# 金融大模型幻觉问题

**一、大模型幻觉定义**

AI“无所不能”的时刻似乎近在咫尺，但真实情况并非如此乐观，“幻觉问题”是大模型在落地金融应用时被讨论的最多的一个问题。所谓大模型幻觉，即大模型产生的看似合理却缺乏事实依据的内容。

目前学界对幻觉的定义与之前的研究一致[1], 将其定性为生成的内容**不遵循原文（Faithfulness）**或者**不符合事实（Factualness）**。在Lei Huang的研究中[2]，将幻觉定为两大类：**事实性幻觉**和**忠实性幻觉**。

事实性幻觉强调生成的内容与可验证的现实世界事实之间**存在差异**，通常表现为事实不一致或捏造。比如用户向大模型询问“2023年中国贷款市场报价利率1年期LPR是多少？”这样的事实问题，正常来说大模型会回答“不知道”或者一个正确的数字如3.45%，但有时大模型会回复一些其他数字，与现实不符，这就属于事实性幻觉。

忠实性幻觉强调从用户的角度解决不一致性问题，通常表现为**指令不一致**、**上下文不一致**。比如，用户给大模型一份解析好的合同文件信息，让大模型分析“这份合同首尾部主体名称是否相同？”，大模型若答非所问如回答“这份合同无知识产权条款”，则为指令不一致的忠实性幻觉；若回复的过程中，并不参考我们给出的合同信息或者不参考自己已产生过的分析，反而参考一些虚假的上文，则为上下文不一致的忠实性幻觉。

1. **大模型幻觉的来源**

幻觉的产生有多方面的原因，我们在研发大模型时，会涉及到**预训练**、**微调**、**对齐**等多个阶段，每个阶段都会是幻觉的引入来源，我们将流程整体划分成四个阶段：**预训练阶段**，**SFT训练阶段**， **RLHF 训练阶段**和**生成推理阶段**。

在预训练阶段中，**有缺陷的数据**和**低效的数据利用方式**会导致大模型缺乏相关知识或内化错误知识。在预训练阶段，LLMs会从大量训练数据中积累大量知识，然后将其存储在模型参数中。当被要求回答问题或完成任务时，LLMs当被要求回答问题或完成任务时，如果**缺乏相关知识**或内化了训练语料中的**错误知识**，LLMs就会产生幻觉。例如，LLMs有时会将虚假的相关性（如位置接近或高度共现的关联）误解为事实知识，即LLM的幻觉与训练数据的分布之间存在很强的相关性。例如，LLM偏向于肯定测试样本，人类生成的语料中也存在幻觉（可反映为过时的、双重的或捏造的表达）。因此，LLMs很容易复制甚至放大这种幻觉行为。

在SFT训练阶段中，**SFT指令数据超出LLM的知识边界**且LLM并**不能准确识别其知识边界**会导致LLM表现出**过度自信**编造毫无根据的答案。由于大模型在预训练期间已经确定了其固有的能力边界，SFT 利用高质量的指令及其对应的回答来增强 LLMs 遵循用户指令的能力，从而在这一过程解锁他们所后天培养的能力。然而，随着 LLM 能力的扩展，出现了一个重大挑战：LLM 的内在能力与注释数据中描述的能力之间可能存在不一致。当对齐数据的要求超出了这些预先定义的能力边界时，LLM就会被训练制作超出自身知识边界的内容，从而加大产生幻觉的风险。与此同时语言模型有时会高估自己的能力，Kadavath 等人[19]进行的实验证明语言模型能够评估自己回答的正确性（自我评估），并确定自己是否知道某个问题的答案。然而，对于非常大的 LLMs 来说，**正确答案和错误答案的分布熵可能是相似的**，这表明 LLMs 在生成错误答案时与生成正确答案时**同样自信**。通过评估流行LLMs识别无法回答或不可知问题的能力显示，即使是最先进的LLM--GPT4，与人类相比也存在明显的性能差距。注意到准确率与信心之间存在相关性，但这种信心往往超过LLM的实际能力，也就是**过度自信**。

在 RLHF 训练阶段中，在**没有先决知识**的条件下进行**错误的对齐过程**会导致LLM表现出**谄媚**生成偏向用户观点的答案。LLM 通常会在预训练后进行对齐过程，它们会在经过精心策划的指令范例上接受进一步训练，以使它们的反应与人类偏好保持一致。然而，如果 LLM 在预培训阶段没有获得相关的先决知识，那么在对指令进行培训时，这实际上是一个错误的对齐过程，会鼓励 LLM 产生幻觉**[20][21]**。

在生成推理阶段中，解码方式的缺陷会导致**LLM产生事实性错误**甚至“**滚雪球**”现象。LLMs 在生成极具创造性和多样性的内容方面表现出了非凡的能力，而这种能力在很大程度上取决于**随机性**在其解码策略中的关键作用。然而解码策略的随机性所带来的多样性是有代价的，因为取样随机性的升高会导致标记概率分布更加均匀，从而增加**从分布尾部取样频率较低的标记**的可能性，导致增加产生幻觉的风险（Aksi-tov 等人，2023 年）。同时在解码阶段，由于语言模型在其注意力机制中经常表现出**局部聚焦**，优先考虑附近的单词，导致**上下文注意力明显不足**直接导致忠实性幻觉，这种在长文本生成的任务中尤其明显，甚至会出现**结构性遗忘**的现象。

1. **缓解幻觉的方法**
2. 、**缓解事实性幻觉**

**1、事实性数据增强**

对于缓解事实性幻觉，最简单且直观的方法就是使用更优质的数据集。然而，随着预训练数据集的不断扩大，人工整理成为一项挑战。鉴于学术或专业领域的数据通常在事实方面是准确的，因此收集专业领域的数据成为一项主要策略。著名的例子包括 **Pile**[23]和 "**textbook-like** "数据源[24]。此外，在Llama 2[25]的工作中，他们证明了在预训练阶段进行**向上采样事实数据**可有效提高 LLM 的事实性，从而减轻幻觉。

1. **Honesty-oriented SFT**

为了减少SFT阶段过度自信的情况，这需要大模型具备自我知识边界认知能力（**拒识能力**）。在**MOSS**的工作中**[26]**，他们提出了**honesty-oriented SFT**，即在sft数据中添加一些**“Sorry，I don't know”**的数据。这种方法在一定程度上减少了过度自信的情况，但是仍然面临着模型**过度拒识**的风险。

**3、RHLF**

在RLHF阶段缓解幻觉的简单思路就是单独再设计一个针对幻觉的reward score，并直接在RLHF阶段优化。然而，由于在**乐于助人**和**诚实**两个对齐目标之间的**权衡失衡**，经过 RL 调整的 LLM 可能会表现出**过度保守**，在缓解了幻觉的同时也有削减大模型的能力的风险[6]。

**4、对比解码（Contrastive decoding）**

Li等人发现较大LM的失败在较小LM中更为普遍，而这种差异恰恰指示了应该偏好哪些文本，并提出了一种基于搜索的新方法--**对比解码（CD）**[9]，通过专家LM和业余LM对token的**预测概率差值**作为决策依据，更倾向于选择差值大的token。通过这种解码方式，可以生成**流畅且词性多样的文本**，同时不影响连贯性，表现显著优于核采样。

但是在 CD 中选择业余模型的主要标准由模型大小决定，然而大小并不一定会抑制业余模型学习事实知识。此外，"一刀切 "的业余模型对于对比不同复杂程度数据集的不同事实知识水平可能不是最佳选择。

**5、DoLa：对比层解码提高了大型语言模型的真实性**

此时，一些研究[10][11]认为，**LLM 的激活空间包含与事实性相关的可解释结构**。基于这一个理论，Chuang 等人从事实知识存储的角度出发，深入研究了如何提高 LLM解码过程的事实性。他们利用转换器 LLM 中事实知识的分层编码，指出低层信息被捕获在较早的层中，而语义信息被捕获在较晚的层中。他们从关于CD的研究中汲取灵感，引入了 DoLa。

DoLa创新性地利用**LLMs中的事实知识通常被证明会被本地化到特定的转换层**这一事实，通过对比投射到词汇空间的后层与前层的对数差异来获得下一个词的分布，提出了一种简单的解码策略来减少预训练 LLMs 的**事实性幻觉**[4]。

这是一种动态选择和对比不同层逻辑信息以完善事实性去编码的策略。DoLa 重视来自高层的知识优势，而淡化来自低层的知识优势，从而展示了其使 LLM 更符合事实的潜力，从而减少了幻觉。相较于CD，DoLa具有更**强适应性**，使其能够迎合标记和语境的复杂性。例如，简单的语境可能只需要早期层，而复杂的语境可能需要中间层或更高层。要实现这一点，CD 需要训练多个较小的 LM，并产生较高的计算成本。相比之下，DoLa 只需要一次前向传递和高效的早期退出，将延迟时间从 ×1.01 降到 ×1.08[4]。

**（二）、缓解忠实性幻觉**

**1、思维链（COT）**

对于忠实性幻觉，早期主要是通过将模型的结果与可验证的知识联系起来，**增强知识边缘的召回能力**，从而提供更强大的防御能力，防止产生忠**实性幻觉**。通常情况下，召回知识的最直接方法是通过 "**思维链**"（Chain-of-Thought）提示让 LLMs 进行推理。Zhong等人认为，只需应用CoT就能提高知识召回率，这大大提高了多跳设置下编辑事实的性能[13]。此外，在COT的基础上，还衍生出了诸如CoT-SC[14]等工作，这些都被证明了在忠实性幻觉缓解上的潜力。但是思维链的局限性也十分显著，首先，思维链必须在模型规模足够大时才能涌现。在 Jason Wei 等的研究[15]中，PaLM 在扩展到 540B 参数时，与思维链提示结合，才表现出了先进的性能。一些小规模模型，思维链并没有太大的影响，能力提升也不会很大。将思维链置入推理过程，会拆解更多的步骤，并用到更多的计算资源，造成**资源浪费**。

**2、上下文预训练（In-Context Pretraining）**

**上下文预训练**[27]是一种在预训练阶段改善幻觉表现的在相关文档序列上训练 LLM，通过简单地改变文档排序，使每个上下文包含相关文档，并直接应用现有的预训练流程，该方法旨在**最大化文本窗口内的相似性**，明确鼓励 LLMs **跨文档边界进行推理**，增强了LLM的上下文一致性，有效缓解了忠实性幻觉。

**3、检索增强（RAG）**

**检索-增强生成**（RAG），在生成过程中以从外部知识源检索到的相关文档为条件，将 LLMs 置于基础之上。通常情况下，RAG 遵循检索-阅读-生成的流水线，相关的上下文文档首先由检索器从外部来源检索，然后由生成器根据输入文本和检索文档生成所需的输出。在此基础上，衍生出了多种减轻幻觉的方法如**迭代检索**和**事后检索**等，进一步降低了产生的幻觉的可能。

**4、自然语言推理链（CoNLI）**

Lei 等人（2023 年）探索了一种通用的后编辑框架[18]，该框架使用**自然语言推理链（CoNLI）**进行幻觉检测，并通过后期编辑以减少推断过程中的忠实性幻觉。这种方法首先在句子和实体层面检测幻觉，然后利用检测反馈完善生成的响应，实际表现很不错。

**5、知识约束解码（KCD）**

Choi等人在KCTS的工作中引入了**知识约束解码（KCD）**[12]，该方法在**冻结 LM** 的基础上应用辅助知识分类器来检测幻觉，并使用其知识基础得分来指导解码过程。通过在解码过程中加入这些分类器，旨在约束生成的文本，确保其忠实于参考知识。

该方法经过训练**可预测幻觉Token的起始点**，并能将序列级定义的知识分类器有效地调整到**Token级**，能够有效的解决上下文不一致导致的忠实性幻觉。

1. **未来展望**

就大模型幻觉缓解，目前而言并没有较为统一的方法，除了上述针对幻觉产生原因进行缓解的方法，我们也留意到诸如如**引入外部知识库**、**专家模型**、**多智体辩论、引入人在循环中**等方式，他们很多已经在实际项目中应用，给我们解决大模型幻觉多了一些启发。

未来关于幻觉问题的解决一定是多维度的，如何结合多种方法，平衡大模型的创造性与可靠性，仍然是一个值得挑战的问题。

参考文献

1. Survey of hallucination in natural language generation
2. Large language models cannot self-correct reasoning
3. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response mode
4. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models
5. Training language models to follow instructions with human feedback
6. Siren’s Song in the AI Ocean:A Survey on Hallucination in Large Language Models
7. A Survey on Hallucination in Large Language Models:Principles, Taxonomy, Challenges, and Open Questions
8. Factuality enhanced language models for open-ended text generation
9. Contrastive decoding: Open-ended text generation as optimization.
10. Discovering latent knowledge in language models without supervision.
11. Relative representations enable zero-shot latent space communication.
12. KCTS: Knowledge-Constrained Tree Search Decoding with Token-Level Hallucination Detection
13. Mquake: Assessing knowledge editing in language models via multi-hop questions.
14. SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS
15. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
16. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding
17. KL-Divergence Guided Temperature Sampling
18. Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations
19. Language models (mostly) know what they know
20. Reinforcement learning for language models
21. Reinforcement learning from human feedback: Progress and challenges.
22. Characterizing Attribution and Fluency Tradeoffs for Retrieval-Augmented Large Language Models
23. The Pile: An 800GB Dataset of Diverse Text for Language Modeling
24. Textbooks Are All You Need
25. Llama 2: Open Foundation and Fine-Tuned Chat Models
26. MOSS: Training Conversational Language Models from Synthetic Data
27. IN-CONTEXT PRETRAINING: LANGUAGE MODELING BEYOND DOCUMENT BOUNDARIES