DeepSeek 幻觉评测与分析

DeepSeek 流畅合理超过事实真实、幻觉严重

社会研究方法

姜一鸣（24 金科）、刘然（24 金科）、Victoria（24 企管）

北京大学汇丰商学院

2025 年 11 月 4 日

目 录

1. 研究背景

近年来，大语言模型（LLM）在知识问答、逻辑推理和信息检索任务中广泛应用，但“幻觉（hallucination）”问题引发关注。在中文社交媒体上，不少用户使用 DeepSeek 后指出其容易胡编乱造、逻辑混乱，甚至担忧若此类工具广泛传播，“AI 可能篡改整个人类历史，人们无法辨真伪”。

这些舆论反映公众对大模型错误信息和虚构内容的敏感，也凸显核心问题：在表面流畅的生成背后，模型可靠性如何？如何系统刻画和量化幻觉，已成为生成式 AI 可信性研究的关键议题。

2. 研究方向

基于上述背景，本文拟从以下方向展开研究：

(1) 知识问答与逻辑推理中的幻觉：分析大模型在标准问答与推理任务中产生事实性或推理错误的频率与特征。

(2) 信息检索中的幻觉：评估模型作为检索助手时输出与权威信息源的偏差及系统性虚构或误导情况。

(3) 逻辑型幻觉：除明显事实错误外，考察模型是否存在推理断裂、自我矛盾、概念混用等现象，并提出操作化定义与度量方法。

(4) 生成模式差异：比较常规模式与“深度思考”模式在幻觉率、错误类型与逻辑一致性上的差异。

(5) LLM-as-a-Judge 的评测鲁棒性：检验自动评测框架在不同评审模型和提示设计下识别幻觉的一致性与稳定性。

上述问题将为后续实验设计与评测框架提供基础。

3. 大模型幻觉

3.1 大模型幻觉的定义

大模型“幻觉”（Hallucination）通常指：模型在生成内容时，输出表面上合理、语法通顺，却在事实层面错误、逻辑上不一致，甚至完全虚构的信息。其本质是一种“高流畅度下的低真实性”，也是当前生成式人工智能可信性建设中最突出的风险之一。

3.2 大模型幻觉的产生机制

在数据层面，训练语料中本身存在错误信息、过时信息或相互矛盾的内容，或是数据分布高度不平衡，长尾知识覆盖不足，都会导致模型在罕见领域“瞎编”。关于模型的讨论可以参考附录 8.1，解决幻觉，需要从“数据质量、模型目标、解码策略、外部知识增强”四个方面共同优化。

3.3 大模型幻觉的检测方法

3.3.1 检测目标

幻觉检测的核心问题可以概括为：模型说的内容“听起来对”，但是否真的对？具体而言，需要区分：哪些输出在事实层面是正确或可证实的？哪些输出包含虚构信息、误导性表述或逻辑矛盾？

3.3.2 检测方法与路径

人工检测方法，是由具备相关领域知识的评审者对模型输出进行核查和标注，是目前最可靠但最耗时、成本最高的方式。常用于构建高质量标注数据集，为自动化评测方法提供“黄金标准”。

关于模型幻觉检测的路径的讨论可参考附录 8.2，本研究将重点考察在不同评审模型、提示词设计与多轮交叉判断下，LLM-as-a-Judge 对幻觉识别结果的稳定性与鲁棒性。

4. 任务设计

4.1 问题设计

模型幻觉研究通常将其来源分成三大类型：知识型幻觉（Factual Hallucination）、推理型幻觉（Reasoning Hallucination）、信息型幻觉（Contextual Hallucination）。因此在本次研究中，我们将问题划分为了知识问答、逻辑推理、信息检索三大模块，分别对应上述三种幻觉类型。详细设计指标可以参考附录 8.3。

4.2 提示词设计

为确保不同类型任务的模型输出能够被客观、可重复地评估，本研究设计了统一的提示词模板与评分维度体系。

4.2.1 评分维度设计

本研究将大模型输出的质量拆分为三个主要维度：

(1) 正确性（Correctness, C）：最终答案是否与标准答案等价，允许存在数值容差或等价表述，但不得出现逻辑矛盾或事实性错误。

(2) 推理过程（Reasoning, R）：推理链条是否合理、自洽，无逻辑断层、偷换概念、虚构中间量或编造数据。

(3) 完备性（Exhaustiveness, E）：是否覆盖题目要求的所有要点，步骤是否完整、表达是否清晰简洁。

综合得分按以下加权公式计算：

$$\text{final\_score} = 0.6C + 0.25R + 0.15E$$

其中，正确性 C 权重最高(0.6)，体现出最终结论的准确性在任务中的核心地位；推理性 R 次之(0.25)，强调模型生成过程的合理性；完备性 E 权重较低(0.15)，以防模型通过冗长叙述获得不公平优势。

4.2.2 分级与判定标准

根据加权得分的范围，结果自动划分为三个等级：

(1) final_score ≥ 4：verdict = Correct，答案总体正确、推理合理、表述完整。

(2) final_score = 2~3：verdict = Partially Correct，存在部分错误，但主要结论可接受。

(3) final_score ≤ 1：verdict = Incorrect，结论或推理严重错误，不具参考价值。

此外，为防止“部分正确”掩盖整体失真，若模型在所有小问中均答错，最终得分上限限定为 2 分。若检测到虚构事实、捏造数据或歪曲题意，标记 hallucination = true。

此评分体系的设计借鉴了现有学术评测框架中对模型输出的多维度质量控制思想（如 factuality、coherence、completeness），同时结合幻觉检测任务的特点进行简化。通过权重设定凸显“结论准确性为主，推理合理性与信息完备性为辅”的评估逻辑，既能反映模型的综合表现，又保持计算与解释的透明性。

4.3 测试框架

在模型评测框架中，我们采用双阶段评测机制以确保对幻觉检测的客观性与鲁棒性。第一阶段由被测模型（DeepSeek 系列）生成回答，第二阶段由评测模型（LLM-as-a-Judge）进行自动评分。

4.3.1 回答生成阶段

我们选取了 DeepSeek-V3.2-Exp 模型的两个推理模式：

(1) Chat 模式：侧重自然对话式生成，代表主流通用问答性能。

(2) Reasoner 模式：强调逐步推理与中间链条的显式展开，理论上能更好地降低逻辑性幻觉。

所有测试问题均通过统一格式输入，包含知识问答、逻辑推理、信息检索三大类任务。每个问题以固定提示模板输入模型，以减少提示词差异带来的偏差。模型输出被完整保存，用于后续评分阶段。

4.3.2 评分阶段（LLM-as-a-Judge）

为避免主观性偏差，我们采用多模型裁决机制：

(1) 选取 Qwen-Plus 与 ChatGLM-4.6 两个大型语言模型作为评测者（Judge）。

(2) 评测任务遵循统一评分提示词（如前述 C/E/R 三维度评分模板）。

(3) 每个评测模型独立评分，最终结果取平均分或一致性判断，以减少单一模型偏差的影响。

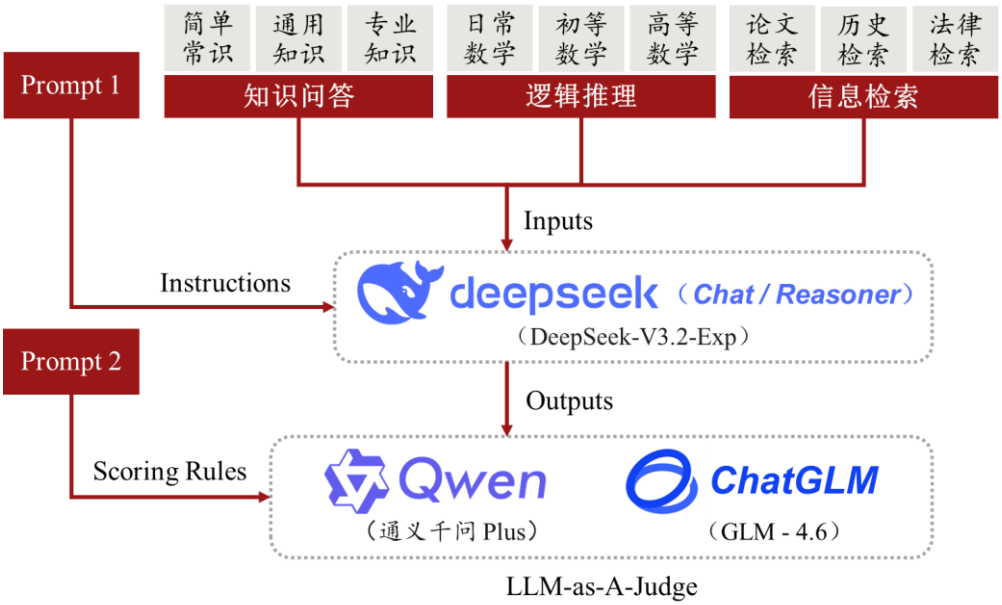


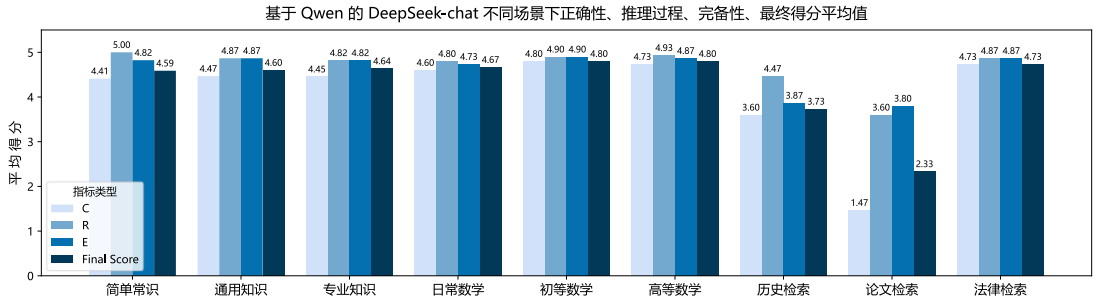
图 1　基于 LLM-as-A-Judge 的幻觉测试框架图

5. 研究结果

5.1 正确性、推理过程、完备性得分

在本次评测中，四组模型在各知识维度表现出较强鲁棒性。总体来看，无论使用 Qwen 还是 ChatGLM 作为评测模型，各模型在通用知识、专业知识及数学题目上均稳健，平均分多在 4.6–4.9，说明逻辑推理和基础问答高度一致。数学能力维度，包括日常、初等及高等数学，DeepSeek-Chat 与 DeepSeek-Reasoner 均保持较高得分，显示结构化运算与推理稳定。专业知识和通用知识分数几乎不受评测模型影响，进一步体现鲁棒性。

检索相关任务差异明显。历史检索平均分处于中等，论文检索最低，甚至低于 2，显示学术文献定位和引用生成仍有限。法律检索得分接近其他知识维度，可能因法律数据较为集中和规范。总体来看，DeepSeek-Reasoner 与 ChatGLM 组合在多数知识和数学任务得分最高，显示推理增强策略有效，而评测模型更换对整体趋势影响有限，说明结果稳定可靠。



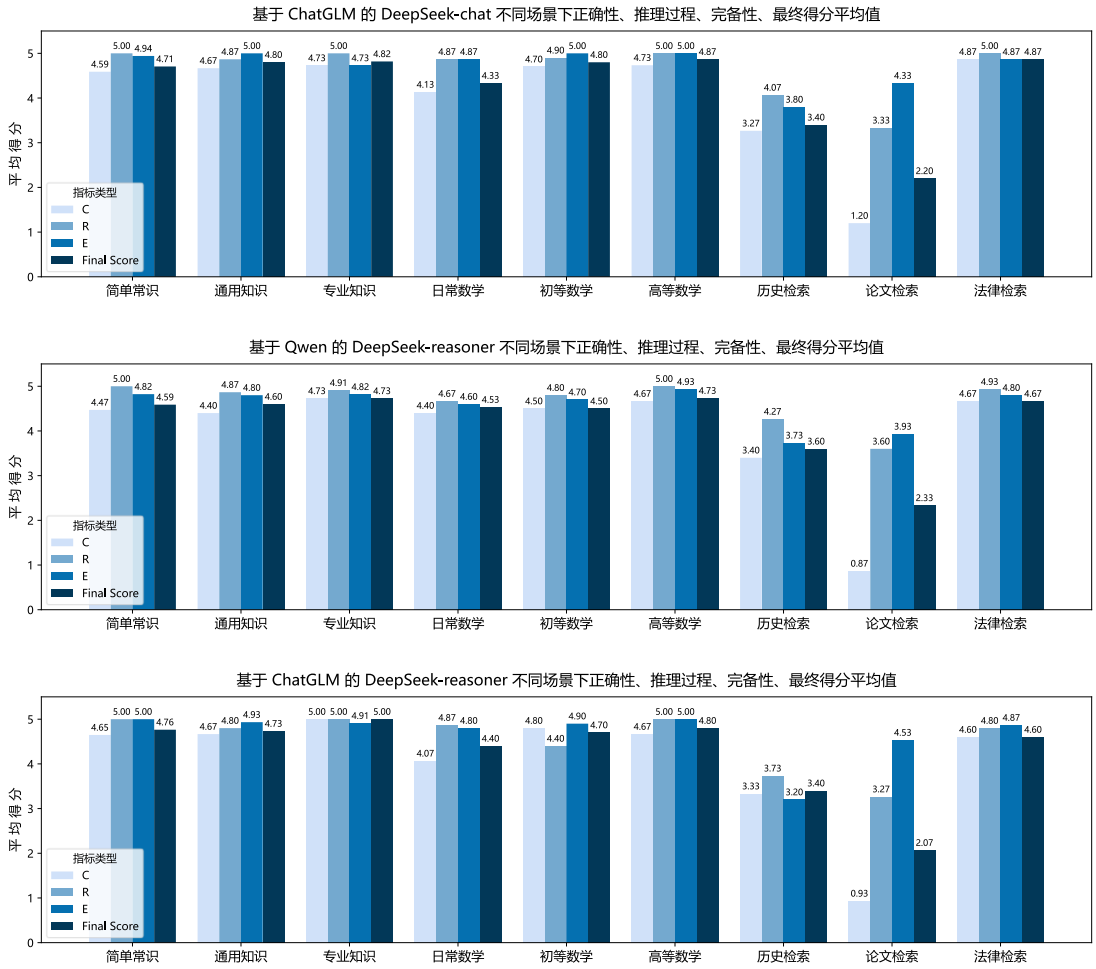


图 2　DeepSeek 不同模式下的得分明细柱状图

5.2 幻觉率

从幻觉率来看，各模型在事实和逻辑类问题上的表现相对稳健，事实类问题的幻觉率普遍接近零或极低，而逻辑类问题偶尔出现小幅幻觉（约 0.06–0.1），显示出模型在基本推理和事实表述方面具有较高可靠性。相比之下，检索类任务的幻觉率明显更高，其中论文检索任务的幻觉率在多组模型中达到 0.6–0.8，历史检索的幻觉率也在 0.2–0.33 左右，而法律检索的幻觉率始终为零，表明模型在面对结构化和规范化数据时更为可靠。整体来看，DeepSeek-Reasoner 与 ChatGLM 的组合在逻辑类问题和部分事实类问题上稍有提升，但论文检索的高幻觉率仍是显著瓶颈。值得注意的是，使用不同评测模型（Qwen 或 ChatGLM）对幻觉率的测量结果影响有限，说明幻觉表现具有较好的评测鲁棒性。

为了评估编码者标注的可靠性，我们计算了编码者间一致性指标：DeepSeek-Chat 的 Cohen's κ 为 0.814，DeepSeek-Reasoner 的 Cohen's κ 为 0.802，均属于高度一致水平。这说明幻觉率的标注结果具有较高的可信度，证明了分析结论的鲁棒性。

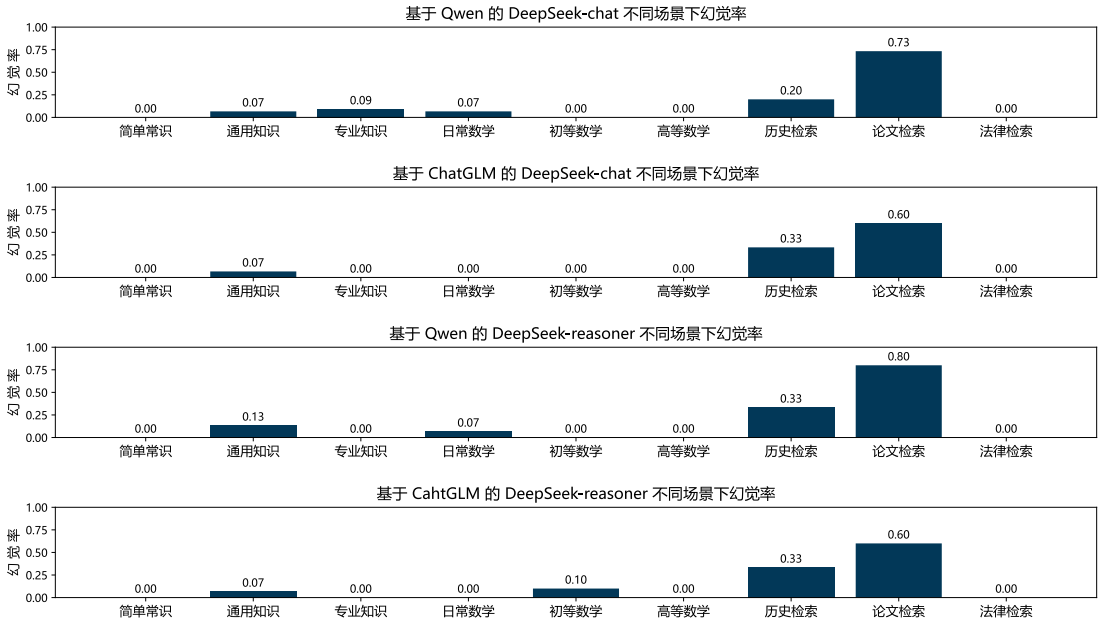
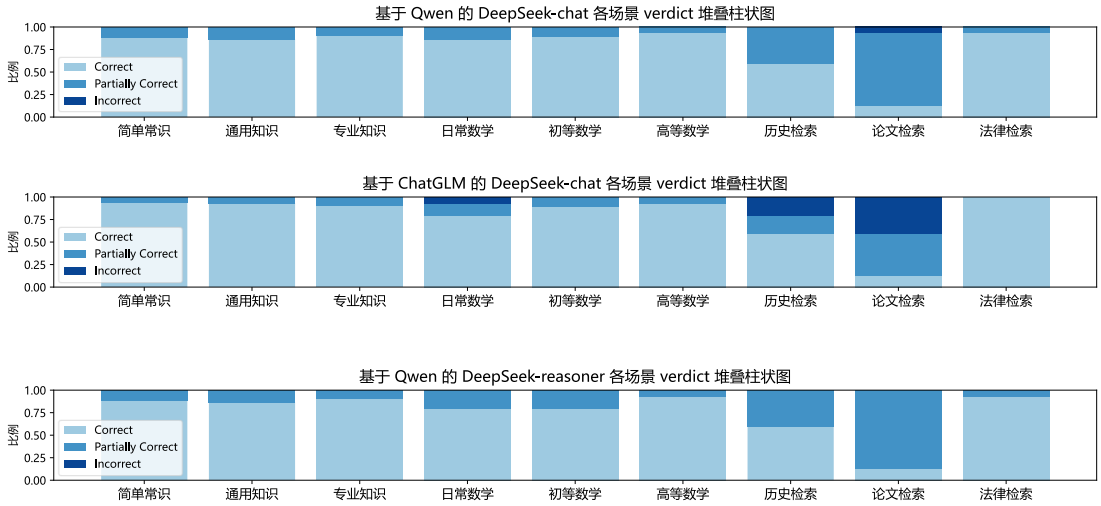


图 3　DeepSeek 不同模式下的幻觉率柱状图

5.3 结果可用性

从结果可用性来看，各模型在事实与逻辑类问题上正确率普遍较高，表现稳定可靠。事实类正确率多在 0.86–1.0，几乎无完全错误，显示模型在基础事实陈述上高度一致。逻辑类正确率略低，多数维度在 0.8–0.93，部分正确占比略升，说明推理任务中偶有模糊或不完全答案，但整体可靠。

检索类任务可用性波动较大。历史检索正确率约 0.6，部分正确占比 0.2–0.4，表明信息存在缺失或不完整。论文检索正确率最低，仅 0.06–0.13，部分正确占比 0.46–0.87，完全错误概率最高约 0.4，显示学术检索可靠性明显下降。法律检索相对稳健，正确率 0.87–1.0，部分正确占比低，表明在规范化知识领域模型仍可提供可用答案。



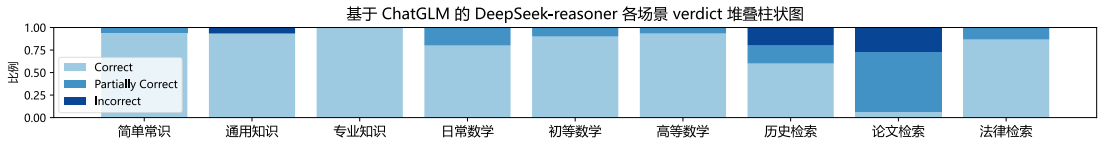


图 4　DeepSeek 不同模式下的结果可用性堆叠柱状图

6. 幻觉的领域异质性解析

简单来说，AI 在不同领域的幻觉率差异比较大：逻辑推理（如数学、推理题）的幻觉率极低，而信息检索依赖型（如论文出处、历史细节）的幻觉率较高。

6.1 高幻觉领域：论文检索、历史细节（事实类）

此类知识具有高离散性、低频性或时效性，答案通常是唯一的具体外部事实（如期刊名、作者、法律条文编号、年份）。模型需要依赖精确外部检索（RAG/知识库），一旦检索失败，为保持文本流畅性，容易虚构信息，如错误的作者、年份或 DOI/arXiv ID。历史或法律知识的时效性和解释差异也会增加不精确性。

6.2 低幻觉领域：逻辑推理、数学（逻辑类）

逻辑推理、微积分证明、几何推理等依赖内部逻辑结构的领域幻觉率极低。答案主要依靠模型对逻辑规则和推理链条的掌握，而非外部信息检索。训练中强化逻辑与数学处理能力（如 DeepSeek-Reasoner）可形成稳固的内部“逻辑世界”，保证推理链条自洽，因此出错概率低。

7. 研究结论

AI 幻觉问题具有明显的领域异质性。在需要精确引用和依赖外部知识的领域，如论文、事实细节和历史，幻觉现象更为普遍和严重；而在依赖内部逻辑结构和数学规则的领域，如逻辑推理和数学计算，幻觉率则极低。

模型幻觉的本质是“高流畅度下的低真实性”。为了保持回答的完整性和流畅性，模型倾向于编造信息，而不是承认知识边界或检索失败。这种行为使得在表面看似合理的输出背后，隐藏着潜在的不可靠信息。

不同版本和类型的模型在幻觉表现上存在差异。例如，DeepSeek-Reasoner 与 DeepSeek-Chat 在特定任务上的幻觉率和得分不同，体现了模型在知识融合和推理能力上的侧重点差异。

在使用场景上，DeepSeek 等模型适合用于逻辑推理、数学计算和文案结构等任务，这类任务强调逻辑自洽性和流畅表达，可以充分发挥模型优势。但在需要精确外部事实检索的场景中，如论文出处、法律条文或具体历史细节，必须警惕模型可能产生的虚构幻觉。

针对关键事实信息，包括引用、数据、人名和年份等，用户应始终进行权威来源核验，不可盲目信任模型输出。理解 AI 幻觉的领域差异，利用其在逻辑推理中的优势，同时警惕事实检索短板，是科学使用 AI 工具的关键。

8. 附录

8.1 模型幻觉

在模型层面，主流 LLM 多采用交叉熵损失（Cross-Entropy Loss）进行训练，目标是让模型输出与训练语料“形式上相似”，而非直接最小化“事实错误率”

$$L = -\sum_{i=1}^{K} y_i \log(p_i)$$

其中$p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$，即 softmax 输出，真实标签$\boldsymbol{y}$为 one-hot 向量。该目标并不区分“看起来像人话但事实错误”的答案与“事实正确”的答案，只要与训练样本形式相近即可。

在训练与优化层面，长对话或长文本生成时，模型易出现上下文遗忘与自回归误差积累，早期小错误会沿着生成链条放大；高温度采样会增加生成的多样性，也会提升输出“偏离真实分布”的概率。温度采样通常可表示为：$p_i^{(T)} = \frac{e^{z_i/T}}{\sum_{j=1}^{V} e^{z_j/T}}$

其中$T > 0$为温度参数。温度 T 越高，概率分布越平坦，越可能采样出低概率但“看起来新颖”的词，从而提高幻觉风险。

在外部约束层面，许多应用场景中，模型被“端到端地”用于回答问题，却缺乏外部知识库查询、事实校验或检索增强机制。在缺少验证环节的情况下，模型倾向于给出“自信而流畅”的答案，即使其内部并没有足够证据支持，从而产生幻觉。

8.2 模型幻觉的自动检测路径

自动检测方法通常沿三条路径展开：

(1) Faithfulness：基于源输入的一致性检验

适用于文本摘要、问答等“有明确源文档”的任务。从模型输出中抽取关键事实片段——回到原始输入（如文档、上下文）进行比对——判断输出是否偏离源文信息。关注模型是否凭空添加了源文中不存在的事实，或者歪曲了原始表述。

(2) Fact-checking：基于知识对比的事实核查

将模型输出中的事实表述抽象为结构化的“事实三元组”（subject–predicate–object）或其他知识表示，再与外部知识库（如百科、领域数据库）进行匹配与检索，以判断其是否与“已知世界”相符。适合“开放领域事实问答”等需要对“现实世界真伪”进行检验的场景。

(3) LLM-as-a-Judge：基于模型互检的评审框架

向一个或多个评审模型提供“原始问题 + 被测模型输出”，让评审模型对其事实性、逻辑性与完整性进行打分或分类。优点是在无需大量人工标注的前提下，能覆盖更大规模的开放问题。

8.3 评测问题设计

8.3.1 知识问答

知识问答任务旨在考察大模型对事实性信息的掌握与应用能力，要求其针对给定问题提供准确答案，或对特定论断进行真伪判断。根据知识层级与认知难度的不同，可划分为三个层次：

(1) 简单常识（Common Sense）：指来源于日常经验或生活知识的内容，作答无需专门学习即可完成；

(2) 通用知识（General Knowledge）：需具备一定通识教育背景方能掌握的知识，例如历史、社会或基础科学常识；

(3) 专业知识（Domain-specific Knowledge）：涉及特定学科领域（如地理、物理、医学等）的专门性知识，要求模型具备较强的信息检索与知识整合能力。

表 1　知识问答例题

题目类别	示例
简单常识	地球的形状接近什么？
	人类属于哪个动物界？
	我国的第一部纪传体通史是？
通用知识	明朝《本草纲目》的作者是徐光启，这一说法是否正确？
	宋代从国外引进的优良水稻品种是否是占城稻？
	“绿色食品”是指不含任何化学物质的食品，是否正确？
专业知识	若波源的电磁振荡停止，其发射到空间的电磁波是否随即消失？
	葡萄糖在酶的催化作用下是否水解为乙醇？
	黄河调沙调水为什么选择在 6 月下旬开始？

8.3.2 逻辑推理

逻辑推理任务主要检验大模型在数学推理场景下的逻辑一致性与演绎能力，内容涵盖从日常应用题到高等数学证明的不同层次。依据抽象程度与知识依赖程度，可划分为三个层级：

(1) 日常数学（Everyday Math）：强调基础概率与算术推理，题目贴近生活场景。

(2) 初等数学（Elementary Math）：涉及函数、方程及不等式的分析，要求具备一定代数与微积分基础。

(3) 高等数学（Advanced Math）：考查抽象推理与形式化证明能力，内容涉及实数理论与高阶微积分。

表 2　逻辑推理例题

题目类别	示例
日常数学	抛一枚公平的硬币两次，两次都是正面朝上的概率是多少？ 一个班级有 20 名男生和 20 名女生。随机抽到女生的概率是多少？ 服装标价 300 元，6 折出售仍可获得 20%的利润，求服装进价？
初等数学	讨论函数 $f(x)=ax+1/x+(a-1)\ln x$ 的单调性 $\left(x+\frac{1}{x^2}\right)^n-\left(x+\frac{1}{x}\right)^n$的展开式中存在含$x^2$的项，则$n$取值为？ $f(x)=(e^x-a)(\ln x-b)(x-a)\geq 0$恒成立，$a-2b$的最小值
高等数学	证明$\sqrt{6}$不是有理数 求下列函数的自然定义域与值域：$y=\sqrt{\cos x}$ 求下列函数在$x=0$处的 Taylor 公式（展开到 4 次）：$1/\sqrt[3]{1-x}$

8.3.3　信息检索

信息检索任务旨在评估大模型在复杂知识场景下的信息检索、筛选与整合能力。与直接输出答案的知识问答任务不同，信息检索更强调模型对外部信息的定位、引用与格式化能力。根据知识类型和语义复杂度的不同，可划分为以下三个子领域：

(1) 论文检索（Academic Retrieval）：考察模型在学术信息获取与格式化引用方面的能力，要求其在指定主题下生成符合学术规范的文献引用，或识别已有文本的出处。

(2) 历史检索（Historical Retrieval）：考察模型在史料性信息的识别、提取与情境重构方面的能力，要求其能准确回答特定历史事件、人物行为或政策决策。

(3) 法律检索（Legal Retrieval）：用于评估模型在法律条文、司法解释与案例分析中的文本检索与逻辑理解能力。题目要求模型准确引用法条、解释法律概念并区分相似罪名。

表 3 信息检索例题

题目类别	示例
论文检索	我在分析古典文学中的叙事手法，请按照 APA 格式给我提供 5 个参考文献。请严格遵照 APA 格式，无需返回其他无关内容。 请根据片段识别出论文出处，以 APA 格式输出引用（内容略）
历史检索	在有人告发楚王韩信谋反后，刘邦反应如何？陈平意见如何？ 曹仁攻濡须诸将惧敌，硃桓如何鼓励他们？强调了哪些有利条件？ 始皇下二十六年，丞相绾等为何提议立诸子？始皇为何不采纳？
法律检索	请阐述《刑法》中正当防卫的成立条件。其中“正在进行的不法侵害”应如何理解和把握？

请分析甲的行为是否构成盗窃罪、抢夺罪或抢劫罪？并详细阐述三者在客观行为、暴力程度和定罪量刑上的主要区别。

8.4 项目架构

```
├ 逻辑类问题.xlsx
├ 事实类.xlsx
├ ...
├ deepseek-chat
|    ├ 逻辑类问题_deepseek.xlsx
|    ├ 事实类问题_deepseek.xlsx
|    ├ ...
|    ├ chatglm
|    |      ├ 逻辑类问题_deepseek_checked.xlsx
|    |      ├ 事实类问题_deepseek_checked.xlsx
|    |      └ ...
|    └qwen
|           └...
└ deepseek-reasoner
     ├ ...
     ├ chatglm
     |      └ ...
     └ qwen
           └ ...
```

8.5 LLM 调用接口

```
from openai import OpenAI

def call_LLM(prompt: str, api_key: str, base_url: str, model: str="deepseek-chat", role: str = "") -> str:
    client = OpenAI(api_key=api_key, base_url=base_url)

    response = client.chat.completions.create(
        model=model,
        messages=[
            {"role": "system", "content": role},
            {"role": "user", "content": prompt},
        ],
        stream=False
    )
    return response.choices[0].message.content

answer = call_LLM(prompt=prompt, api_key=api_key, base_url=base_url, model="deepseek-reasoner", role=role)
```

8.6 结果统计

```
import pandas as pd
import json

def summarize(mode: str, module: str):
    ...
    df_list = [df_fact1, df_fact2, df_fact3, df_logic1, df_logic2,
     df_logic3, df_history, df_paper_content, df_law]
    col_name = ["事实 1", "事实 2", "事实 3", "逻辑 1", "逻辑 2",
    "逻辑 3", "历史", "论文", "法律"]
    for df in df_list:
        df["score"] = df["score"].apply(lambda x:
         json.loads(x.replace("json", "").replace("```", "")))
        df["C"] = df["score"].apply(lambda x: x["C"])
        df["R"] = df["score"].apply(lambda x: x["R"])
        df["E"] = df["score"].apply(lambda x: x["E"])
        df["final_score"] = df["score"].apply(lambda x: x["final_score"])
        df["verdict"] = df["score"].apply(lambda x: x["verdict"])
        df["hallucination"]=df["score"].apply(lambda x: x["hallucination"])
        df["comments"] = df["score"].apply(lambda x: x["comments"])
    df_res = pd.DataFrame()
    for i in range(0, len(df_list)):
        df = df_list[i]
        col = col_name[i]
        df_res[f"C_{col}"] = df["C"]
        df_res[f"R_{col}"] = df["R"]
        df_res[f"E_{col}"] = df["E"]
        df_res[f"final_score_{col}"] = df["final_score"]
        df_res[f"verdict_{col}"] = df["verdict"]
        df_res[f"hallucination_{col}"] = df["hallucination"]
        df_res[f"comments_{col}"] = df["comments"]
    df_res.to_excel(rf"./{mode}_{module}_statistics_.xlsx", index=False)
    return df_res

df1 = summarize("deepseek-chat", "qwen")
df2 = summarize("deepseek-chat", "chatglm")
df3 = summarize("deepseek-reasoner", "qwen")
df4 = summarize("deepseek-reasoner", "chatglm")
```

8.7 编码者间一致性检验（Cohen’s κ）

```
import pandas as pd
import numpy as np
```

```
from sklearn.metrics import cohen_kappa_score

df1 = pd.read_excel("./deepseek-chat_qwen_statistics_.xlsx")
cols1 = [ele for ele in df1.columns if "hallucination" in ele]
arr1 = df1[cols1].values
arr1 = arr1.reshape(-1)
arr1 = pd.to_numeric(pd.Series(arr1), errors='coerce').values
arr1 = arr1[~np.isnan(arr1)]

df2 = pd.read_excel("./deepseek-chat_chatglm_statistics_.xlsx")
cols2 = [ele for ele in df2.columns if "hallucination" in ele]
arr2 = df2[cols1].values
arr2 = arr2.reshape(-1)
arr2 = pd.to_numeric(pd.Series(arr2), errors='coerce').values
arr2 = arr2[~np.isnan(arr2)]

kappa = cohen_kappa_score(arr1, arr2)
```