

# 大模型技术及开发

## 大模型幻觉及缓解

主讲人：陈小军

时间：2025.5.30

01

# 什么是AI幻觉

---

2025

# 什么是AI幻觉

## 案例1



今天被 Deepseek 耍得团团转，给我推荐了一个阿布扎比的本地市场在迪拜，道歉后，给推荐的一个特色商场根本不存在。好，给他第三次机会，推荐一个直面滨海大道与海岸线的咖啡馆，实际是这样的🤔🤔

收起

- 设计概述:
  - 项目背景: 本项目旨在打造一个集休闲、娱乐、餐饮于一体的综合性商业空间。
  - 设计目标: 打造一个具有现代感、时尚感、舒适感的商业空间。
  - 设计理念: 以人为本，注重细节，追求品质。
- 项目概况:
  - 项目名称: 阿布扎比滨海大道与海岸线咖啡馆
  - 项目地址: 阿布扎比滨海大道与海岸线
  - 项目规模: 约 1000 平方米
- 设计思路:
  - 设计理念: 以人为本，注重细节，追求品质。
  - 设计思路: 通过空间布局、色彩搭配、材质选择等手段，打造一个具有现代感、时尚感、舒适感的商业空间。
- 项目成果:
  - 项目成果: 项目已于 2023 年 10 月 1 日正式开业，受到广大顾客的热烈欢迎。

Abu Dhabi Mall 项目设计说明

1. 项目背景

名称	地址	建筑面积	设计单位
Abu Dhabi Mall	阿布扎比	100000	阿布扎比设计局
Abu Dhabi Mall	阿布扎比	100000	阿布扎比设计局
Abu Dhabi Mall	阿布扎比	100000	阿布扎比设计局



## 什么是AI幻觉

# 案例2

**Whisper: OpenAI的自动语音识别 (ASR) 系统**

**行业应用: 医疗系统中, 将患者与医生的对话问诊过程音频, 转写为文字病例, 有超过30000名临床医生和40个医疗系统使用**

**发现: 100多个小时的Whisper转录样本, 其中约有一半内容存在幻觉**

原音频: “嗯, 她的父亲再婚后不久就去世了”

转录文本: “没关系。只是太敏感了, 不方便透露。她确实在65岁时去世了”

**结果: 2.6W多份自动转录病例中, 几乎每本都存在瞎编和幻觉问题, 对患者健康和医疗系统产生严重负面影响**

## ▶ 什么是AI幻觉

**学术概念：**指模型生成与事实不符、逻辑断裂或脱离上下文的内容，本质是统计概率驱动的“合理猜测”

**总结：**一本正经地胡说八道

**事实性幻觉：**指模型生成的内容与可验证的现实世界事实不一致

**忠实性幻觉：**指模型生成的内容与用户的指令或上下文不一致

**提问：**糖尿病患者可以通过吃蜂蜜代替糖吗？

### 回答

### 分析

#### 事实性幻觉

是的，蜂蜜是天然的，可以帮助糖尿病患者稳定血糖水平。

**错误：**蜂蜜虽然是天然食品，但仍然含有大量果糖和葡萄糖，会升高血糖水平，不适合糖尿病患者代替糖使用。

#### 忠实性幻觉

蜂蜜富含维生素和矿物质，对提高免疫力很有帮助，因此是一种健康的食品。

**偏题：**回答内容虽无事实错误，但与提问“糖尿病患者是否可以用蜂蜜代替糖”无关，未忠实于用户意图。



## ▶ AI为什么会产生幻觉

- **数据偏差：**训练数据中的错误或片面性被模型放大（如医学领域过时论文导致错误结论）
- **泛化困境：**模型难以处理训练集外的复杂场景（如南极冰层融化对非洲农业的影响预测）
- **知识固化：**模型过度依赖参数化记忆，缺乏动态更新能力（如2023年后的事件完全虚构）
- **意图误解：**用户提问模糊时，模型易“自由发挥”（如“介绍深度学习”可能偏离实际需求）



## AI幻觉的潜在风险

- **信息污染风险：** 由于DeepSeek的低门槛和普及度高，大量AI生成内容涌入中文互联网，加剧了虚假信息传播的“雪球效应”，甚至污染下一代模型训练数据
- **信任危机：** 普通用户难以辨别AI内容的真实性，可能对医疗建议、法律咨询等专业场景的可靠性产生长期怀疑
- **控制欠缺：** DeepSeek的**对齐**工作较其他闭源大模型有所欠缺，其开源特性也允许使用者随意使用，可能会成为恶意行为的工具
- **安全漏洞：** 若错误信息被用于自动化系统（如金融分析、工业控制），可能引发连锁反应

.....

## ▶ AI幻觉的创造力价值

### ➤ 科学发现：从“错误”到突破的范式跃迁

- 蛋白质设计：大卫·贝克团队利用AI“错误折叠”启发新型蛋白质结构，获2024诺贝尔化学奖。
- 认为AI幻觉是“从零开始设计蛋白质”的关键
- 发表的多篇论文都将“幻觉”纳入标题当中

Article | Published: 01 December 2021

### **De novo protein design by deep network hallucination**

[Ivan Anishchenko](#), [Samuel J. Pellock](#), [Tamuka M. Chidyausiku](#), [Theresa A. Ramelot](#), [Sergey Ovchinnikov](#), [Jingzhou Hao](#), [Khushboo Bafna](#), [Christoffer Norn](#), [Alex Kang](#), [Asim K. Bera](#), [Frank DiMaio](#), [Lauren Carter](#), [Cameron M. Chow](#), [Gaetano T. Montelione](#) & [David Baker](#) ✉





## ▶ AI幻觉的创造力价值

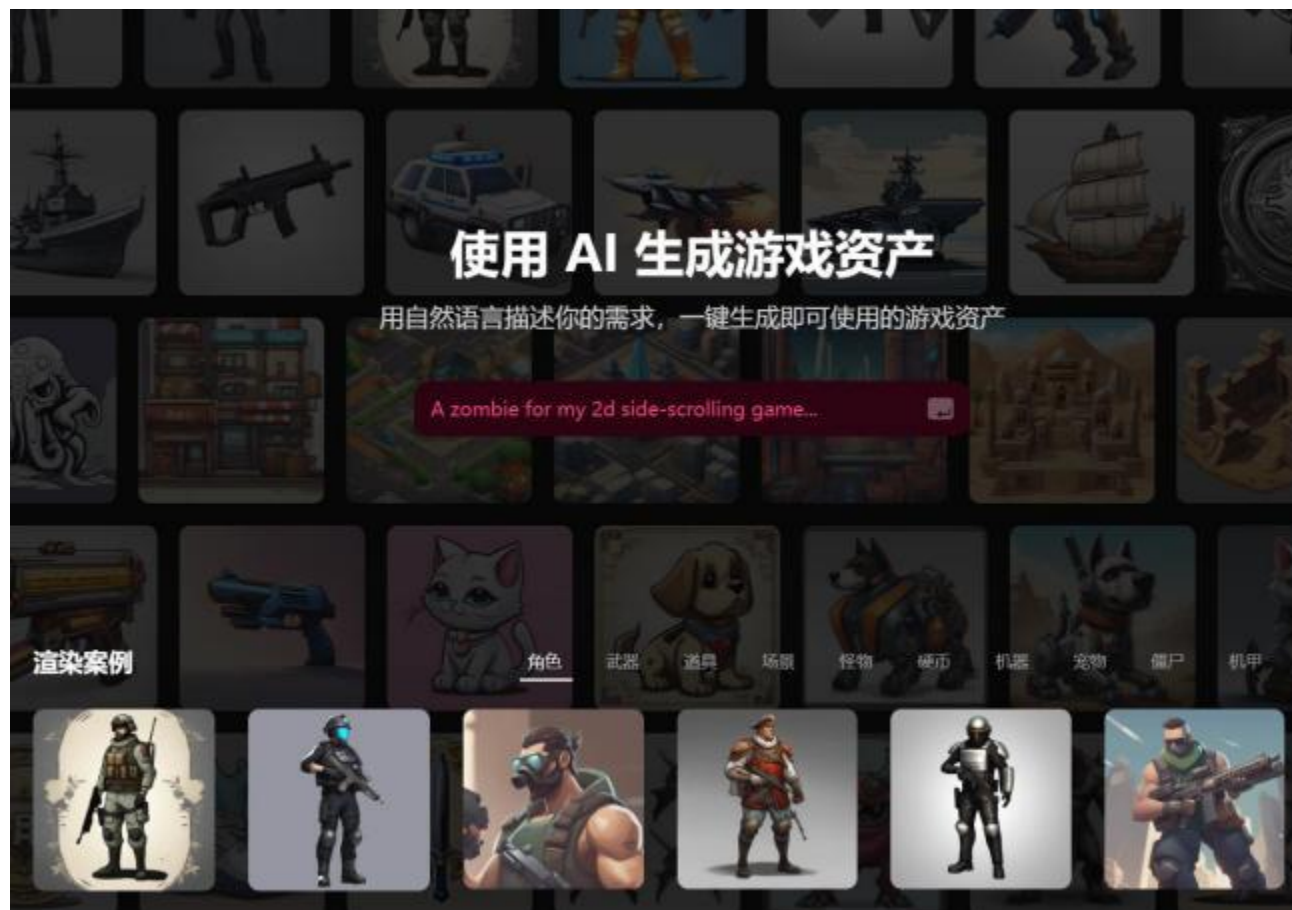
### ➤ 文艺与设计：突破人类思维定式的“超现实引擎”



## ▶ AI幻觉的创造力价值

### ➤ 娱乐与游戏：创造新的视觉和听觉体验

- AI生成的虚拟环境和角色设计为游戏开发人员提供了无限的可能性，增强了玩家的沉浸感和探索欲
- AI幻觉还被用于生成故事、对话和诗歌，为游戏和文学创作提供灵感



## AI幻觉的创造力价值

### ➤ 技术创新：从“缺陷”到方法论的转化

- DeepMind团队发现，AI在图像分割任务中产生的“超现实边界”虽不符合真实场景，却意外提升了自动驾驶系统对极端天气（如浓雾、暴雨）的识别精度。

### ➤ 新型科研范式：科学界正构建“AI幻觉-实验验证-理论重构”的三阶段研究流程。

- 加州理工学院团队通过AI生成虚构导管设计，最终通过新型人工智能技术优化后的新设计，在实验中证实将向上游游动的细菌数量减少了100倍，形成“疯狂创意→理性筛选”的创新闭环。



02

# 幻觉的分类

---

2025

# AI幻觉的分类

## 1. 事实性问题 (Factuality Issues):

- **事实性错误 (Factual Errors):** 模型回答与客观事实不一致。
  - 例子: “加拿大最大的城市是渥太华。” (实际是多伦多, 渥太华是首都)
- **事实性虚构 (Factual Fabrication):** 模型回答在真实世界无法验证, 纯属编造。
  - 例子: “独角兽起源于英国。”

## 2. 忠诚度问题 (Faithfulness Issues):

- **违背指令 (Violating Instructions):** 模型回答没有遵循用户的明确指令。
  - 例子: 用户要求“用一句话总结”, 模型输出了三段话。
- **违背上下文 (Inconsistent with Context):** 模型回答与提供的上下文信息不一致或矛盾 (在RAG场景下尤为突出)。
  - 例子: 上下文: “报告指出, 公司去年营收增长了15%。” 模型回答: “根据报告, 公司去年营收有所下降。”

## 3. 自我矛盾 (Self-Contradiction):

- 模型在同一段回答中, 前后内容出现逻辑矛盾。
  - 例子: “我非常喜欢阴天, 因为阳光明媚的天气总能让我心情愉悦。”





## 事实性幻觉评测

### ➤ 常识错误

提问	正确答案	DeepSeekV3	DeepSeek R1	Qianwen2.5-Max	豆包
“天上百鸟朝凤凰，朝凤凰，地上葵花向 <b>太阳</b> ，向太阳，哎， <b>北京</b> 城里的 <b>毛主席</b> ，我们永远跟您走，幸福万年长罗！”这段歌词出自哪首歌曲	这段歌词出自藏族民歌《北京有个金太阳》。	这段歌词出自歌曲《 <b>北京的金山上</b> 》。	歌曲**《 <b>阿佤人民唱新歌</b> 》**（1972年），表达少数民族对领袖的拥护。	这段歌词出自歌曲《 <b>东方红</b> 》。	“天上百鸟朝凤凰，朝凤凰，地上葵花向太阳，向太阳，哎，北京城里的毛主席，我们永远跟您走，幸福万年长罗”出自《 <b>北京的金山上</b> 》。



# 事实性幻觉评测

## ➤ 逻辑陷阱

**提问：**为什么一向见钱眼开的小明仍然会被金钱蒙住双眼



已深度思考 (用时 19 秒)

小明虽然一向见钱眼开，但过度贪婪反而导致其判断力丧失，最终被金钱蒙蔽。

嗯，用户问的是“为什么一向见钱眼开的小明仍然会被金钱蒙住双眼？”  
个成语的意思。“见钱眼开”通常指看到钱就眼睛发亮，形容人贪财，  
为贪图金钱而失去了判断力，做出错误决定。

用户的问题在于，既然小明本来就是见钱眼开的人，为什么还会被  
已经很贪财了，应该更懂得如何利用金钱，而不是被其蒙蔽。可能的  
贪婪反而导致他失去了理性，无法正确评估风险，或者被眼前的利益

接下来，我需要验证这个解释是否合理。

骗，这时候他的贪婪让他无法冷静分析，

另外，可能要考虑“见钱眼开”和“被金钱蒙蔽”  
而导致的盲目。也就是说，小明的贪财和

还需要考虑是否有其他因素，比如小明可

这时候，即使他平时贪财，但缺乏应对能

总结起来，可能的答案是：虽然小明平时  
择。或者，他的贪财性格使他在面对金钱

### 4. 机会成本的误判

被金钱蒙眼往往伴随着：

- 忽视时间价值（如为加班费牺牲健康）
- 错失人际关系等无形财富积累
- 透支信用资本导致发展受限

这种现象本质是“金钱认知失调”：当物质追求与个人能力、社会环境产生结构性矛盾时，越执着于金钱反而  
越容易被其反噬。如同物理学的逃逸速度原理，当对金钱的渴望超过某个临界值，理性判断就会脱离正常轨  
道。

### 1. 贪婪的陷阱

见钱眼开体现对金钱的渴望，但过度贪婪会引发非理性行为。例如：

- 追逐短期暴利时忽视风险（如参与传销、赌博）
- 为快速获利突破道德底线（如受贿、诈骗）
- 陷入“损失厌恶”心理，为挽回损失投入更多沉没成本

## 事实性幻觉评测

### ➤ 虚构事件

提问	正确答案	DeepSeekV3	DeepSeek R1	Qianwen2.5-Max	豆包
水浒传中李逵为什么要大闹五台山?	李逵并没有大闹五台山,事实上,《水浒传》中大闹五台山的是花和尚鲁智深。	水浒传中李逵大闹五台山是因为他喝醉了酒,误闯了寺庙。	误以为宋江强抢民女(实为他人冒充),怒闯五台山讨说法。	李逵大闹五台山是因为他误以为鲁智深在寺庙中受欺负,于是出手相助,引发混乱。	水浒传中大闹五台山的是鲁智深,不是李逵。鲁智深大闹五台山是因为他出家后难以忍受寺里的清规戒律,又喝酒吃肉,还打伤了寺里的僧人等。

03

# 幻觉的来源

---

2025



# AI幻觉的来源

## ➤ 来自数据的幻觉 (Data-induced Hallucinations)

### • 数据源缺陷:

- **知识边界/过时:** 模型训练数据截止于某个时间点，对新知识未知。
  - 来源例子: 模型训练数据截止到2022年，无法准确回答2024年发生的事件。
- **低质/错误数据:** 训练数据本身包含错误、噪声或偏见。
  - 来源例子: 训练数据中包含“地球是平的”等错误论述。
- **数据稀疏/领域知识匮乏:** 某些领域或主题的数据量不足。
  - 来源例子: 模型在高度专业的医学或法律问题上更容易幻觉。

### • 数据利用缺陷:

- **上下文理解不佳 (召回取巧/失败):** 模型未能充分理解或利用提供的长上下文，可能只关注开头结尾，或错误关联信息。
  - 来源例子: 在处理一篇长篇报告摘要时，模型可能遗漏中间的关键数据点。





## ▶ AI幻觉的来源

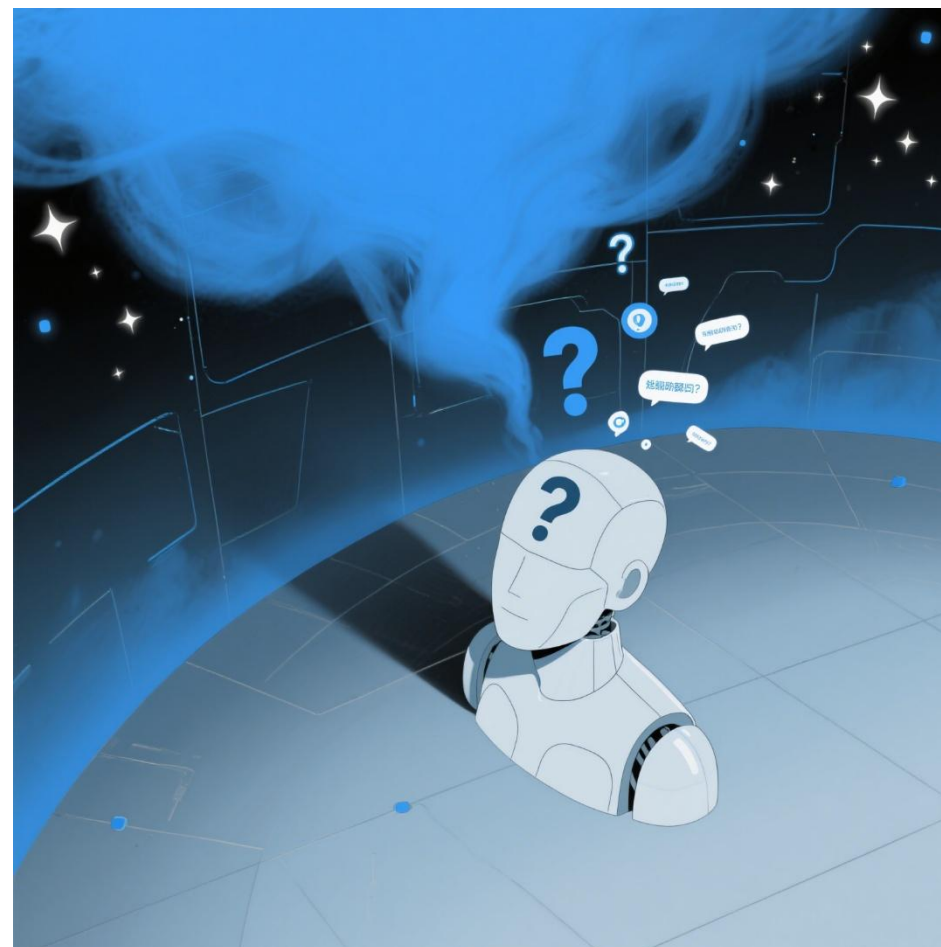
### ➤ 来自训练的幻觉 (Training-induced Hallucinations)

#### • 训练架构/策略:

- **注意力机制的缺陷:** 可能导致对输入信息的不当加权。
- **预训练与微调目标不一致:** 导致模型在特定任务上表现不佳。

#### • 偏好对齐问题:

- **过度修正/对齐税:** RLHF等对齐过程可能为了“乐于助人”或“无害”而牺牲部分事实性，在不确定时倾向于编造而非承认不知。
  - **来源例子:** 为了显得合作，模型可能会猜测一个它并不知道的答案。



## ▶ AI幻觉的来源

### ➤ 来自推理的幻觉 (Inference-induced Hallucinations)

#### • 解码策略:

- **随机解码的固有问题:** 贪心搜索可能导致重复、平庸的回答, 而采样方法 (如Temperature、Top-k, Top-p) 虽能增加多样性, 但也可能引入不准确的内容。
  - **来源例子:** 较高的Temperature设置可能让模型生成更“有创意”但也更不靠谱的回答。

#### • 解码过程的信息损失/错误累积:

- 在生成长文本时, 早期微小的偏差或错误可能被放大, 导致后续内容偏离事实。



## ▶ 推理能力与幻觉率存在双向作用机制

### 🧠推理能力增强 → 幻觉率降低

DeepSeek V3: 提问 → 回答

DeepSeek R1: 提问 → 思维链 → 回答

#### • ☒ 逻辑准确性与错误减少

强推理能力帮助避免逻辑错误。例如在数学题中，多步推理让模型更可能得出正确结论而非臆测。

#### • ☒ 上下文理解与信息关联

推理能力使模型更精准把握上下文，减少断章取义。例如在问答中能排除干扰项，降低幻觉风险。

### ⚠️ 推理能力增强 → 幻觉率上升

#### • ✖ 逻辑过度外推

模型在已有事实间构建“超合理”但虚构的联系  
如：1990年发明A技术 → 自动推测1995年获诺奖（虚构）

#### • ✖ 认知置信度错位

推理能力强的模型更倾向生成“自信错误”答案，而非坦诚回答“不知道”

#### • ✖ 错误前提下的正确推理

基于错误设定，仍进行逻辑严密但方向错误的推理



# 推理与幻觉的关系测试

评价方法	DeepSeek R1	DeepSeek V3	GPT-o1	GPT-4o
Vectara HHEM 2.1	14.3%	<b>3.9%</b>	2.4%	<b>1.5%</b>
Google FACTS (GPT-4o + Claude 3.5 Sonnet)	4.37%	<b>2.99%</b>	<b>1.00%</b>	1.49%
Google FACTS (GPT-4o + Gemini 1.5 Pro)	3.09%	<b>1.99%</b>	<b>0.90%</b>	1.39%
Google FACTS (Claude 3.5 Sonnet + Gemini 1.5 Pro)	3.89%	<b>2.69%</b>	<b>1.39%</b>	1.89%



# 幻觉高发场景

场景类别	具体场景	示例	风险等级	防护建议
知识边界模糊	开放域生成	续写未完结的经典文学作品	高	添加创作范围限制+事实性标注
	未来事件预测	预测2030年科技突破细节	极高	声明预测性质+概率分布呈现
复杂推理	多跳推理任务	追溯企业高管早期职业轨迹	高	分步验证+外部知识库检索
	数学证明延伸	要求证明未解决的数学猜想	极高	中断机制+当前研究进展说明
技术性诱发	长文本生成	小说连续章节生成	中	阶段一致性检查+人物属性维护
	多轮对话	复杂业务流程咨询	高	对话历史摘要+关键事实复核
数据引用	矛盾数据源	不同版本的实验数据引用	中	矛盾点对比+最新研究成果优先
情感驱动	安慰性回应	重症患者寻求治疗方案建议	极高	情感剥离响应+理论应用提示
特殊领域	医疗诊断	根据症状描述提供诊断建议	极高	明确非专业建议+医疗数据库
	法律咨询	解释特定法条适用范围	高	司法辖区限定+法律条文引用
	金融预测	给出具体股票买卖建议	极高	风险提示+历史回报率说明



04

# 幻觉的检测

---

2025

# AI幻觉检测的主要方法

## 1. 事实性检测:

- **外部知识库/搜索引擎验证:** 将模型输出的关键信息点与权威数据源（如Google搜索、维基百科、专业数据库）进行比对。
- **模型不确定性评估:**
  - **需要模型参数:** 分析输出Token的概率分布、熵值等。
  - **无需模型参数:** 多次生成看答案一致性（Self-CheckGPT），或利用N-gram重叠度等指标。

## 2. 忠诚度检测 (常用于RAG):

- **信息重叠度:**
  - N-gram、实体、关系元组在模型回答和源上下文之间的重叠程度。
- **分类任务:**
  - **NLI (自然语言推理):** 训练一个模型判断“回答”是否被“上下文”所支持/蕴含 (Entailment)。
  - **QA任务:** 从模型回答中抽取事实，构建QA对，然后在上下文中寻找答案，判断是否一致。
- **大模型Prompt自评:** 指示另一个（或同一个）大模型评估当前回答是否忠实于给定上下文。



05

# 幻觉的缓解

---

2025

## ▶ 数据与训练层面的缓解 (主要面向开发者)

### 数据层面

#### 提升数据质量:

- **清洗数据:** 例如, 在训练一个客服聊天机器人时, 删除其训练数据中的重复条目、修正错误标注的数据, 或剔除无关信息。
- **去除偏见:** 例如, 确保用于贷款申请审批模型的训练数据在不同人群中具有均衡的代表性, 以避免产生带有偏见的预测结果。
- **增加高质量和多样化的训练数据:** 例如, 为一个图像识别模型添加在不同光照条件、不同角度拍摄的, 以及包含多种物体变体的图像。

#### 知识增强:

- **引入知识图谱:** 例如, 通过引入医学知识图谱来增强医疗诊断AI的训练数据, 以提高其对疾病、症状和治疗之间关系的理解。
- **事实数据库进行数据增强:** 例如, 使用历史事件数据库来丰富语言模型的训练数据, 使其能够更准确地回答与历史相关的问题。





# 数据与训练层面的缓解 (主要面向开发者)

## 训练层面

### •改进训练策略:

- **课程学习 (Curriculum Learning):** 例如, 训练一个数学解题AI时, 首先让其学习简单的算术问题, 然后逐步增加难度至代数和微积分, 模仿人类的学习过程。
- **多任务学习 (Multi-Task Learning):** 例如, 训练单个模型同时执行多种相关的自然语言处理任务, 如情感分析、命名实体识别和机器翻译, 使模型能够学习到更具泛化性的表示。

### •优化对齐技术:

- **改进SFT (Supervised Fine-Tuning):** 在大型语言模型的SFT阶段, 使用更精心策划和更高质量的指令-回应数据对, 以使其输出更符合期望的行为。
- **改进RLHF (Reinforcement Learning from Human Feedback):** 在RLHF中开发更精细的奖励模型, 不仅评估回应的有用性, 还评估其事实准确性和无害性。
- **引入过程监督 (Process Supervision):** 不仅仅因为模型最终给出了数学问题的正确答案而奖励它, 也对它推理过程中的每一个正确中间步骤给予奖励。
- **事实性奖励信号 (Factual Reward Signals):** 设计一个奖励函数, 专门惩罚模型生成与可信知识库相矛盾的陈述。

### •模型编辑 (Model Editing):

- **直接修改模型参数以纠正或更新特定知识点:** 例如, 如果一个语言模型错误地陈述了某个国家的首都, 可以使用模型编辑技术直接调整其内部权重, 使其提供正确的首都信息, 而无需完全重新训练。



## ► 推理层面的缓解 (主要面向开发者)

### 优化解码策略

- **事实核采样 (Factual Nucleus Sampling):** 例如，当语言模型生成关于某一历史事件的文本时，事实核采样会优先选择与既定历史事实一致的词语和短语，而不是允许更多推测性或虚构性的延续。
- **对比解码 (Contrastive Decoding):** 例如，通过对比大型通用模型和小型专业模型（如一个在医学文本上微调过的模型）的输出来生成回应。最终输出会倾向于两者都支持的信息，或者使用小型模型来优化大型模型的输出。另一个例子：使用正面提示（“描述一只猫”）和负面提示（“不要提及狗”）来引导生成。

### 后处理与验证

- **CoVe (Chain-of-Verification):**
  - **例子:** 用户提问：“阿司匹林有哪些副作用？”
    - 模型生成初步答案：“阿司匹林可能引起胃部不适。”
    - 模型规划验证问题：“根据医学文献，阿司匹林的常见副作用有哪些？”、“阿司匹林是否会导致胃部不适？”
    - 模型执行验证（例如，查询医学数据库或可靠的网页来源）。
    - 模型综合结果生成最终答案：“根据[来源X]等医学资料，阿司匹林的常见副作用包括胃部不适、胃灼热和容易瘀伤。因此，阿司匹林确实可能引起胃部不适。”

## ▶ 普通用户应对AI幻觉的三种方式

### ➤ 联网搜索

“ 请使用联网功能 ”、联网功能选项

大模型	通用性测试幻觉率	事实性测试幻觉率
DeepSeekV3	2%→0% (下降2%)	29.67%→24.67% (下降5%)
DeepSeekR1	3%→0% (下降3%)	22.33%→19% (下降3%)

注：黑色为未开启联网搜索，红色为开启联网楼

## ▶ 普通用户应对AI幻觉的三种方式

### ➤ 双AI验证/大模型协作

例如，利用DeepSeek生成答案后，再应用其他大模型进行审查，相互监督，交叉验证

请判断以下内容的真实性：

在《水浒传》中，李逵大闹五台山的原因主要是为了救出被囚禁的宋江。

具体情节如下：

1. **宋江被囚**：宋江因误入五台山，被寺中的和尚误认为是强盗，将其囚禁。
2. **李逵得知消息**：李逵得知宋江被囚后，心急如焚，决定前往五台山救人。
3. **大闹五台山**：李逵性格鲁莽，到达五台山后，不顾一切地冲进寺庙，与和尚们发生冲突，最终成功救出宋江。李逵的行为反映了他对宋江的忠诚和义气，同时也展现了他冲动、鲁莽的性格特点。



Qwen2.5-Max

以上内容并不符合《水浒传》原著的情节，因此是**不真实的**。以下是具体分析：

#### 1. 宋江被囚五台山？

在《水浒传》中，并没有宋江被囚禁于五台山的情节。宋江确实曾因杀人而逃亡，但他的主要经历包括在清风寨、江州等地的冒险，以及后来被刺配江州等情节。五台山并不是宋江的重要活动地点。

五台山在《水浒传》中与鲁智深的故事密切相关。鲁智深因打死镇关西后出家，在五台山文殊院当和尚，但他因性格豪放、不守清规戒律，多次闹事，最终被迫离开五台山。五台山的情节主要是围绕鲁智深展开的，与宋江无关。

## ▶ 普通用户应对AI幻觉的三种方式

### ➤ 提示词工程

**知识边限定：通过时空维度约束降低虚构可能性（本质：约束大模型）**

1. **时间锚定法：**“基于2023年之前的公开学术文献，分步骤解释量子纠缠现象” → 规避未来时态虚构
2. **知识锚定法：**“基于《中国药典》回答，若信息不明确请注明“暂无可靠数据支持” → 限定权威来源
3. **领域限定符：**“作为临床医学专家，请列举FDA批准的5种糖尿病药物” → 添加专业身份限定
4. **置信度声明：**“如果存在不确定性，请用[推测]标签标注相关陈述” → 减少绝对化错误断言
5. **上下文提示：**“根据《2024全球能源转型报告》（国际能源署，2024年1月发布）显示：” 2030年光伏发电成本预计降至0.02美元/千瓦时，但储能技术突破仍是普及瓶颈。“请基于此数据，分析中国西部光伏基地发展的三个关键挑战，并标注每个挑战与原文结论的逻辑关联。 → 嵌入权威数据片段
6. **生成参数协同控制：**“请以temperature=0.3的严谨模式，列举2024年《柳叶刀》发表的传染病研究”

## ▶ 普通用户应对AI幻觉的三种方式

### ➤ 提示词工程

**对抗性提示：强制暴露推理脆弱点，用户可见潜在错误路径（本质：大模型自我审查）**

1. **植入反幻觉检测机制：** "请用以下格式回答： - 主要答案（严格基于公开可验证信息） - [反 事实检查] 部分（列出可能导致此答案错误的3种假设） "
2. **预设验证条件，迫使模型交叉检查信息：** "请先回答 "量子纠缠能否证明灵魂存在？ "，然后从以下角度验证答案的可靠性： 1. 物理学界主流观点； 2. 近五年相关论文数量； 3. 是否存在可重复实验证据。 "
3. **链式验证：** 请完成以下验证链： 1. 陈述观点： \_\_\_\_\_ 2. 列出支撑该观点的三个权威数据源  
3. 检查每个数据源是否存在矛盾信息 4. 最终结论（标注可信度等级）



# 普通用户应对AI幻觉的方式总结

## 1. 清晰、具体的提问 (Clear Prompts):

- 提供充足背景信息，避免模棱两可的问题。
- **技巧:** “假设你是一位[角色]，请根据[上下文信息]，用[特定风格]回答关于[明确主题]的问题。”

## 2. 批判性思维与交叉验证 (Critical Thinking & Cross-Verification):

- **不轻信:** 对模型的回答（尤其是关键信息）持怀疑态度。
- **核查:** 利用搜索引擎、权威网站、书籍等多种渠道进行事实核查。

## 3. 要求引用与溯源 (Request Citations):

- 如果系统支持（如某些RAG应用），要求模型提供信息来源。

## 4. 调整参数 (Parameter Tuning):

- **Temperature:** 调低温度（如0.1-0.3）通常能获得更稳定、事实性更强的回答；调高则更具创造性但也更易幻觉。

## 5. 分解复杂问题 (Decompose Complex Questions):

- 将一个复杂问题拆解成若干简单的小问题，逐步引导模型。

## 6. 利用特定提示技巧 (Prompting Techniques):

- **“逐步思考” (Think step-by-step):** 鼓励模型展示推理过程。
- **“请仅根据我提供的上下文回答”:** 在RAG场景中强调忠诚度。
- **“如果不知道，请说不知道”:** 鼓励模型承认知识边界。

06

# 抑制幻觉的相关技术

---

2025



## 应对AI幻觉的技术方案

- **检索增强生成 (RAG - Retrieval Augmented Generation):**
  - **核心:** 在生成回答前, 先从外部知识库中检索相关信息, 并将其作为上下文提供给LLM。
  - **优势:**
    - **提升事实性:** 基于检索到的实时、准确信息生成回答。
    - **可解释性/可溯源:** 可以追溯答案来源。
    - **知识更新便捷:** 只需更新外部知识库, 无需重新训练大模型。
  - **流程:** 用户提问 -> 检索器检索 -> 上下文注入 -> LLM生成。
- **外部知识库:** 可以是文档集合、数据库、知识图谱等。
- **精细训练:** 针对不同任务类型进行具体的微调或强化
- **评估工具:** 开发高效的自动化AI幻觉识别工具, 对生成内容进行及时验证

## 如何应对AI幻觉

### **数据为基 (Data-centric):**

- 高质量预训练数据、事实增强、偏见缓解。
- 构建和维护高质量的外部知识库 (用于RAG)。

### **模型为核 (Model-centric):**

- 改进模型架构 (如注意力机制)。
- 模型编辑技术。

### **训练为钥 (Training-centric):**

- 先进的对齐算法 (RLHF, DPO, PPO)。
- 指令微调、过程监督。

### **推理为辅 (Inference-centric):**

- 解码算法优化 (如对比解码、事实核采样)。
- 提示工程 (CoT, ToT, Self-Ask)。
- 后处理验证 (CoVe, Self-Reflection)。

### **检索为援 (Retrieval-centric):**

- RAG及其各种变体 (Advanced RAG, Modular RAG)。

### **评估为尺 (Evaluation-centric):**

- 开发更可靠的幻觉检测基准和方法。
- 持续监控和迭代。

### **人机协同 (Human-in-the-loop):**

- 用户反馈回路, 专家审核。



“

**AI幻觉像一面棱镜，既折射出技术的局限性，也投射出超越人类想象的可能。与其追求“绝对正确”，不如学会与AI的“想象力”共舞——因为最伟大的创新，往往诞生于理性与狂想的交界处。**

**——DeepSeek R1**



2025

谢谢大家

时间: 202X.X