Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 156-167
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151016

Hans汉斯

大语言模型幻觉研究综述

何金骅[1]，王洪俊[1,2]

[1]北京信息科技大学计算机学院，北京
[2]拓尔思信息技术股份有限公司，北京

收稿日期：2025年11月26日；录用日期：2025年12月31日；发布日期：2026年1月9日

摘　要

大语言模型作为当前人工智能研究的核心方向，已在各类下游任务中都展现出显著的性能优势。然而，大模型产生的幻觉问题已成为其在高可靠性场景中应用的瓶颈，引发了学术界和工业界广泛关注。本文对大模型幻觉进行全面系统的回顾，首先，系统阐释大语言模型及其幻觉的定义，构建大语言模型幻觉的分类体系，从多个维度分析大模型幻觉产生的原因；其次，对有监督微调，检索增强等消减大语言模型幻觉的方法进行了综述；最后，在分析现有方法局限性的基础上，对未来消减大语言模型幻觉的研究方向进行了展望。旨在为构建更可靠、更可信的大语言模型提供理论参考与实践指引。

关键词

大语言模型，幻觉，有监督微调，检索增强

A Survey of Hallucination in Large Language Models

Jinxing He[1], Hongjun Wang[1,2]

[1]College of Computer Science, Beijing Information Science and Technology University, Beijing
[2]TRS, Beijing

Received: November 26, 2025; accepted: December 31, 2025; published: January 9, 2026

Abstract

As a central focus of contemporary artificial intelligence research, Large Language Models (LLMs) have demonstrated significant performance advantages across a multitude of downstream tasks. However, the issue of hallucination has emerged as a critical bottleneck to their application in high-reliability scenarios, attracting widespread attention from both academia and industry. This paper provides a comprehensive and systematic review of hallucinations in LLMs. It begins by systematically

文章引用: 何金骅, 王洪俊. 大语言模型幻觉研究综述[J]. 人工智能与机器人研究, 2026, 15(1): 156-167.
DOI: 10.12677/airr.2026.151016

elucidating the definitions of LLMs and their hallucinations, followed by systematically establishing a classification system for such hallucinatory phenomena and conducting a multi-dimensional analysis of their underlying causes. Subsequently, it surveys methods for mitigating LLM hallucination, such as supervised fine-tuning and retrieval-augmented generation. Finally, based on an analysis of the limitations of existing methods, this review offers an outlook on future research directions for mitigating LLM hallucinations. The aim is to provide a theoretical reference and practical guidance for the development of more reliable and trustworthy Large Language Models.

Keywords

Large Language Models, Hallucinations, Supervised Fine-Tuning, Retrieval-Augmented Generation

1. 引言

近年来，用户数据需求持续增长、算力资源快速提升，算力价格稳步下调[1]。Transformer 架构的提出，以其独特的自注意力机制，成功突破了传统循环神经网络(RNN)和卷积神经网络(CNN)在处理长序列依赖时的计算瓶颈[2]。这一突破推动自然语言处理领域实现从百万参数级模型到亿万参数级模型的跨越，标志着大语言模型时代的到来。这类模型凭借其庞大的参数量，能够从海量的数据中学习到丰富的特征表示与复杂的语言规律，展现出强大的泛化能力与上下文理解能力。它们在传统自然语言处理领域的各类下游任务中均实现了卓越的性能。在智能助手、文本生成、信息抽取等多个领域有着广泛的应用[3]-[5]。然而，随着大语言模型的应用日益深入，其潜在的问题与挑战也逐渐显现，其中模型幻觉问题尤为突出。大语言模型可能生成看似合理但事实上不准确、无依据甚至完全虚构的内容，这种错误的隐蔽性在于其输出往往逻辑自洽，极易对用户产生误导。这严重影响了大模型在事实敏感场景下的可靠性与安全性，在法律领域，幻觉可能杜撰出不存在的判例法规；在医疗咨询中，可能提供错误的药物建议；在金融与军事决策中，一个基于幻觉的分析可能导致灾难性的后果[6]-[8]。有效缓解大语言模型的幻觉，不仅是技术优化的需求，更是保障其安全性、可靠性的核心前提，成为推动大模型技术进一步落地与深化的关键问题。

2. 大语言模型相关概念

2.1. 大语言模型

大语言模型采用自回归的方式生成文本，即在给定初始提示词 p (prompt)的条件下，逐个预测下一个词元(token)，并将已生成的部分作为新的上下文输入，自回归地进行后续预测。整个过程从左到右按顺序进行，每一步都依赖于之前已生成的所有内容，最终生成语义连贯的内容[9]-[11]。

假设要生成一个长度为 T 的文本序列 $x = (x_1, x_2, \ldots, x_T)$，其中每个 x_i 是第 i 个 token。给定提示词 p，模型通过链式法则将联合概率分解为一系列条件概率的乘积，完整的文本联合概率如式(1)所示：

$$P(x|p)\prod_{i=1}^{T}P(x_i|x_1,x_2,\cdots,x_{i-1},\theta) \tag{1}$$

在实际大模型的生成过程中，当生成到第 i 步时，模型会输出一个基于当前上下文的条件概率分布，

其中$(x_1, \ldots, x_{i-1})$表示为$x < i$，表示前$i-1$个已生成的token。该分布覆盖整个词汇表V，即对每个候选词元$x_i \in V$计算概率，这个过程如式(2)所示：

$$P\left(x_i \middle| x_{<i}, \theta\right) \tag{2}$$

在自回归生成的每一步，为了从条件分布中选择具体的token，模型在输出下一个token的时候使用贪婪搜索策略，该策略在每一步中选择模型输出的具有最高条件概率的token，并将该token作为模型最终的输出，这个过程如式(3)所示：

$$x_i = \underset{w \in V}{\operatorname{argmax}} P\left(w \middle| x_{<i}, \theta\right) \tag{3}$$

大模型的输出方式为一个token接着一个token生成，直至输出一个语义连贯且完全符合提示词指令的完整文本序列。

2.2. 大模型幻觉

在大型语言模型研究领域，幻觉(Hallucination)已成为一个关键性技术术语，大模型幻觉指模型生成表面合理却缺乏事实依据的输出内容[12]。这种表面合理性常表现为流畅的语法结构、符合语境的逻辑衔接，专业的表述风格。LLM幻觉的核心在于构造并呈现非事实性内容，且不会明确标识其输出的推测性质[13]。因此，用户在缺乏领域知识或验证手段的情况下，极易将模型生成的虚构内容误认为事实。为应对这一挑战，本文系统综述了LLM幻觉的主要分类体系，并通过典型实例阐释各类幻觉的生成机制与表现特征，旨在为读者提供可操作的辨识工具，提升用户在模型交互中对输出内容真实性的评估能力。

现有研究对大模型幻觉进行了分类，幻觉分类方式主要可以总结为两类，内部幻觉与外部幻觉[14]-[18]。事实性幻觉与忠实性幻觉[19] [20]。

内部幻觉：内部幻觉是指大模型生成的文本与用户提供的输入信息或上下文信息相矛盾的现象。这类错误的核心特征在于其无需参照外部知识，根源在于生成内容内部存在矛盾。

外部幻觉：外部幻觉是指生成文本与训练数据不一致，且无法被输入上下文证实或证伪的现象。外部幻觉的根源在于模型训练数据的知识局限性以及模型无法准确识别自身知识边界。表现为输出现实中不存在的实体、事实或事件，此类幻觉错误需借助外部知识源进行判断。

事实性幻觉：事实性幻觉是指大语言模型生成内容与客观事实相悖的现象。该类幻觉背离现实世界知识体系或经过权威验证的信息源，生成结果与可验证信息在真实性层面存在偏差。

忠实性幻觉：忠实性幻觉是指大模型输出内容偏离输入提示或所提供上下文的现象。其核心特征在于输出与输入之间的不一致性，即模型输出内容可能具备较好的内部逻辑与表面合理性，却未能正确反映用户在输入中明确指定的信息或意图。

本文基于大语言模型在实际应用中的特点，从忠实性幻觉和事实性幻觉两个维度，将大模型幻觉细分为四个类别，分别为上下文不一致，逻辑不一致，事实不一致，事实捏造，分类情况如表 1 所示。针对上述大模型幻觉的分类内容，本文为每一种幻觉类型提供示例，如表 2 所示。

Table 1. Classification of large language model hallucinations
表 1. 大语言模型幻觉分类

幻觉类型	划分维度	划分边界
上下文不一致	输入忠实度	输出内容与用户输入信息中明确陈述的事实或信息直接矛盾
逻辑不一致	输入忠实度	输出内容有内部逻辑不一致问题
事实不一致	事实准确度	输出内容与现实真实信息矛盾
事实捏造	事实准确度	输出内容无法根据现实真实信息知识验证

Table 2. Examples of large language model hallucinations
表 2. 大语言模型幻觉示例

幻觉类型	用户输入	模型输出	说明
上下文不一致	总结以下关于埃菲尔铁塔的段落：埃菲尔铁塔是位于法国巴黎战神广场的一座锻铁格子塔。它建于1887年至1889年间，是1889年世界博览会的核心展品，其设计最初曾遭到法国一些主要艺术家和知识分子的批评。	埃菲尔铁塔是一座位于伦敦的标志性建筑。这座宏伟的钢塔建于20世纪初，其创新的设计在当时就受到了公众和评论家的普遍赞扬。它至今仍是工业时代工程与艺术结合的典范。	大模型的总结与所提供的上下文不一致。它错误地陈述该塔是“钢”结构的并位于“伦敦”，建于“20世纪初”，并且“受到普遍赞扬”。而段落中明确提到它是一座“锻铁”塔，位于“巴黎”，建于“1887年至1889年”间，并且最初遭到了一些法国艺术家和知识分子的“批评”。
逻辑不一致	如果一个披萨被切成8片，我吃了3片，还剩多少片	总共有8片，吃了3片意味着剩下8 − 3 = 5片，所以，还剩下4片。	模型的回应存在逻辑不一致。它正确计算出“剩下5片”，但随后立即得出了相矛盾的结论“还剩下4片”。
事实不一致	电话是谁发明的	托马斯·爱迪生以发明电话而闻名。	大模型的回答属于事实不一致，因为贝尔被公认为发明了第一台实用电话，而托马斯·爱迪生则因发明电灯而闻名。
事实捏造	尼斯湖水怪的历史是怎样的	1933年，伦敦自然历史学会对尼斯湖进行了广泛调查，并正式确认存在一种大型未知水生生物，并将大多数目击事件归因于幸存的蛇颈龙。	大模型的回应属于事实捏造，因为没有经过验证的现实世界记录表明伦敦自然历史学会曾进行过此类调查或正式确认该生物的存在。蛇颈龙理论仍然是一种推测且未经证实的假设。

2.3. 幻觉产生原因

本文探讨了产生大模型幻觉产生的原因，大模型的幻觉源于训练数据，模型架构和模型学习知识的过程，以及用户提供的提示词。

2.3.1. 训练数据

大模型在大量的数据上进行预训练，大模型的能力来源于对数据的学习，数据的质量好坏严重影响模型的性能。大模型训练的数据中的噪声数据包含虚假数据，重复数据，偏见数据等。噪声数据会严重影响模型的性能，让模型学习到错误的语料信息，导致模型输出错误的结果[21]-[23]。大语言模型的预训练数据主要抓取自互联网、书籍、新闻等通用文本，虽然覆盖面广，但对于法律、医疗、金融、教育等高度专业化的垂直领域[6]-[8] [24]，其用于训练大模型的专业领域的数据远远不够，导致大模型面对垂直领域问题时容易产生幻觉。大型语言模型的预训练数据是静态的，其知识存在一个“截止日期”。当被问及超出其训练数据范围的新事件或信息时，模型无法获取到实时的数据，输出捏造的错误的内容，产生幻觉[25] [26]。

2.3.2. 模型架构及训练过程

幻觉产生的根本原因之一在于大语言模型的自回归特性，此类模型的核心任务是进行词元预测，即根据已生成的文本序列，预测并输出概率最高的下一个词元。这一过程的输出直接目标并非保证事实的准确性，而是保证语言上的连贯性与高可能性[19]。

模型训练过程面临着训练阶段与推理阶段数据分布不一致性。在训练中，模型以真实的上文为条件进行预测，而在推理时，它必须以自身生成的、可能存在偏差的上文为条件进行后续预测，这种条件分布的偏移会引发误差的级联传播，导致生成内容产生偏差[27]。当大模型在训练过程中指定针对一些性能指标过度优化，也可能会导致大模型在一些领域产生幻觉。当大型语言模型的能力被对齐到超出其训练数据所能充分支持的范围时，尤其是在其训练数据不足的专业领域，会产生幻觉[28]。

大语言模型在文本生成过程中采用采样策略，这些策略为输出引入了随机性元素。例如，大模型输出阶段可调节超参数温度(temperature)，较高的温度设置可以增强创造力，但也会因为倾向于选择低概率的词元，而显著增加产生幻觉的风险[29]。

2.3.3. 用户提示词

用户的提示词也是造成大模型幻觉的一个重要因素，用户的提示词会诱导加剧大模型产生幻觉。当用户提示中嵌入蓄意注入或无意间包含的虚假内容时，模型倾向于基于这些错误信息进行连贯性阐述与扩展，从而生成与事实不符的内容。这一过程本质上是大模型表现出一种过度确认的倾向，会优先追求输出内容的说服力与流畅性。这一特性会加剧基于提示词提供的虚假内容所带来的影响，使模型输出的内容呈现出专业性和准确性，从而提升了其可信度，对用户造成更强的误导性[30] [31]。

3. 大语言模型幻觉消减方法

本文系统探讨大语言模型的幻觉消减方法。当前研究可主要归结为模型层面与系统应用层面两大类，其关系如图 1 所示。在大模型层面的方法包括有监督微调(SFT supervised fine-tuning)，基于人类反馈的强化学习(RLHF reinforcement learning with human feedback)，系统层面包括提示词技巧(PT prompt trick)，检索增强技术(RAG retrieval augment)。以下重点介绍每个方法及其应用。

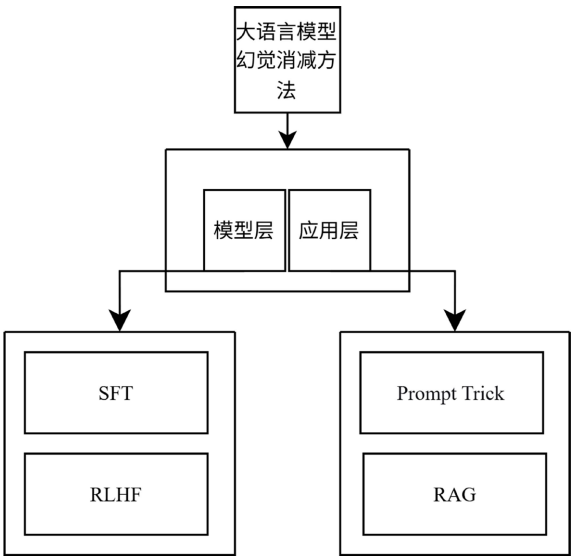

Figure 1. Hallucination mitigation methods for large language models
图 1. 大语言模型幻觉消减方法

3.1. 模型层

3.1.1. 有监督微调

有监督微调是大模型适应下游任务的关键，通过人工标注的高质量数据对模型进行微调。根据下游特定任务设计好损失函数，该损失函数用于衡量大语言模型的输出和真实标签之间的差异，通过最小化损失函数，模型参数进行调整，使模型的输出更加逼近真实标签[32]-[35]。

例如，Luo 等[36]提出 Code Evol-Instruct 指令进化结合 SFT 的方法，以 StarCoder 15B 和 CodeLlama-34B-Python 为预训练基础模型，通过迭代进化 Code Alpaca (约 20 k 初始样本)生成了约 78 k 样本，用该数据集微调后的 WizardCoder 模型，其性能超越所有开源代码生成大模型。Zhou 等[37]为优化 LawGPT 模型在下游法律任务中的表现，先构建含 200 K 开源犯罪相关样本、20 K JEC-QA 法律问答样本及 80 K 经 GPT-3.5 Turbo 优化的高质量样本，共 300 K 知识驱动指令数据集 DLFT。以 DLFT 为训练数据，过程中采用 LoRA 技术优化模型参数，经 SFT 后的 LawGPT，在零样本设置下的 8 个法律任务中性能优于开源 LLaMA 7B，降低了模型在输出法律任务问题的幻觉，验证了 SFT 对提升模型法律任务性能的作用。Li 等[38]开发医疗对话模型 ChatDoctor 时，以 Meta 的 LLaMA-7B 模型进行微调，先通过斯坦福 Alpaca 提供的数据进行微调，模型获取基础对话能力，再用 10 万条 HealthCareMagic 真实医患对话数据进一步微调以强化医疗专业能力，同时搭配在线(如 Wikipedia)与离线(基于 MedlinePlus 的疾病数据库)自主知识检索模块；最终模型可准确回答新疾病(如猴痘)、新药(如 Daybue)及专业诊疗问题，在 BERTScore (0.8444 ± 0.0185)、召回率(0.8451 ± 0.0157)、F1 值(0.8446 ± 0.0138)等性能指标均优于现有模型。该模型能够辅助医生问诊、提升医疗资源匮乏地区的医疗咨询可及性。通过在高质量、大规模的领域特定数据上进行微调，能够有效提升模型在特定领域生成能力。但微调模型的效果高度依赖于微调数据的质量与规模。

SFT 高度依赖于微调数据的质量与规模，若数据中存在噪声、偏见或虚假信息，模型不仅无法消减幻觉，反而可能固化这些错误；同时，获取大规模、高质量的专业领域标注数据成本高昂，限制了其在垂直领域的应用。SFT 作为一种外部对齐方法，并未触及大模型概率化生成范式的固有局限，它通过模仿正确答案来约束行为，而非改变模型以语言连贯性而非事实准确性为自回归生成目标的本质。SFT 的可靠性提升是有限的，主要局限于已训练的数据分布，面对新知识或复杂推理问题时，模型仍可能产生幻觉。

3.1.2. 基于人类反馈的强化学习

RLHF 是一种通过人类主观反馈来指导强化学习过程的模型训练方法。RLHF 的核心是用人类对模型输出的评价(通常表现为对多个输出的排序)来训练一个奖励模型，以替代难以设计的奖励函数，引导模型的行为更符合人类的价值观和偏好。RLHF 通常包括三个步骤，先用高质量数据监督微调，再用人类反馈训练奖励模型，最后用强化学习算法优化模型[39]-[43]。

Iacovides 等[44]提出金融情感分析框架 FinDPO，以 Llama-3-8B-Instruct 为基础模型，结合直接偏好优化(DPO)的人类偏好对齐与低秩适配(LoRA)技术，将三个公开金融新闻数据集的 32,970 条样本转为(偏好，非偏好)训练对，通过 DPO 损失更新模型权重，结果表明，F1 分数达 0.846 (平均超 FinGPTv3.3 11%)，66.64%年化回报及 2.03 夏普比率。Dai 等[45]在 RLHF 框架上提出双维度解耦改进，拆解人类偏好标注为独立有用性排序与无害性评估 + 安全标签以生成专属数据集，再分别训练 RLHF 的奖励模型(RM，量化有用性)与新增的成本模型(CM，识别有害性)，最终通过拉格朗日动态约束，在 RLHF 微调中实现最大化奖励。实验中 Alpaca-7B 经三轮 Safe RLHF 微调后，有害响应率从 53.08%降至 2.45%。Yang 等[46]提出中文医疗大模型 Zhongjing 及 CMtMedQA 数据集，以解决通用模型医疗领域短板。Zhongjing 基于 Ziya-LLaMA 模型，采用预训练(多源医疗语料)→SFT (四类数据融合)→RLHF (专家标注 + PPO 算法)全流程

训练，CMtMedQA 数据集包含 14 科室共 7 万条多轮医患对话。在 CMtMedQA (多轮对话)与 Huatuo-26M (单轮对话)两大测试集上，Zhongjing 在专业性、流畅性、安全性三大维度及九种细分能力上全维度超越 BenTsao、DoctorGLM、HuatuoGPT 等基线模型。

RLHF 虽能有效对齐模型行为与人类偏好，但其内在缺陷与应用难点突出。其多阶段流程需要大量高质量的人类反馈数据，导致成本高昂且复杂。同时，该方法过度依赖人类主观、可能存在偏见的人类反馈，易训练出有偏差的奖励模型，误导模型优化方向。

3.2. 应用层

3.2.1. 提示词技巧

提示词工程作为控制大语言模型行为、缓解幻觉现象的关键技术路径，相比依赖模型微调、架构优化等传统方法，无需额外的大规模算力投入或数据标注成本，更为经济高效且落地门槛更低。提示词工程是指通过设计和优化用户输入的提示词，为大模型提供特定的上下文、明确的指令和期望的输出格式，从而引导其生成更准确、更符合预期的结果。这种方法不仅能有效降低大模型的幻觉发生率，还能解决输出结果发散、格式不统一、任务适配性差等问题，广泛适用于文本生成、信息抽取、数据分析、代码辅助、多轮对话等各类场景[47]-[49]。

Wei 等[50]提出思维链(Chain of Thought, CoT) CoT 的核心是打破传统提示词中直接输出答案的模式，通过在提示词中加入如“请分步推理”，“说明思考过程”等语句，让模型在输出最终的内容前，以连贯有逻辑顺序的文本形式输出中间推理步骤。这种创新方法的核心依据是大语言模型本质上是为预测下一个 token 序列而设计的，而非进行显式的推理过程。通过在提示词中清晰描述必要的推理步骤，能够引导大语言模型生成更具推理性和逻辑性的输出。

除思维链提示词外，通过明确设计系统提示词(用于引导大模型行为的特殊指令信息)来约束模型输出。在系统提示词中明确纳入“禁止传播虚假或无法验证信息”的指令，可以从源头缓解幻觉的生成。Touvron 等[51]在其开发的 Llama 2-Chat 模型中便采用了这一思路，其设计的系统提示词明确规定“若你不知道某一问题的答案，请切勿生成虚假的信息”。

提示词技术虽经济高效且门槛低，但其内在缺陷与应用难点同样突出。作为一种外部引导，它并未触及模型概率化生成的根本局限，仅在输出层面施加约束，治标不治本。其效果高度依赖用户设计，稳定性差，且在面对恶意提示词时尤为脆弱。需要恶意提示词识别，恶意提示词改写等后处理工作。该方法在解决事实性幻觉方面能力有限，难以独立承担高可靠性场景下的保障任务。

3.2.2. 检索增强技术

检索增强技术(RAG)的核心设计是让检索组件与生成模型协同作用，实现动态知识注入与内容生成的闭环。其工作流程如图 2 所示。RAG 主流方案在推理阶段无需微调，通过向量检索、混合检索等技术从外部知识库召回相关文档块，并将其作为上下文整合到用户提示词中。大语言模型基于这些可验证的信息，结合用户查询生成有依据的回答，显著提升事实准确性并降低幻觉率[52]-[56]。

为了解决传统 RAG 在全局性问题上的不足，Edge 等[57]提出的 GraphRAG 方法，通过构建层次化知识图谱索引并采用 Map-Reduce 机制，其流程分为两阶段，离线构建阶：利用大模型从源文档中提取实体、关系等元素构建知识图谱，再通过社区检测算法将其划分为层次化社区，并为每个社区预生成摘要，形成结构化索引。在线查询阶段：针对用户查询，系统并行地利用社区摘要生成多个部分答案，排序并融合这些答案，最终汇总成一个覆盖全局视角的最终答案，从而实现了对整个语料库的全局性意义构建。

为解决 GraphRAG 存在的 token 开销过大，动态更新成本高昂和检索效率低下的问题。Guo 等[58]提

出了一种融合图增强文本索引与双层检索范式的方法 LightRAG。该方法先通过 LLM 从文本中抽取实体和关系构建知识图谱，并为每个节点和边生成高效的键值对索引。在检索时，其双层检索范式结合了图与向量，它先从查询中提取局部和全局关键词，再利用向量数据库将局部关键词匹配图中的实体，将全局关键词与关联的关系进行匹配，最后通过图结构扩展到一跳邻居节点，通过图结构扩展，引入已匹配实体和关系的邻居节点，从而在保证检索精度的同时，丰富了信息的广度和深度。还通过其增量更新算法，仅需将新数据的子图与原图合并即可完成更新，有效避免了 GraphRAG 的全量重建开销，实现了高效、低成本且适应性强的 RAG 系统。

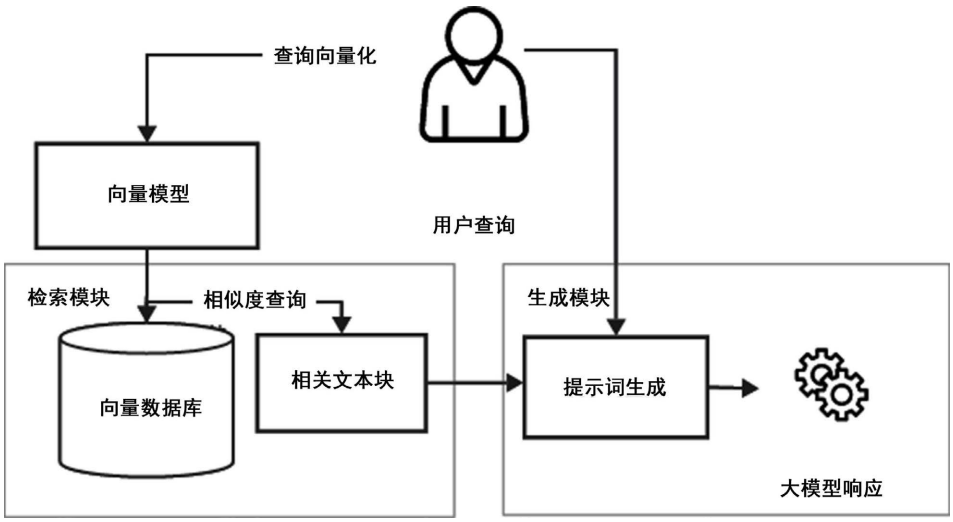

Figure 2. Diagram of the standard RAG workflow
图 2. 标准 RAG 工作流程图

RAG 的有效性高度依赖于外部知识库的完整性、时效性与全面性。RAG 系统的最终输出质量受到检索模块性能的根本性制约。在检索过程返回不相关的文本块时，大模型会基于错误的上下文进行推理，输出错误的内容，导致模型产生幻觉。

在实际应用中，对于诸如 GraphRAG 等方法，其复杂的知识图谱索引构建涉及大规模的离线预处理，以及知识库的动态更新则面临显著的经济花费与计算，在实际应用中开销较大，使得知识库的构建与维护成为一项资源密集型工作。

如表 3 所示，从数据依赖、计算成本、消减幻觉效果、可解释性等多个维度对 SFT、RLHF、PT、RAG 等主流方法进行比较。

Table 3. Comparison of different hallucination reduction methods
表 3. 不同幻觉消减方法比较

方法	数据依赖	计算成本	消减幻觉效果	可解释性
SFT	高度依赖人工标注的高质量数据，效果受数据质量与规模制约	需设计特定损失函数，参数调整优化	对特定领域幻觉效果显著	较低
RLHF	需要高质量人类反馈数据，需专家参与	包含三个复杂阶段，资源投入大	提升特定领域能力上效果显著	中等

续表

PT	无需额外大规模数据标注，依赖用户提示词设计能力	经济高效，门槛低	缓解忠实性幻觉，对事实性幻觉效果有限	较高
RAG	高度依赖外部知识库相关性与准确性	离线向量数据库索引构建，在线检索，向量数据库更新维护	显著提升事实准确性	较高

4. 未来研究方向

4.1. 恶意提示词的智能识别

含有诱导性词汇的提示词是导致大模型产生幻觉的重要因素，当前有思维链、系统提示等方法用于缓解幻觉，但这些方法仅能实现推理透明度的提升或输出的初步约束，无法有效识别提示词中的前提错误与恶意指令。现有技术难以精准区分提示词中的正常需求与诱导性指令，导致模型易被恶意提示引导生成虚假内容。

未来研究方向在于设计恶意提示词识别算法，深入分析用户提示词的意图特征，构建识别体系，实现对恶意提示词的精准识别，阻断恶意提示词引发的模型幻觉，提升模型的抗干扰能力与输出可靠性。未来工作需从识别用户输入的提示词为出发点，设计多维度恶意提示词识别算法。该算法不仅需捕捉“编造”“虚构”等显性诱导词汇，还需深度挖掘语义层面的隐性特征(如句子主谓关系矛盾、事实断言缺乏权威来源支撑)，同时融合上下文动态特征(如用户多轮对话中提示词的变化)与用户行为特征(如是否通过反复修改提示词规避模型约束)，构建多特征融合的识别模型，以显著提升恶意提示词的识别精度与泛化能力。构建意图分类–风险分级–响应适配的识别体系。通过意图分类模块，基于语义理解与行为特征区分正常需求与诱导需求。依据恶意提示词危害程度等指标，划分为低、中、高三个风险等级，针对不同风险等级，实现对恶意提示词的精细化处置。强化恶意提示词识别算法与其他幻觉消减方法的协同联动。当系统识别出高风险提示词时，可直接触发模型生成流程的阻断机制，无需启动后续的检索增强和事实校验环节，减少不必要的计算资源消耗、提升处理效率，从输入源头遏制幻觉风险的产生。

4.2. 模型内部可靠性机制的构建

当前主流的幻觉缓解方法大部分依赖于外源性知识检索，此类方法虽可通过外部事实辅助提升事实准确性，但未能触及大模型概率化生成范式的固有局限与知识表征体系的结构性缺陷。未来研究的核心在于构建模型内部可靠性机制，赋予模型独立于外部提供的事实性校准，自主修正输出内容可靠性的能力，从生成逻辑底层抑制幻觉的产生。

融合知识图谱与模型推理的深度集成是未来技术路径。该方法通过将领域知识图谱(如医疗领域的药物–疾病关系网络)编码为可微分的注意力约束。该方法可设计构建 GNN-Transformer 混合架构，将知识图谱节点转化为位置偏置向量并注入自注意力层，使生成过程实时关联可信知识源。这一技术的核心在于实现模型功能范式的转变，从依赖统计规律的概率预测器转变为具备知识验证能力的推理系统。然而，该路径面临显著技术挑战：知识图谱的动态更新需求与模型增量训练之间存在固有冲突，新知识的实时补充易引发灾难性遗忘。

另一关键研究方向是构建细粒度不确定性评估模块，该模块通过多维度量化模型对生成片段的置信度，包括预测分布熵值、上下文一致性及任务敏感度等指标，并在置信度低于预设阈值时主动中断生成链路。此机制对抑制误差级联传播具有决定性作用，可有效阻断初始微小偏差在自回归生成过程中造成

的影响。该技术需解决三大核心问题：评估指标的可靠性保障、实时推理的效率优化，以及模型过度自信表征的识别与校准。

5. 结语

大语言模型幻觉问题已成为制约其可信应用的关键瓶颈，对人工智能的安全落地构成严峻挑战。本文系统梳理了幻觉的定义、分类体系与产生原因，并全面综述了从模型层到应用层的各类消减方法。尽管现有方法，如有监督微调与检索增强技术，在一定程度上缓解了幻觉现象，但它们多依赖外部知识辅助，尚未从根本上解决大模型概率化生成范式的固有缺陷。因此，未来的研究重心应转向构建模型内部的可靠性机制，使模型具备自主事实校准与修正能力，从生成逻辑的底层抑制幻觉的产生。相信通过持续的理论突破与技术创新，构建真正可靠、可信的大语言模型将从愿景走向现实，为人工智能的健康发展奠定坚实基础。

参考文献

[1] Coyle, D. and Hampton, L. (2024) 21st Century Progress in Computing. *Telecommunications Policy*, **48**, Article 102649. https://doi.org/10.1016/j.telpol.2023.102649

[2] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[3] Zhao, W.X., Zhou, K., Li, J., *et al*. (2023) A Survey of Large Language Models. arXiv:2303.18223.

[4] Hadi, M.U., Qureshi, R., Shah, A., *et al*. (2023) A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. Authorea Preprints.

[5] Minaee, S., Mikolov, T., Nikzad, N., *et al*. (2024) Large Language Models: A Survey. arXiv:2402.06196.

[6] Dahl, M., Magesh, V., Suzgun, M. and Ho, D.E. (2024) Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, **16**, 64-93. https://doi.org/10.1093/jla/laae003

[7] Liu, Z., Huang, D., Huang, K., Li, Z. and Zhao, J. (2020) FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4513-4519. https://doi.org/10.24963/ijcai.2020/622

[8] Kim, Y., Jeong, H., Chen, S., *et al*. (2025) Medical Hallucinations in Foundation Models and Their Impact on Healthcare. arXiv:2503.05777.

[9] Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L.S. and Wong, D.F. (2025) A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Computational Linguistics*, **51**, 275-338. https://doi.org/10.1162/coli_a_00549

[10] Trummer, I. (2024) Large Language Models: Principles and Practice. 2024 *IEEE* 40*th International Conference on Data Engineering* (*ICDE*), Utrecht, 13-16 May 2024, 5354-5357. https://doi.org/10.1109/icde60146.2024.00404

[11] Wang, Z., Chu, Z., Doan, T.V., Ni, S., Yang, M. and Zhang, W. (2024) History, Development, and Principles of Large Language Models: An Introductory Survey. *AI and Ethics*, **5**, 1955-1971. https://doi.org/10.1007/s43681-024-00583-7

[12] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., *et al*. (2025) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, **43**, 1-55. https://doi.org/10.1145/3703155

[13] Joshi, S. (2025) Mitigating LLM Hallucinations: A Comprehensive Review of Techniques and Architectures. https://doi.org/10.2139/ssrn.5267540

[14] Elchafei, P. and Abu-Elkheir, M. (2025) Span-Level Hallucination Detection for LLM-Generated Answers. arXiv:2504.18639.

[15] Orgad, H., Toker, M., Gekhman, Z., *et al*. (2024) LLMs Know More than They Show: On the Intrinsic Representation of llm Hallucinations. arXiv:2410.02707.

[16] Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., *et al*. (2025) Hallulens: LLM Hallucination Benchmark. *Proceedings of the* 63*rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Vienna, 27 July-1 August 2025, 24128-24156. https://doi.org/10.18653/v1/2025.acl-long.1176

[17] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., *et al*. (2023) Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, **55**, 1-38. https://doi.org/10.1145/3571730

[18] Sun, W., Shi, Z., Gao, S., Ren, P., De Rijke, M. and Ren, Z. (2023) Contrastive Learning Reduces Hallucination in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 13618-13626. https://doi.org/10.1609/aaai.v37i11.26596

[19] Li, J., Chen, J., Ren, R., *et al*. (2024) The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. arXiv:2401.03205.

[20] Chen, S., Zhang, F., Sone, K. and Roth, D. (2021) Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Online, 6-11 June 2021, 5935-5941. https://doi.org/10.18653/v1/2021.naacl-main.475

[21] Gautam, A.R. (2025) Impact of High Data Quality on LLM Hallucinations. *International Journal of Computer Applications*, **187**, 35-39.

[22] Carlini, N., Tramer, F., Wallace, E., *et al*. (2021) Extracting Training Data from Large Language Models. 30*th USENIX security symposium* (*USENIX Security* 21), 11-13 August 2021, 2633-2650.

[23] Sheng, E., Chang, K., Natarajan, P. and Peng, N. (2021) Societal Biases in Language Generation: Progress and Challenges. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), Online, 1-6 August 2021, 4275-4293. https://doi.org/10.18653/v1/2021.acl-long.330

[24] Ho, H., Ly, D. and Nguyen, L.V. (2024) Mitigating Hallucinations in Large Language Models for Educational Application. 2024 *IEEE International Conference on Consumer Electronics-Asia* (*ICCE-Asia*), Danang, 3-6 November 2024, 1-4. https://doi.org/10.1109/icce-asia63397.2024.10773965

[25] Zhu, C., Chen, N., Gao, Y., Zhang, Y., Tiwari, P. and Wang, B. (2025) Is Your LLM Outdated? A Deep Look at Temporal Generalization. *Proceedings of the* 2025 *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Volume 1: Long Papers), Albuquerque, 29 April-4 May 2025, 7433-7457. https://doi.org/10.18653/v1/2025.naacl-long.381

[26] Karpowicz, M.P. (2025) On the Fundamental Impossibility of Hallucination Control in Large Language Models. arXiv:2506.06382.

[27] Pozzi, A., Incremona, A., Tessera, D. and Toti, D. (2025) Mitigating Exposure Bias in Large Language Model Distillation: An Imitation Learning Approach. *Neural Computing and Applications*, **37**, 12013-12029. https://doi.org/10.1007/s00521-025-11162-0

[28] Kirk, R., Mediratta, I., Nalmpantis, C., *et al*. (2023) Understanding the Effects of RLHF on LLM Generalisation and Diversity. arXiv:2310.06452.

[29] Waldo, J. and Boussard, S. (2024) GPTs and Hallucination: Why Do Large Language Models Hallucinate? *Queue*, **22**, 19-33. https://doi.org/10.1145/3688007

[30] Xu, X., Kong, K., Liu, N., *et al*. (2023) An LLM Can Fool Itself: A Prompt-Based Adversarial Attack. arXiv:2310.13345.

[31] Rawte, V., Priya, P., Tonmoy, S.M., *et al*. (2023) Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness. arXiv:2309.11064.

[32] Parthasarathy, V.B., Zafar, A., Khan, A., *et al*. (2024) The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv:2408.13296.

[33] P, M. and Velvizhy, P. (2025) A Comprehensive Review of Supervised Fine-Tuning for Large Language Models in Creative Applications and Content Moderation. 2025 *International Conference on Inventive Computation Technologies* (*ICICT*), Kirtipur, 23-25 April 2025, 1294-1299. https://doi.org/10.1109/icict64420.2025.11005111

[34] Xu, L., Xie, H., Qin, S.Z.J., *et al*. (2023) Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. CoRR.

[35] Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., *et al*. (2025) Parameter-Efficient Fine-Tuning in Large Language Models: A Survey of Methodologies. *Artificial Intelligence Review*, **58**, Article No. 227. https://doi.org/10.1007/s10462-025-11236-4

[36] Luo, Z., Xu, C., Zhao, P., *et al*. (2023) Wizardcoder: Empowering Code Large Language Models with Evolinstruct. arXiv:2306.08568.

[37] Zhou, Z., Shi, J.X., Song, P.X., *et al*. (2024) LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model. CoRR.

[38] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S. and Zhang, Y. (2023) Chatdoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (Llama) Using Medical Domain Knowledge. *Cureus*, **15**, e40895. https://doi.org/10.7759/cureus.40895

[39] Ouyang, L., Wu, J., Jiang, X., *et al*. (2022) Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, **35**, 27730-27744.

[40] Christiano, P.F., Leike, J., Brown, T., *et al*. (2017) Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, **30**, 4299-4307.

[41] Kaufmann, T., Weng, P., Bengs, V., *et al*. (2024) A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925.

[42] Wang, Z., Bi, B., Pentyala, S.K., *et al*. (2024) A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv:2407.16216.

[43] Srivastava, S.S. and Aggarwal, V. (2025) A Technical Survey of Reinforcement Learning Techniques for Large Language Models. arXiv:2507.04136.

[44] Iacovides, G., Zhou, W. and Mandic, D. (2025) Findpo: Financial Sentiment Analysis for Algorithmic Trading through Preference Optimization of LLMs. *Proceedings of the* 6*th ACM International Conference on AI in Finance*, Singapore, 15-18 November 2025, 647-655. https://doi.org/10.1145/3768292.3770367

[45] Dai, J., Pan, X., Sun, R., *et al*. (2024) Safe RLHF: Safe Reinforcement Learning from Human Feedback. *The Twelfth International Conference on Learning Representations*, Vienna, 7-11 May 2024, 47991-48018.

[46] Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, Y., *et al*. (2024) Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 19368-19376. https://doi.org/10.1609/aaai.v38i17.29907

[47] Feldman, P., Foulds, J.R. and Pan, S. (2023) Trapping LLM Hallucinations Using Tagged Context Prompts. arXiv:2306.06085.

[48] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. (2023) Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, **55**, 1-35. https://doi.org/10.1145/3560815

[49] Schulhoff, S., Ilie, M., Balepur, N., *et al*. (2024) The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv:2406.06608.

[50] Wei, J., Wang, X., Schuurmans, D., *et al*. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, **35**, 24824-24837.

[51] Touvron, H., Martin, L., Stone, K., *et al*. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

[52] Lewis, P., Perez, E., Piktus, A., *et al*. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, **33**, 9459-9474.

[53] Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q. and Liu, Z. (2024) Evaluation of Retrieval-Augmented Generation: A Survey. In: Zhu, W., *et al*., Eds., *Communications in Computer and Information Science*, Springer Nature, 102-120. https://doi.org/10.1007/978-981-96-1024-2_8

[54] Hu, Y. and Lu, Y. (2024) Rag and Rau: A Survey on Retrieval-Augmented Language Model in Natural Language Processing. arXiv:2404.19543.

[55] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., *et al*. (2024) A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the* 30*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, 25-29 August 2024, 6491-6501. https://doi.org/10.1145/3637528.3671470

[56] Zhao, P., Zhang, H., Yu, Q., *et al*. (2024) Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473.

[57] Edge, D., Trinh, H., Cheng, N., *et al*. (2024) From Local to Global: A Graph Rag Approach to Query-Focused Summarization. arXiv:2404.16130.

[58] Guo, Z., Xia, L., Yu, Y., Ao, T. and Huang, C. (2025) Lightrag: Simple and Fast Retrieval-Augmented Generation. *Findings of the Association for Computational Linguistics*: *EMNLP* 2025, Suzhou, 4-9 November 2025, 10746-10761. https://doi.org/10.18653/v1/2025.findings-emnlp.568