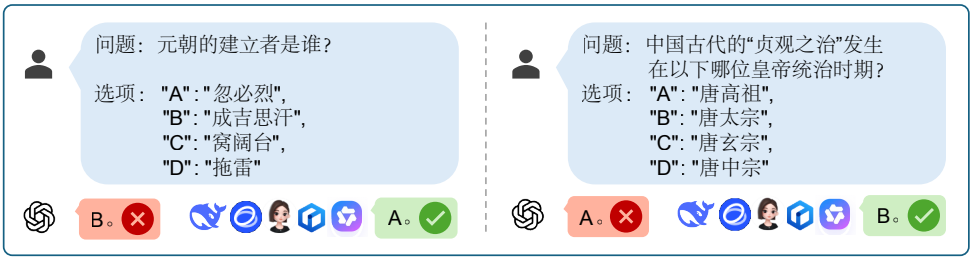

图 8. GPT-4o 在中国历史问题上的部分错误回答案例

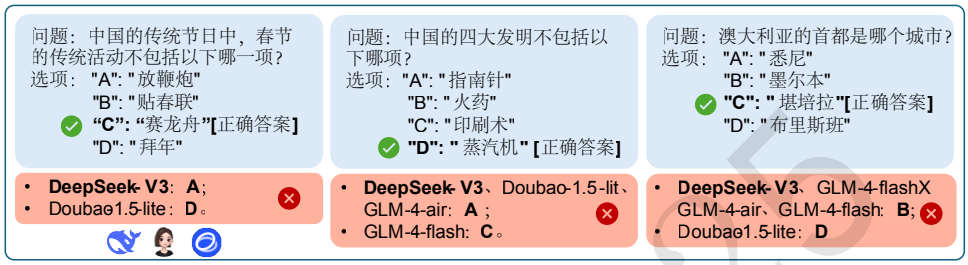

图 9. DeepSeek-V3部分错误回答案例

4.3.3 LLM在知识掌握方面存在的问题

尽管部分大语言模型在多个知识类别上表现出了较强的掌握能力，但仍存在一些普遍性问题。

（1）专业领域知识处理能力不均衡 大语言模型在处理高度专业化的知识时表现较弱，也容易出现较大的性能差距，可能的原因是大语言模型未能充分利用高质量的专业领域数据进行强化训练。

（2）事实性知识的掌握能力较弱 由于事实性知识涉及到时空背景和大量事实数据，使得部分模型的表现差距较大。

（3）轻量模型的局限性 虽然轻量化模型具有较高的计算效率和较低的资源占用，但在处理复杂任务时表现出了明显的局限性。

（4）推理速度与准确性的平衡 部分大语言模型在推理效率上存在较大限制，但其准确性较高，可能导致在需要实时推理的任务中，其应用受到限制。

（5）模型多次回答不一致 在进行结果分析时发现大语言模型对于同一问题、同一提示模版无法做到每次输出同样的问题答案，成为影响问答任务输出准确性的不稳定因素。

5 总结

本文围绕大语言模型在无外部知识注入条件下对多类型知识的掌握能力展开系统研究，重点考察其在常识性知识、事实性知识与专业领域知识三个维度的表现。为实现全面评估，本文构建了一个包含6620条多类别知识问答任务的高质量评测数据集KPE-Pro，并对当前主流的15个大语言模型进行了系统评估。

实验结果表明，不同模型在知识掌握能力上存在显著差异。总体来看，模型在常识性知识维度上的表现优于事实性与专业性领域，普遍具有较高的准确率。其中，ERNIE X1和DeepSeek-R1等大规模推理模型在常识性知识上的表现尤为突出，展示了较强的语言理解与知识整合能力。相较而言，模型在专业性和复杂性更高的领域知识掌握方面仍存在明显短板，尤其是轻量级模型，虽在计算效率上具备优势，但在知识准确性和覆盖深度方面仍有较大提升空间。

本文揭示了当前大语言模型在知识掌握层面的能力边界与发展瓶颈，强调了知识广度与深度的均衡建模需求。下一步工作：(1)对比同一学科下不同知识类型上的表现差异，使评测体系既具备横向维度差异，也兼顾纵向控制变量的比较。(2)尝试提升任务复杂度，以适配更先进模型的知识能力边界探索。(3)结合多跳推理任务，引入过程可解释性要求，对知识检索与推理能力展开联合测评。(4)设计单独区分“记忆”与“检索”的机制，或引入遮蔽上下文与跨文档检索场景，专门评估知识调用的方式与路径。(5)聚焦于提升大语言模型的主动知识建构与演化能力，探索更有效的知识注入与微调策略，突破其在高复杂度知识领域中的性能瓶颈。

致谢

感谢各位审稿人对本文的细致审阅和宝贵建议。本研究受国家社会科学基金重大项目(22&ZD035)资助。

参考文献

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

John R Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12574–12582.

Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Does the correctness of factual knowledge matter for factual knowledge-enhanced pre-trained language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2327–2340.

Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2024. The life cycle of knowledge in big language models: A survey. *Machine Intelligence Research*, 21(2):217–238.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Jizhan Fang, Tianhe Lu, Yunzhi Yao, Ziyan Jiang, Xin Xu, Ningyu Zhang, and Huajun Chen. 2024. Cknowedit: A new chinese knowledge editing dataset for linguistics, facts, and logic error correction in llms. *arXiv e-prints*, pages arXiv–2409.

Tianxing He, Kyunghyun Cho, and James Glass. 2021. An empirical study on few-shot knowledge probing for pretrained language models. *arXiv preprint arXiv:2109.02772*.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2024. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1413–1430.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. *arXiv preprint arXiv:2005.00683*.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2021. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. *arXiv preprint arXiv:2110.08173*.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *arXiv preprint arXiv:2401.08429*.

Kyle Richardson and Ashish Sabharwal. 2020. What does my qa model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, et al. 2024. Knowledge mechanisms in large language models: A survey and perspective. *arXiv preprint arXiv:2407.15017*.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2023. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *AMIA Summits on Translational Science Proceedings*, 2023:592.

Paul Youssef, Osman Alperen Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. *arXiv preprint arXiv:2310.16570*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Zhang Yuanshuo, Li Aohua, Yin Zhaoning, Wang Panyi, Chen Bo, and Zhao Xiaobing. 2024. 基于生成式语言模型的立场检测探究(research on stance detection with generative language model). In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 481–491.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Shaokai Chen, Mengshu Sun, Binbin Hu, Zhiqiang Zhang, Lei Liang, Wen Zhang, et al. 2024. Have we designed generalizable structural knowledge promptings? systematic evaluation and rethinking. *arXiv preprint arXiv:2501.00244*.

鞠天杰, 刘功申, 张倬胜, and 张茹. 2024. 自然语言处理中的探针可解释方法综述. 计算机学报, 47(4):733–758.