计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(1): 123−146, 2026

大语言模型幻觉检测方法综述

李自拓[1] 孙建彬[1] 陈广州[1] 方馨悦[2,3] 崔瑞靖[1] 田植良[2,3] 黄 震[2,3] 杨克巍[1]
[1]（国防科技大学系统工程学院 长沙 410073）
[2]（国防科技大学计算机学院 长沙 410073）
[3]（并行与分布计算全国重点实验室（国防科技大学） 长沙 410073）
（lizituonudt@nudt.edu.cn）

Survey of Hallucination Detection Methods for Large Language Models

Li Zituo[1], Sun Jianbin[1], Chen Guangzhou[1], Fang Xinyue[2,3], Cui Ruijing[1], Tian Zhiliang[2,3], Huang Zhen[2,3], and Yang Kewei[1]
[1]（*College of Systems Engineering*, *National University of Defense Technology*, *Changsha* 410073）
[2]（*College of Computer Science and Technology*, *National University of Defense Technology*, *Changsha* 410073）
[3]（*National Key Laboratory of Parallel and Distributed Computing* (*National University of Defense Technology*), *Changsha* 410073）

Abstract In recent years, large language models (LLMs) have made significant strides in the field of natural language processing (NLP), demonstrating impressive capabilities in both language understanding and generation. However, despite these advancements, LLMs still face numerous challenges in practical applications. One such issue that has garnered extensive attention from both the academic and industrial communities is the problem of hallucinations. Effectively detecting hallucinations in large language models is a critical challenge for ensuring their reliable, secure, and trustworthy application in downstream tasks such as text generation. We provide a comprehensive review of methods for detecting hallucinations in large language models. Firstly, we introduce the concept of large language models and clarify the definition and classification of hallucinations. Then we systematically examine the characteristics of LLMs throughout their entire lifecycle, from construction to deployment, and delve into the mechanisms and causes of hallucinations. Secondly, based on practical application requirements and considering factors such as model transparency in different task scenarios, we categorize hallucination detection methods into two types: those for white-box models and those for black-box models. A focused review and in-depth comparison of these methods are provided. Then we analyze and summarize the current mainstream benchmarks for hallucination detection, laying a foundation for future research in this area. Finally, we identify various potential research directions and new challenges in the detection of hallucinations in large language models.

Key words hallucination detection; large language model; factual consistency; text generation; natural language processing (NLP)

摘 要 近年来，大语言模型（large language models, LLMs）在自然语言处理（natural language processing, NLP）等领域取得了显著进展，展现出强大的语言理解与生成能力。然而，在实际应用过程中，大语言模型仍然面临诸多挑战。其中，幻觉（hallucination）问题引起了学术界和工业界的广泛关注。如何有效检测大语言模型幻觉，成为确保其在文本生成等下游任务可靠、安全、可信应用的关键挑战。该研究着重对大语言模型幻觉检测方法进行综述：首先，介绍了大语言模型概念，进一步明确了幻觉的定义与分类，系

收稿日期：2025-02-06；**修回日期**：2025-09-16
基金项目：国家自然科学基金项目（72471238）；国家自然科学基金重点项目（72231011）；国家自然科学基金青年科学基金项目（72401287）
This work was supported by the National Natural Science Foundation of China (72471238), the Key Program of the National Natural Science Foundation of China (72231011), and the National Natural Science Foundation of China for Young Scientists (72401287).

统梳理了大语言模型从构建到部署应用全生命周期各环节的特点,并深入分析了幻觉的产生机制与诱因;其次,立足于实际应用需求,考虑到在不同任务场景下模型透明度的差异等因素,将幻觉检测方法划分为针对白盒模型和黑盒模型2类,并进行了重点梳理和深入对比;而后,分析总结了现阶段主流的幻觉检测基准,为后续开展幻觉检测奠定基础;最后,指出了大语言模型幻觉检测的各种潜在研究方法和新的挑战。

关键词 幻觉检测;大语言模型;事实一致性;文本生成;自然语言处理

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550069 **CSTR:** 32373.14.issn1000-1239.202550069

近年来,大语言模型(large language models, LLMs)的发展为自然语言处理(natural language processing, NLP)领域带来了革命性的变革[1-3]。大语言模型通常基于Transformer架构,通过在海量多样化数据上进行训练,形成拥有数十亿甚至数万亿参数的庞大模型,从而具备了深层次的语言理解和生成能力。这使其能够在摘要生成、机器翻译、开放性问答等任务上展现出卓越表现[4-6]。

然而,研究表明,大语言模型在诸多任务中存在“幻觉”(hallucination)现象[7-9]。这一问题显著影响了模型生成内容的准确性和可信性。例如,在生成文本时,模型可能会凭空捏造不存在的事实,错误地引用信息或构建不合理的逻辑[10]。这对大语言模型的实际应用构成了潜在风险,尤其在军事、法律和金融等高敏感领域更为突出[11-13]。因此,如何有效检测幻觉成为确保大语言模型可靠、安全、可信的关键挑战。

幻觉检测(hallucination detection)作为一种后处理方法,旨在通过系统地评估和验证生成的文本内容,识别其中存在的事实性偏差、不合逻辑的内容或其他错误输出[14]。传统NLP领域的幻觉检测方法主要依赖于规则、统计模型或简单的机器学习技术来识别文本中的异常或错误[15]。但是,由于大语言模型的生成能力建立在海量数据和庞大参数的基础上,其生成内容具有高度的语言流畅性和表层语义一致性,这使得幻觉现象更加隐蔽,检测难度显著增加。传统NLP领域的幻觉检测方法已难以直接适用于大语言模型。因此,研究逐渐转向针对大语言模型的幻觉检测方法[16-17]。

尽管近年来幻觉检测研究不断深入,技术方法和测评标准层出不穷,但系统性总结和分析却依然不足。这在一定程度上阻碍了对幻觉现象本质的深入认识,并限制了幻觉检测方法的发展。因此,对现有幻觉检测方法进行全面梳理和归纳,深入探讨其共性与差异,对于构建更精准、更高效的幻觉检测方法具有重要意义。鉴于此,本文在充分借鉴与吸收文献[14−17]研究成果的基础上,结合近年来大语言模型发展所引发的幻觉问题新趋势,力求在以下3个方面做出进一步的拓展与深化:

1)针对当前幻觉现象日益复杂化的问题,本文基于近期文献成果,进一步明确了大语言模型幻觉的边界,并进行了更加细致的分类,力求覆盖更多的幻觉类型及其交叉叠加特性。不同于已有综述多基于早期定义以及分类体系较为粗略的局限,本文在分类粒度与类型覆盖方面进行了扩展。同时,系统性梳理了大语言模型从构建、预训练、微调、对齐到推理的全生命周期,深入分析并总结了幻觉现象的生成机制及潜在诱因。

2)立足于实际应用需求,根据模型透明度及对内部信息的访问能力,提出了新的幻觉检测方法分类框架,系统分析了针对白盒模型和黑盒模型的检测方法在各类应用场景中的特点、优势与局限性,为实际任务中方法选择与优化提供了理论参考。

3)深入研究现有的幻觉检测基准,本文从幻觉类型、应用任务、数据构建方式与评价指标等多维度进行了更加细致的分类与特性分析,进一步揭示了当前基准体系中存在的标准不统一、覆盖度有限等问题。此外,总结了当前研究中存在的不足与面临的挑战,并展望了未来大语言模型幻觉检测的研究方向。

1 相关概念

1.1 大语言模型

大语言模型通过条件概率分布建模,逐步生成语法正确且语义连贯的文本[18-19]。其核心目标在于生成符合给定条件分布的自然语言序列。

设生成一段长度为T的文本$\boldsymbol{x}=(x_1,x_2,\cdots,x_T)$,其中$x_t$表示生成的第$t$个元素(如单词或子词)。生成过程通常以用户提供的提示$\boldsymbol{p}$作为条件,记为$\boldsymbol{x}\sim LLM(\cdot|\boldsymbol{p})$。对于文本$\boldsymbol{x}$中的单个元素$x_t$,若生成过程建模为

基于条件概率分布的递归生成，则每个元素的生成依赖于已生成序列的条件概率：

$$x_t \sim LLM(\cdot|x_1, x_2, \cdots, x_{t-1}, \boldsymbol{p})。 \tag{1}$$

完整文本的联合概率分布可表示为

$$P(\boldsymbol{x}|\boldsymbol{p}) = \prod_{t=1}^{T} P(x_t|x_1, x_2, \cdots, x_{t-1}, \boldsymbol{p})。 \tag{2}$$

这种分解形式利用了语言模型的自回归特性[20]，通过条件概率逐步生成每个单词或子词，最终生成一段符合给定提示$\boldsymbol{p}$的完整文本。

1.2 幻觉及分类

大语言模型具有交互性强、生成自由度高、泛化能力强等特点。从用户提示语的角度来看，大语言模型产生幻觉会受到提示语的显著影响。考虑到大语言模型较强的用户交互性带来的潜在风险，在文献[15−17]的基础上进一步明确了大语言模型幻觉的定义，即在输入提示合理的情况下，模型生成的文本虽然语言形式和结构上正确，但在语义、事实性、逻辑或上下文一致性上存在偏差、不真实或虚构内容。值得注意的是，幻觉是大语言模型的一个特定类型的错误，不能将其视为大语言模型的“全部问题”。

归因于其开放性应用场景、高度自由的生成能力，以及训练数据的多样性和潜在局限性，大语言模型的幻觉更加复杂。在缺乏明确任务约束或事实验证机制时，模型容易生成偏离输入语义的虚构信息或包含逻辑、事实错误的内容。此外，高度开放的任务场景进一步加剧了生成内容中幻觉现象的多样性和复杂性。

为了更加系统地描述和研究这些复杂的幻觉行为，现有研究对幻觉进行了分类。目前，对于幻觉的分类方式主要可归纳为3种：1)内在幻觉(intrinsic hallucination)和外在幻觉(extrinsic hallucination)[15,17-24]；2)忠诚幻觉(faithfulness hallucination)和事实幻觉(factuality hallucination)[17]；3)忠诚幻觉、事实幻觉和上下文冲突幻觉(context-conflicting hallucination)[16]。然而，这些分类方式在面对大语言模型的复杂应用场景时仍然存在一定的局限性，难以全面覆盖所有可能的幻觉类型[25]。

本文基于近期文献的研究成果[15-24]，结合大语言模型在NLP任务中的应用特点，从输入忠诚度和知识准确度2个维度，将幻觉细化为4类，分别为语义忠诚性幻觉(semantic faithfulness hallucination, SFH)、事实一致性幻觉(factual consistency hallucination, FCH)、上下文一致性幻觉(contextual consistency hallucination, CCH)和外部依赖性幻觉(external dependency hallucination, EDH)，如表1所示。

Table 1 Classification of Large Language Model Hallucination

表1 大语言模型幻觉分类

类型	划分维度	划分边界说明
语义忠诚性幻觉	输入忠诚度	输出偏离用户提问的核心意图或语义
事实一致性幻觉	知识准确度	错误的知识源于模型内部记忆
上下文一致性幻觉	输入忠诚度	输出未偏离用户语义，但内容存在上下文对话历史矛盾或不连贯
外部依赖性幻觉	知识准确度	错误的知识源于外部检索知识库

输入忠诚度反映模型在响应提示或对话历史时的语义对齐程度，该维度主要包含语义忠诚性幻觉和上下文一致性幻觉2类，其侧重点在于用户视角的输入对齐；而知识准确度则衡量模型输出与客观事实之间的一致性，关注生成内容是否违背真实世界知识，对应于事实一致性幻觉与外部依赖性幻觉，体现模型在事实维度上的知识可靠性。

为便于读者理解这4种幻觉的区别，对这4种幻觉进行形式化描述，并给出示例加以说明，如图1所示。

语义忠诚性幻觉是指大语言模型未能准确捕捉提示词的语义约束，从而生成偏离输入信息核心含义或用户意图的内容。当发生语义忠诚性幻觉时，可表示为

$$\exists x_t, \arg\max_{x_t} P(x_t|x_1, x_2, \cdots, x_{t-1}, \boldsymbol{p}) \notin \mathcal{T}(\boldsymbol{p}), \tag{3}$$

其中$\mathcal{T}(\boldsymbol{p})$是由提示$\boldsymbol{p}$所定义的可接受输出的合理集合，而$\arg\max_{x_t} P(\cdot)$是生成过程中实际选择的输出项。若生成的$x_t$不在提示$\boldsymbol{p}$所定义的合理集合$\mathcal{T}(\boldsymbol{p})$中，则说明模型未能忠实于提示语，从而产生语义忠诚性幻觉。例如，图1(a)中用户要求列出COVID-19的主要症状，但模型的生成内容与症状无关，明显歪曲了用户的实际需求。

事实一致性幻觉是指基于大语言模型内部知识生成的内容与客观事实存在不一致，包括包含虚假信息或误导性陈述等现象。设真实世界知识$\mathcal{K}$，则生成文本的条件概率分布可表示为

$$P(\boldsymbol{x}|\boldsymbol{p}) = \prod_{t=1}^{T} P(x_t|x_1, x_2, \cdots, x_{t-1}, \boldsymbol{p}, \mathcal{K}_{\mathrm{LLM}}), \tag{4}$$

其中$\mathcal{K}_{\mathrm{LLM}}$表示模型内部的知识。当$x_t$满足$x_t \notin \mathcal{K}$时，事实一致性幻觉发生。换言之，如果$x_t$不属于真实知识集合$\mathcal{K}$，则模型生成的信息与客观事实之间存在偏离。此类幻觉的示例如图1(b)所示，大语言模型

Fig. 1 Examples of hallucination in LLMs

图 1 大语言模型幻觉示例

错误地将电影《阿甘正传》中的主角描述为莱昂纳多·迪卡普里奥而非汤姆·汉克斯，这一结果显然违背了客观事实，与真实世界知识不符。

上下文一致性幻觉是指以模型已理解用户意图为前提，生成内容中上下文之间不一致的现象，表现为在文本流中前后信息矛盾或信息脱节。上下文一致性幻觉发生的条件可以形式化为

$$\begin{aligned}&\exists(t_1,t_2),\ t_1<t_2,\\&P(x_{t_2}|x_1,\cdots,x_{t_1},\cdots,x_{t_2-1})=0\wedge C(x_{t_1},x_{t_2})=0,\end{aligned}\tag{5}$$

其中$P(x_{t_2}|\cdots)=0$表示x_{t_2}与先前上下文完全矛盾，而$C(x_{t_1},x_{t_2})$表示x_{t_1}和x_{t_2}之间的逻辑关系。在图 1(c)中，用户明确表明自己是软件工程师，但在第 2 轮问答中，模型错误地将用户的身份理解为医生。这种错误反映了模型未能保持语境逻辑的一致性，进而导致上下文矛盾。

外部依赖性幻觉是指涉及外部检索增强生成机制时，由于外部信息检索错误或生成内容与检索信息之间存在冲突等原因，导致生成内容偏离事实的现象。随着检索增强生成技术(retrieval-augmented generation, RAG)在大语言模型中的广泛应用，外部依赖性幻觉的风险日益凸显[25-26]。为更精准地描述此类现象，本文提出外部依赖性幻觉的概念，以刻画因外部数据库支持引发的幻觉问题。这种幻觉的具体表现为：外部检索信息与真实知识相冲突，或生成文本$\boldsymbol{x}$未能准确映射检索信息$\boldsymbol{r}$。在给定提示$\boldsymbol{p}$和外部检索信息$\boldsymbol{r}$的条件下，生成文本的条件概率分布表示为

$$P(\boldsymbol{x}|\boldsymbol{p},\boldsymbol{r})=\prod_{t=1}^{T}P(x_t|x_1,x_2,\cdots,x_{t-1},\boldsymbol{p},\boldsymbol{r}),\tag{6}$$

其中$\boldsymbol{r}$为检索到的信息，可能包含关键信息r_j。发生外部依赖性幻觉的充分条件定义为 2 种情形之一：1)外部检索信息与真实知识相冲突，即$\exists r_j, r_j\notin\mathcal{K}$；2)外部检索信息与真实知识不冲突，但生成内容与检索信息存在语义冲突，即$\exists r_j,\ r_j\in\mathcal{K}$且$SemErr(x_t,r_j)=1$，其中$SemErr(x_t,r_j)$为语义冲突判定函数，当$SemErr(x_t,r_j)=1$表示$x_t$和$r_j$存在语义冲突。如图 1(d)所示，检索到的内容明确指出“妊娠 24～30 周为获取图像质量最佳的推荐时间窗口”，然而，语言模型生成的回答对这一具体时段进行了模糊扩展，未能准确复述检索内容，从而导致事实偏离。

语义忠诚性幻觉与上下文一致性幻觉的主要区别在于其错误来源的不同。具体而言，前者体现为模型未能准确解析用户显式提示中的语义意图，导致生成内容偏离任务核心要求，常表现为答非所问、主题脱离等问题；而后者则是在模型已基本理解当前提示的前提下，输出内容与历史对话或上下文语

境中的隐含信息发生冲突，常见表现包括身份错位、语义矛盾与语境割裂。而事实一致性幻觉与外部依赖性幻觉均关注模型生成内容与真实世界知识之间的偏差，两者的核心区别在于知识来源的不同：前者侧重于模型基于其内部记忆生成的内容是否违背事实；后者则聚焦在模型在调用外部检索知识库时，生成内容是否与客观事实存在不一致。

此外，由于检索信息通常以“上下文注入”的方式提供给语言模型，在出现外部依赖性幻觉时，生成内容与检索内容之间的不一致在形式上可能体现为一种上下文矛盾。然而，两者在本质上仍有所区别，即上下文一致性幻觉更侧重于模型对“内部语境”的维持能力，而外部依赖性幻觉则强调“外部信息整合”过程中的事实偏差。

综上所述，尽管不同类型幻觉在具体表现上可能存在一定交叉，但它们在偏差来源、语义对齐对象与建模目标上的侧重点各不相同。由此可见，厘清这一分类边界，是实现大语言模型幻觉精准检测的关键前提。

2 幻觉产生原因

传统的幻觉检测方法已难以满足当前大语言模型的需求，这一挑战促使检测范式发生显著转变。为有效检测大语言模型的幻觉现象，须结合大语言模型特点，深入分析幻觉的生成机制，明确其成因与产生过程，从而针对性地设计出更加精确的检测策略。对大语言模型从构建、预训练、微调、对齐到部署应用的全生命周期进行系统性梳理，是挖掘各阶段潜在幻觉诱因的基础和前提。为此，本文系统梳理了大语言模型的生命周期，并深入分析幻觉的生成机制与诱因，为构建先进的幻觉检测方法提供依据。

2.1 模型架构设计

大语言模型的架构设计主要基于 Transformer 及其衍生变体[27-28]，如 BERT（bidirectional encoder representations from Transformer）[29]，GPT（generative pre-trained Transformer）[30-33]，T5（text-to-text transfer Transformer）[34]，XLNet[35] 等。这些模型在多种 NLP 任务上展现了卓越的语言理解与生成能力。在架构设计过程中，幻觉现象的产生与模型规模、建模范式和注意力机制 3 个关键因素密切相关。

1）模型规模。基础模型的选择及其参数规模在很大程度上决定了模型对知识的捕获、表达及推理能力。Elaraby 等人[36] 指出，小规模开源模型由于参数受限，难以有效建模复杂知识、理解上下文和捕捉长距离依赖，因此在生成过程中更易导致“幻觉”现象的发生。例如，可以部署在消费级 PC（personal computer）上的大语言模型通常大小为 7 B 或 8 B，如 LLaMA 7 B[37]。相比于规模较大的模型，使用这些规模较小的模型可能会导致更多的幻觉[38]。

2）建模范式。除了模型规模，建模范式也是导致大语言模型产生幻觉的因素。基于 Transformer 架构的大型语言模型，如 GPT 系列和 LLaMA（LLM meta AI）系列等广泛采用单向自回归建模，主要应用于 NLP 任务。然而，这种因果语言建模范式使模型仅依赖时间序列，难以全面理解双向上下文，限制了知识整合与全局语义建模能力。这一局限性在一定程度上增加了生成内容语义脱节的风险[39]。

3）注意力机制。注意力机制是构建大语言模型时的重要考虑因素，直接影响模型的表达能力、计算效率以及信息处理方式[40]。不同注意力机制在上下文建模、长序列处理及复杂依赖关系捕获方面表现出显著差异[41]。全局注意力虽能覆盖完整序列，但在长序列中易出现权重稀释，导致对关键信息的聚焦能力下降，增加生成噪声与幻觉风险。例如，当模型处理长篇文章时，可能会因权重分布的稀释而忽略文中的核心论点，转而关注无关细节，从而影响文本生成的准确性和一致性。相比之下，软注意力机制通过连续加权求和来平滑地捕捉上下文信息，但随着序列增长，权重分布趋于均匀，同样削弱了对重要信息识别能力，进而加剧幻觉现象[42-43]。由此可见，通过分析模型内部的注意力模式，尤其是识别注意力失衡或分散的区域，有助于判断生成内容是否偏离了核心语义或事实依据，从而提升幻觉检测的准确性。

2.2 模型预训练

大语言模型预训练是指通过大规模无监督文本数据学习语言的语义、语法及知识结构，以构建具备泛化能力的基础模型的过程[15]。由此可见，大语言模型的能力基础主要来源于预训练数据[44]。显然，数据质量对于模型性能具有重要的影响[45]。低质量数据往往可能引入偏差、噪声和幻觉现象，削弱模型生成内容的准确性与可靠性。因此，为了更深入理解模型预训练过程中的幻觉诱因，从真实性、平衡性、偏见性、及时性和专业性等角度探讨数据质量对幻觉产生的影响机制。

1）虚假数据。由于预训练数据来源广泛并缺乏较为严格的数据审核与验证机制，这些数据中不可

避免地存在未经验证的虚假信息、主观观点及错误知识[15-16]。利用此类数据进行预训练将影响模型的知识表示能力,进而引发幻觉问题。研究表明[46],训练数据中错误样本的存在会显著影响词元的贡献分布,最终导致模型输出内容偏离事实,产生幻觉。基于这一发现,Filippova[47]从训练数据中删除与事实不符的错误样本,显著减少了幻觉的发生。

2)重复数据。大语言模型在预训练阶段具有记忆训练数据的内在倾向,且该特性随着模型规模的增加而变得更为显著[48-50]。当预训练数据呈现不平衡特征,存在大量重复或冗余信息时,模型可能过度拟合这些重复特征,而非通过抽象和泛化来理解语言的深层语义结构[51-52]。这导致模型在生成内容时优先回忆训练中反复出现的信息,而忽略上下文的实际需求,导致生成的内容偏离事实或缺乏逻辑[53]。

3)偏见数据。数据偏见性指的是训练数据中包含了某种偏向或倾向,如性别歧视、表述方式倾向等。当包含偏见的数据被模型在预训练阶段学习并记忆,会影响生成内容的客观性与准确性,导致幻觉现象的发生[54-55]。例如,Wan 等人[56]发现,大型语言模型中的性别偏见会在推荐信生成任务中显现。具体而言,ChatGPT 在生成推荐信时,对男性表现出“有领导能力”的特质,而对女性则表现出“受欢迎”或“善于合作”的特质。这种现象与训练数据中社会性别刻板印象的存在密切相关。

4)过时数据。预训练数据往往反映的是模型训练时的历史信息,缺乏对实时知识的动态更新。由于大语言模型本身无法自动更新知识库,生成内容时难以对过时信息进行验证或修正。当面对需要动态更新知识或时效性要求高的任务时,模型会在无约束条件下填补信息空白,凭借其统计关联性生成过时的内容或捏造事实,从而引发幻觉[57-58]。

5)专业知识匮乏。此外,预训练数据中专业领域的知识覆盖不足,导致通用模型在处理垂直领域任务时面临领域知识表示不准确等问题。即使经过后续微调,模型在医学[59]、法律[12]、金融[60]等专业性强的领域中,仍难以基于事实逻辑生成精准且可靠的内容,幻觉现象尤为突出。

2.3 模型微调

由于预训练主要依赖于下游任务无关的粗粒度数据,模型在特定任务中的适配性较弱,难以直接解决实际问题。因此,为提升模型在具体任务上的表现需要进行微调。然而,微调过程通常仅在局部层面调整参数,难以全面重构模型中已存储的知识表示,导致旧知识与新知识之间的冲突未被有效解决,从而引发幻觉现象[61-63]。

2.4 模型对齐

尽管经过微调的模型在特定任务上表现出色,它们仍可能生成与人类期望、需求或价值观不符的内容。因此,为了提高模型的安全性、可靠性和伦理规范性,需要进行模型对齐。大语言模型的对齐技术通常采用强化学习与人类反馈相结合的方式,将人类反馈纳入到模型优化过程中,以指导大语言模型输出高质量且无害的内容[4]。对齐过程主要包括 2 个关键阶段,即监督微调(supervised fine-tuning, SFT)和基于人类反馈的强化学习(reinforcement learning with human feedback, RLHF)。

1)能力边界有限。在 SFT 阶段,通过高质量的人工标注数据对模型进行微调,使其行为更贴近人类期望,实现定向优化[4]。与模型微调类似,如果对齐数据的需求超出当前模型的能力边界,模型有可能生成超出自身知识范围的内容,从而增加幻觉产生的风险[64]。例如,使用模型生成医疗诊断的相关建议时,尽管对齐数据要求模型给出专业化的意见,但由于预训练和微调阶段,模型并未学习某些特定的伦理规范,即能力边界有限,则易生成违背医学伦理的有害内容[65-66]。

2)信念错位。在 RLHF 阶段,通过收集人类反馈来构建奖励模型,并利用强化学习优化模型行为,使其生成更符合人类期望和需求的内容。然而,近期研究表明,经过 RLHF 对齐的模型有时倾向于安抚用户,以牺牲真实性为代价,表现出迎合用户意见的行为[67-68],这种现象被称为“信念错位”或“阿谀奉承”[69]。其原因可能在于 RLHF 中使用的奖励模型过于强调用户满意度,导致模型优先考虑生成令人信服的谄媚回复,而非事实正确的回复[70]。

2.5 模型推理

模型完成训练与对齐后,进入推理阶段,响应用户输入并生成文本。鉴于外部环境的不可追溯性及其对特定任务的高度依赖,模型的输出可能受到多重因素的干扰,本节重点探讨大语言模型在推理过程中的特点,剖析模型推理阶段主要存在的随机采样、错误积累和过度自信等问题。

1)随机采样。如前所述,采样和编码是模型推理阶段的关键环节。采用具有更大不确定性的采样算法可能会使大语言模型更容易产生幻觉[71]。例如,在核采样中,较高的概率阈值扩大了词汇选择范围,增加生成多样性的同时,也提高了选择低概率词的

可能性，易导致上下文逻辑或事实偏离。此外，采样温度是影响幻觉产生的另一个重要因素。Renze 等人[72]在多项选择题问答任务中发现，GPT-3.5 温度从 0 上升至 1.0 时性能变化不大，但超过 1.0 后准确率急剧下降至 0。研究表明，模型的解码策略通常难以在事实性和多样性之间取得平衡，采样随机性的增加虽提升了内容多样性，却同时加剧了幻觉现象[70-73]。值得注意的是，采样引入的不确定性不仅是幻觉产生的诱因，也为幻觉检测提供了重要线索。通过熵值等不确定性指标可以有效度量生成内容的真实性。

2）错误积累。在推理阶段，当模型输出错误内容时，它往往倾向于维持生成内容的一致性，而非修正错误。这种现象导致最初的错误信息被逐步扩展并反复强化，进而加剧幻觉的程度，形成所谓的“幻觉滚雪球”效应[74]。其根本原因在于模型生成内容时高度依赖上下文的连贯性与一致性，往往未能有效区分错误信息与正确内容，并且缺乏强有力的事实性校验机制[75-81]。长思维链（long chain-of-thought，Long CoT）方法被提出，引入“反思-修正”机制[82]，引导模型在生成过程中自我检测并修正推理链中的错误，从而在一定程度上缓解“幻觉滚雪球”现象。

3）过度自信。此外，大语言模型在感知其事实知识边界方面存在一定困难，并倾向于表现出过度自信。Ren 等人[83]通过先验与后验判断分析发现，模型在回答前高估自身能力，回答后亦倾向于认为答案正确，且自评正确率与实际准确率存在显著偏差。Wen 等人[84]进一步揭示了大规模模型和小规模模型在置信度估计上的行为差异，其中大语言模型（如 LLaMA-3-70 B）在简单任务上倾向于低估自己的表现，而在困难任务上倾向于高估自己的表现；相比之下，小模型（如 Gemma2-9 B）在不同任务中均表现出一致的过度自信。该现象表明，置信度可作为幻觉检测的重要特征，通过识别置信度异常的输出模式，以识别潜在幻觉。

此外，鉴于大语言模型存在知识更新滞后等问题，研究中常引入 RAG 机制。然而，即使利用 RAG 等技术，模型仍有可能会产生毫无根据或与检索到的参考文献中提供的信息相矛盾的陈述[15,85]，尤其是针对侧重于生成内容多样性的 NLP 任务。这与 RAG 检索质量低有关。具体而言，检索系统可能返回与输入问题相关性较低或覆盖不足的文档，导致模型在生成答案时基于不完整或错误的语境，增加幻觉的发生。亦或模型在整合多个检索文档时，未能正确调和知识冲突，从而导致生成内容不一致或不准确[86-88]。

图 2 总结了大语言模型在从设计到推理的全生命周期中可能诱发幻觉的关键节点和影响因素。表 2 展示了该坐标系中不同诱因对模型能力的影响。例如，“重复数据”可能对大语言模型“拥有知识”的能力造成较大的负面影响，而“知识冲突”则影响模型的“拥有知识”“理解问题”“表达知识”能力。

Fig. 2 The cause of the large language model hallucination generation

图 2 大语言模型幻觉产生的原因

Table 2 Analysis of the Hallucinations Causes

表 2 幻觉诱因分析

模型能力	※	■	⋆	▲	△	▼	▽	◆	◇	◎	○	★	☆	⊙	●	◢
拥有知识	×	×	×	×	×	×	×	×	×	×					×	×
理解问题	×	×	×								×					
表达知识	×	×	×								×	×	×	×		

注：第 1 行符号含义如图 2 所示，“×”表示诱因通过影响模型的该能力导致幻觉的产生。

由表 2 可见，幻觉的产生源于多种因素的交互作用，具有难以追溯、复杂多维、强耦合的特点，难以从单一维度剖析其最终成因。这将导致在大语言模型的推理过程中，尤其是在长文本生成、多轮对话等复杂任务中可能出现多种类型幻觉叠加的现象。

目前，尚缺乏对该现象的系统定义与深入研究。本文基于现有研究，将其初步归纳为复合幻觉（composite hallucination）。复合幻觉是指多种类型的幻觉在同一生成片段中同时存在、交叉叠加的现象，表现为语义偏离、事实错误与语境冲突等问题的混合共现。图 3 展示了大语言模型在回答中同时出现语义忠诚性幻觉和事实一致性幻觉的示例。具体而言，用户询问屈原的文学成就及其是否参与了统一战争，模型却错误地将屈原描述为军事家并参与秦国统一六国，既偏离了提问意图，又与历史事实不符。

3 幻觉检测方法

大语言模型的幻觉问题不仅削弱了模型的可信度，还可能导致用户产生错误认知。因此，幻觉检测

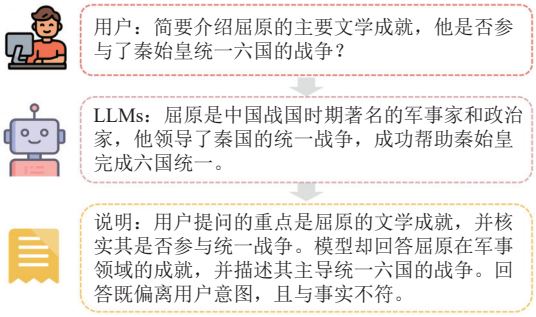


Fig. 3 Examples of composite hallucination in LLMs

图 3 大语言模型复合幻觉示例

已成为确保大语言模型可靠、可信的关键技术。本节梳理了相关文献，综述了面向大语言模型的幻觉检测方法。

现有研究表明，幻觉类型、模型应用场景及用户需求的不同，使得幻觉检测的步骤和方法存在显著差异[16-17]。立足于实际应用需求，考虑到在不同任务场景下大语言模型透明度的差异，以及检测者对模型内部信息的访问能力，将幻觉检测方法划分为针对白盒模型的检测方法和针对黑盒模型的检测方法。

由于复合幻觉具有类型多源与特征叠加的复杂性，其检测面临更高难度。现有研究主要针对单一类型幻觉进行识别。因此，本文所综述的幻觉检测方法均针对单一幻觉类型。

3.1 针对白盒模型的幻觉检测

白盒模型提供了较高的透明性和可解释性，通过访问模型内部状态，能够深入了解大语言模型的推理过程和生成机制。针对白盒模型的幻觉检测方法则是利用模型生成并输出文本的过程中所产生的状态信息来识别生成文本中存在的幻觉，包括隐藏层激活值、logits 值、熵值、注意力权重和梯度。针对白盒模型的幻觉检测框架如图 4 所示。

1)隐藏层激活值。隐藏层激活值是模型内部状态的核心数据，直接影响模型的性能和输出质量。因此，Rateike 等人[89]对隐藏层激活值进行分布差异的统计检验，计算幻觉文本与非幻觉文本的激活分布之间的偏离程度，以检测生成文本中是否存在事实一致性幻觉。具体而言，该方法通过分析隐藏层激活值的左尾、右尾和双尾分布，定位异常的激活单元及其对应的输入特征，从而捕获可能导致幻觉的中间表示模式。模型深层包含更多的事实信息，而浅层分布中噪声更多[90]。利用这一特点，Chuang 等人[91]将浅层和深层的隐藏层激活值投影到词表空间以生成 pseudo logits，通过对比数值差异检测生成内容是否偏离模型内在知识，即是否存在幻觉。深层与浅层的分布差异越显著，说明生成内容存在事实性幻觉的概率越大。

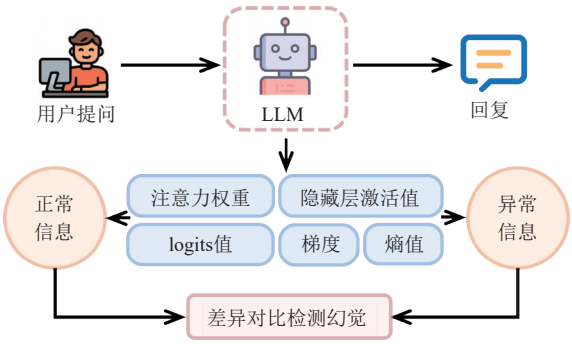


Fig. 4 Hallucination detection framework for white-box models

图 4 针对白盒模型的幻觉检测框架

2)logits 值。logits 值是语言模型在生成过程中，输出层对词汇表中每个候选词所计算的未归一化分数。它表示模型在当前生成状态下对候选词的“相对倾向性”或“原始置信度”。通过对 logits 值的分析与归一化处理，能够获得生成序列的概率分布、不确定性及置信水平，从而为幻觉检测提供重要依据。

聚焦检测高风险幻觉内容可以在生成过程的早期阶段纠正错误，从而防止这些错误在后续生成中累积。为了高效识别潜在幻觉，Varshney 等人[92]通过识别文本中的高权重词汇来缩小幻觉检测范围，并利用 logits 输出值计算每个关键概念的生成概率值，选取最小值作为该概念的不确定性得分。然而，该方法仅关注单词级或概念级的不确定性，只依赖词汇权重可能无法充分捕捉上下文层面的语义偏差。Chen 等人[93]通过分析大语言模型在每个位置生成词元时的输出概率分布，利用最大概率等统计特征衡量其生成决策的置信度，从而估计单个词元发生幻觉的潜在风险。

序列的对数概率能够有效反映生成序列的累积置信度。较低的对数概率通常表明模型在生成过程中多次选择不确定性高的词，反映出模型对生成内容缺乏明确支持。在这种情况下，生成内容更可能包含事实一致性幻觉。然而，由于较长的序列会累积更多的对数概率项，其整体对数概率值通常偏低，这种现象可能会引起模型不确定性评估结果的偏差。为解决这一问题，许多研究[94-96]引入了长度归一化策略，通过对对数概率进行归一化处理，消除序列长度对不确定性评估的影响，从而提高针对不同长度的生成序列的幻觉检测准确率。

3)熵值。熵值被用于评估语言模型生成文本的

质量和多样性。高熵往往意味着生成过程存在较大的不确定性，模型对当前输出的置信度较低，这种情况下生成的内容更可能偏离事实，进而导致幻觉。鉴于此，Xiao 等人[97]对生成序列中每个词的熵值进行分析，发现高熵值往往对应事实一致性幻觉内容，从而能够根据熵值检测幻觉。同样，Su 等人[98]提出了一种实时的事实一致性幻觉检测方法，通过对生成内容中的命名实体进行概率和熵分析，能够精确定位幻觉来源。而 Van Der Poel 等人[99]进一步验证了，当条件熵较高时，模型更可能生成与源文档内容不一致的幻觉内容。因此，他们利用模型的预测概率分布计算条件熵，进而检测语义忠诚性幻觉。此外，还通过条件熵动态调整解码策略，在高不确定性时减少幻觉生成，显著提升了生成内容的忠诚性。

4）注意力权重。除此之外，注意力权重也是能够反映模型内部状态的信息。Chuang 等人[100]假设幻觉通常发生在模型更多关注其生成内容而非提供的上下文时。基于这一假设，他们提出了一种基于注意力权重的幻觉检测方法。其原理是通过计算模型生成过程中每层每个注意力头的注意力权重分布，得到上下文与生成词元之间的注意力权重比例，以量化生成内容对上下文的依赖程度，从而识别幻觉内容。Sriramanan 等人[101]通过分析单个响应的注意力图变化，设计了轻量、高效的检测指标，实现了在无需多次采样条件下的幻觉检测。

5）多维特征。以上研究均基于单一特征进行幻觉检测，但由于单一特征难以全面表征幻觉的复杂性，在信息利用丰富度上存在一定的局限性。通过结合多种特征进行幻觉检测，可以弥补单一特征对幻觉类型覆盖不足的问题，进而提高检测的鲁棒性和准确性。例如，Zablocki 等人[102]利用隐藏层激活值与注意力权重 2 种特征，识别其中偏离正常文本的显著模式。基于此，通过采用回归分析和主成分分析等统计方法，建立内部状态与事实一致性幻觉风险之间的相关性模型，进而实现事实一致性幻觉检测。考虑到梯度特征能够捕获模型对输入敏感性的细粒度变化，Hu 等人[103]融合隐藏层激活值与梯度的双重特征，以建模生成内容与提示词之间的相关性。实验表明，该方法在提高语义忠诚性幻觉检测精度方面表现出显著优势。此外，Snyder 等人[104]提取了 4 种与模型生成相关的特征，用于训练事实一致性幻觉检测的分类器，包括 Softmax 概率分布、特征归因分数、自注意力分数和全连接层激活值。但实验表明不同特征的组合并未显著提升分类性能。其原因可能在于不同特征之间可能存在较大的信息冗余，导致组合特征未能为分类器提供额外显著的信息增益。对于外部依赖性幻觉，ReDeEP 利用注意力权重和 logits 值，解耦了外部上下文和参数知识对内容生成的贡献来检测幻觉[105]。

总体来看，针对白盒模型的幻觉检测方法通过深入分析内部状态，实现了较高的透明度和细粒度检测能力。然而，其检测效果受限于模型特征选择和上下文建模能力。同时，实时分析隐藏层激活值等特征对计算资源要求较高，可能不适用于大规模或高效的应用场景。

3.2 针对黑盒模型的幻觉检测

出于对技术保护、风险控制和资源管理等因素的综合考虑，GPT-4[32]，Gemini[106]等大语言模型普遍采用闭源策略，仅通过 API 接口对外提供服务，导致针对白盒模型的幻觉检测方法将不再适用。而针对黑盒模型的幻觉检测方法不需要了解模型的内部结构，能够适用于各种类型的大语言模型，具有高度的灵活性和可拓展性等优点。

针对黑盒模型的幻觉检测以外部验证为核心，旨在不依赖模型内部结构或参数信息的情况下，通过分析输入与输出的映射关系并结合外部知识源评估生成内容的真实性、逻辑性和一致性。根据外部资源的可用性，将幻觉检测方法进一步细分为零资源（zero-resource）和非零资源（non-zero-resource）两种类别。

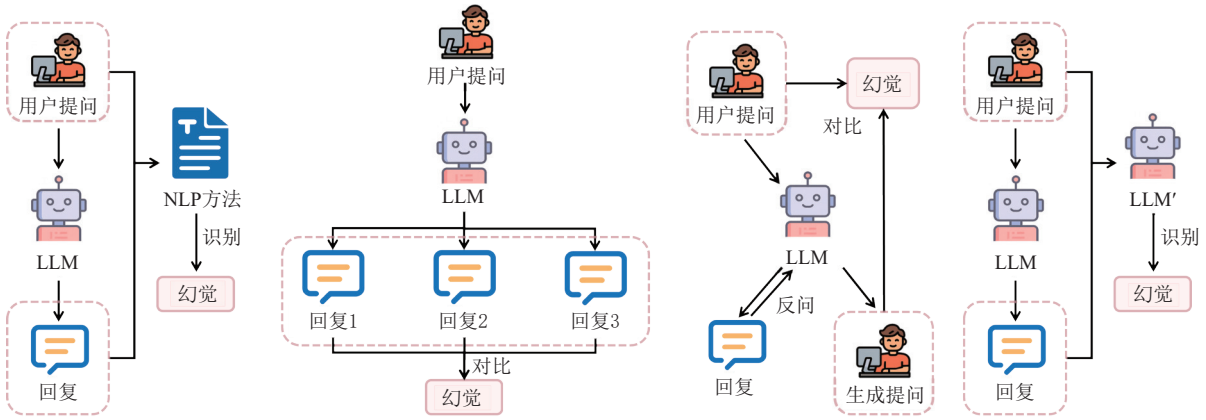
3.2.1 零资源幻觉检测方法

“零资源”一词表示没有用于验证的外部资源或辅助工具。因此，零资源幻觉检测方法是指在不依赖外部知识库或数据源的情况下，通过分析大语言模型的输入与输出，检测生成内容是否存在幻觉。此类方法强调利用输入设计或模型本身的信息来完成检测，而非引入外部知识支持。针对黑盒模型的零资源幻觉检测框架如图 5 所示。

1）传统方法迁移

由于幻觉是 NLP 中的共性问题，且传统 NLP 任务中幻觉检测方法本身具备一定的通用性，因此可以将传统方法的检测思路迁移至大语言模型上。

Bhamidipati 等人[107]创新性地将幻觉检测任务形式化为自然语言推理（natural language inference, NLI）任务，通过使用 NLI 模型来检测输入和输出之间的单向蕴涵和双向蕴涵等语义关系。这使其能够深入分析生成文本与源文本之间的语义一致性，而不仅仅依赖于表层相似度。此外，Rashad 等人[108]将输入



(a) 传统方法迁移 (b) 生成多次响应 (c) 反向验证 (d) 基于大模型

Fig. 5 Zero-resource hallucination detection framework for black-box models

图 5 针对黑盒模型的零资源幻觉检测框架

文本和生成文本分别构建为知识图谱，并进行知识图谱对齐，以衡量生成内容与输入内容的语义一致性。然而，当生成大规模文本时，计算成本显著增加。为此，Sansford 等人[109]提出了一种幻觉检测方法——GraphEval，仅需调用 1 次大语言模型来构建知识图谱，显著提高了检测效率。同时，该方法将知识图谱和 NLI 模型相结合，能够直接检测幻觉发生的具体位置，增强了幻觉检测的可解释性。此外，Durmus 等人[110]提出通过对摘要句子中关键信息进行掩蔽，生成对应的“标准”答案，将掩蔽后的句子转换为自然语言问题。然后使用预训练的问答模型从原文中提取答案，并与摘要生成问题的“标准”答案进行匹配，根据答案匹配的准确性，计算 $F1$ 分数作为摘要的忠实度分数。如果模型从原文中提取的答案与摘要中的答案匹配较差，说明摘要中可能存在幻觉信息。

即便如此，利用传统幻觉检测方法识别大语言模型的幻觉仍存在泛化能力差等不足。传统方法通常针对特定任务设计，难以适应大语言模型多样化的应用场景，表现出较弱的跨领域和跨任务泛化能力。由此可见，仅依赖传统方法难以满足大语言模型在广泛生成任务中的实际需求，亟需设计更加通用且高效的幻觉检测方法，以适应更多样化的任务。

2)生成多次响应

除了传统的幻觉检测方法外，还可以利用大语言模型支持多轮问答的特性，通过设计系统化的问答交互流程，以有效检测生成内容中的幻觉。

一类代表性的方法是通过多次生成响应，分析样本一致性以评估生成内容的可靠性。其原理在于当一个语言模型对某主题或事实信息具有较强的内在知识或支持时，通过随机采样生成的响应内容应在语义、事实或逻辑等维度均具有较高的一致性；相反，如果生成内容涉及幻觉，模型在多次采样生成的结果之间往往会存在显著差异。基于这一思路，衍生出了许多幻觉检测方法。

Farquhar 等人[111]从语义熵的角度分析多次随机采样生成的响应在语义上的相似性，进而检测生成内容的语义一致性幻觉。语义熵是度量生成的文本输出分组为语义等价类分布情况的指标。该方法解决了传统熵计算不适合语言生成任务的问题。而 Manakul 等人[112]通过调整随机种子或使用 Top-K 采样等多种解码策略生成多次采样响应。然后，构造了 5 种变体分别从语义、事实和逻辑等不同维度来检测问答任务中的幻觉。此外，Elaraby 等人[36]提供了一种轻量化检测方法——HALOCHECK，即对多次采样生成的内容进行句子级别的蕴涵分析。通过捕捉这些样本之间的冲突量化生成样本之间的一致性，一致性得分越低，幻觉风险越高。针对长文本的幻觉检测方法往往将长文本分成多个事实，并单独比较每对事实的一致性。然而，这些方法很难实现多个事实之间的对齐，且忽略了多个上下文事实之间的依赖关系。Fang 等人[113]提取生成文本的知识三元组并建模图结构，捕捉三元组之间的依赖关系，以增强对多重上下文的建模能力。

3)反向验证

此外，还有一类方法在正向生成答案后进行了反向生成，即通过生成的答案重新构造查询，并评估生成的查询与原始查询的一致性。应用此类方法的前提是语言模型的参数存储了实体及其相关知识。通过将模型生成的内容转化为查询语句，验证模型是否能返回与初始生成内容一致的实体。若生成内

容包含幻觉，转化为查询时将导致搜索条件错误，从而无法检索到正确的实体。

Yang 等人[114]设计了 2 种反向验证方法，即基于问题生成的反向验证和基于实体匹配的反向验证。前者通过提示语言模型，根据生成内容构造一个问题，要求模型回答该问题并返回实体，判断返回的实体是否与原始实体一致；后者将生成内容中的信息重写为一系列特征要求，并提示模型返回符合这些要求的实体，要求模型报告返回实体与要求的匹配程度，如果匹配度低于设定阈值，则判定生成内容为事实幻觉。类似地，文献 [115−118] 生成问题再回答问题，利用 NLI 模型或 $F1$ 分数等方法评估答案的一致性。InterrogateLLM 方法[119]通过正向生成与反向验证的双向机制，使用嵌入模型将原始查询和重构查询转换为向量，计算原始查询与重构查询之间的余弦相似度，进而检测幻觉。受模拟法律中交叉质询机制的启发，Cohen 等人[120]基于交叉询问原理的零资源黑盒幻觉检测方法，将语言模型生成的事实性检测建模为 2 个模型之间的交互，其中一个语言模型生成声明或回答，另一个语言模型通过提问来验证这些回答。为提升生成内容中知识三元组的一致性验证能力，Fang 等人[113]围绕每个知识三元组（头实体、关系、尾实体）提出了 3 项子任务：基于问题生成的头实体验证、关系重建以及尾实体选择。通过构建反向重建-验证机制，该方法实现了对生成知识的细粒度一致性校验。

上述方法具有操作简单、可拓展性强等优势，能够解决传统一致性检测中可能出现的“遗漏问题”。但是，无论是多次生成采样还是反向验证，这些方法都对计算资源要求较高，尤其是在参数较多的大语言模型和大规模数据集上进行应用时，可能面临性能瓶颈。

4）基于大语言模型的检测

由于大语言模型参数中编码了广泛的事实知识，大语言模型常被作为事实检查的工具。同时，大语言模型具备较强的指令跟踪能力，即根据用户提供的具体指令完成相关任务的能力[121-122]。结合这一特性，模型不仅可以生成内容，还能够对自身生成的内容进行检查和评估[123-125]。

Gao 等人[124]使用 ChatGPT 等大语言模型作为自动化评估工具，通过提供详细的任务指令和评分标准，使模型能够模拟人类的评估过程，检测生成内容中是否存在事实一致性幻觉。类似地，Adlakha 等人[123]验证了大语言模型作为自动评估工具的潜力。具体而言，通过为大语言模型提供清晰的指示，如明确的评分标准和任务背景，结合模型生成的内容与知识源内容，模型能够生成与人类评估高度一致的评估结果。然而，Adlakha 等人也指出，当大语言模型作为评估工具时，其性能仍然受到任务复杂性和输入质量的影响，例如在高多样性或语言模糊性的任务中，模型可能出现评估偏差。为了进一步提高检测的准确性，Jain 等人[125]通过选择具有代表性的上下文示例并嵌入到提示中，使大语言模型能够模仿人类的评分模式，对生成文本进行高效且准确的打分。实验表明，大语言模型在一致性和相关性评估方面表现出与人类标注者高度相关的性能，并能够捕捉生成内容中的事实性错误和逻辑缺陷。

在提高幻觉检测可解释性方面，通常是在检测过程中结合思维链或验证链（chain of verification）等逐步推理法，显式地分解复杂任务或逐步验证生成内容的逻辑一致性和事实性。例如，Luo 等人[126]为 ChatGPT 提供源文档和生成摘要，并结合 CoT 技术[80]，引导模型逐步推理，通过详细解释生成推理过程后作出判断，检测生成内容是否存在事实一致性幻觉。Dhuliawala 等人[127]引入了验证链，根据查询和初始回答生成一系列验证问题，以检测回答中的事实性错误。Luo 等人[128]从输入指令中提取出核心概念，要求模型对核心概念进行解释与推理，并量化模型对概念的熟悉度，以此度量模型输出内容的不确定性，从而预防潜在的事实一致性幻觉生成。

此外，Agrawal 等人[129]通过间接查询（indirect queries）进行幻觉检测。其核心思想是通过提出开放性问题而非直接验证问题来检测幻觉。与直接查询相比，间接查询通过生成多个关于引用内容的细节性回答，能够捕捉更为复杂的幻觉现象。单一模型可能对特定任务或数据有偏差，通过集成不同模型进行验证[130]，能够减少这种偏差的影响。

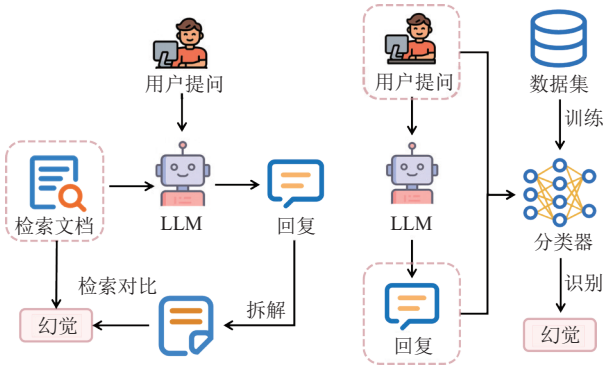
尽管此类方法自动化程度高、易于实现、拓展性强，但是高度依赖于语言模型本身的推理能力和生成质量。如果模型内部知识库存在漏洞或矛盾，这些方法可能无法有效识别幻觉。

综上所述，在外部资源不可用、不切实际或计算成本高昂的条件下，零资源幻觉检测方法具有明显的优势。此类方法提供了一种轻量化且具可扩展性的解决方案。然而，由于缺乏外部知识支持，此类方法高度依赖输入设计的质量及大语言模型的语言分析能力。当输入设计过于复杂或模型的表达能力不足[131]时，检测结果可能无法有效识别复杂或隐性的

幻觉。

3.2.2 非零资源幻觉检测方法

非零资源幻觉检测方法依赖于外部知识源或辅助工具,通过引入知识库等外部资源,对生成的内容进行幻觉识别。本节将非零资源幻觉检测方法划分为基于外部数据库的检测方法和基于分类器的检测方法2类。针对黑盒模型的非零资源幻觉检测框架如图6所示。



(a) 基于外部数据库 (b) 基于分类器

Fig. 6 Non-zero-resource hallucination detection framework for black-box models

图6 针对黑盒模型的非零资源幻觉检测框架

1)基于外部数据库的检测方法

基于外部数据库的检测方法通过将生成内容与外部数据库进行比对验证,以识别生成内容中可能存在的幻觉现象。

生成内容通常需经过预处理,以提高其与外部数据库匹配分析的准确性和有效性。这一环节尤为重要,因为未经处理的生成内容往往存在较大差异,难以直接与知识库对比分析。为此,一些研究致力于处理生成文本,以优化匹配效率并提升分析可靠性。例如,Son等人[132]通过建模幻觉风险,定量评估生成内容与知识库之间的偏离程度。然而,生成文本通常较为冗长,并缺乏对具体事实明确而精细的定义粒度,同时在事实核查过程中也面临证据不足的问题。为应对上述挑战,Min等人[133]首先将生成文本逐条分割为原子事实单元,并利用搜索系统从指定知识库中提取相关证据。随后,他们逐一验证每个事实单元的支持性,从而实现了细粒度的事实一致性幻觉检测。类似地,Chern等人[134]开发了一种与任务和领域无关的框架FacTool。该框架通过将生成内容分解为独立的事实性声明,结合外部知识库的检索、文本相似度匹配以及NLI技术,逐一对文本和事实进行匹配分析,从而检测生成内容中的幻觉。此外,Refchecker框架[135]引入三元组提取方法,将复杂文本解构为语义独立的单元,逐项验证三元组是否受参考文献支持。与此同时,Mishra等人[136]从幻觉检测的分类视角出发,提出了细粒度的幻觉分类方法,涵盖多种复杂错误类型。该方法依托外部知识库检索与对比,并结合编辑模型,对错误进行精准检测与修复,进一步完善了生成内容的质量控制机制。Li等人[137]使用逻辑推理规则对知识库中的事实进行转换和扩展。这种逻辑推理能够构建更复杂、更有覆盖力的推断事实,以用于生成更多场景化的测试用例。

除此之外,外部数据库的质量对幻觉检测效果至关重要。为了提高幻觉检测精度,Bayat等人[138]提出了一种依赖外部知识库和网络检索的检测方法。该方法利用知识图谱进行结构化查询,获取直接支持的高置信度证据,并通过网络检索补充知识图谱不足的事实验证。这种双重验证策略结合了结构化知识和动态开放领域信息,使得能够在多种生成任务中有效检测事实一致性幻觉,并为修订提供依据。以上方法通常假设检索到的证据是可靠的,并且在分析时未细分证据类别,可能导致误判。考虑到这一点,Halu-J通过证据分类过滤掉完全无关的内容,提取部分无关证据中的有用部分,并深入分析高度相关证据[139]。对于误导性证据,模型设计允许其被理解为“高度相关但混淆性的内容”,避免误判的同时提高了检测的鲁棒性。

现有基于外部数据库的检测方法往往依赖单一知识源,导致知识覆盖不足、无法处理多种证据类型等问题。为此,Zhao等人[140]通过从多个证据源中检索和整合信息,提高了检测结果的准确性和可靠性。Zhang等人[141]通过结合多类型知识,解决了知识来源单一的问题,增强了检测的泛化性。通过利用结构化或可靠的外部知识库,显著提高幻觉检测能力,尤其在开放领域和高复杂性任务中提供权威、可验证的依据。然而,当前的幻觉检测方法通常需要检索大量相关证据,存在响应时间过长的缺点。为了降低计算成本,Wang等人[142]基于贝叶斯序贯分析,通过逐步检索文档实时评估当前证据是否足够,从而动态决定是否继续检索更多文档在同等准确率的前提下减少了平均检索文档数。

2)基于分类器的幻觉检测方法

基于分类器的幻觉检测方法是通过构造合适的幻觉数据集对分类器进行训练,利用训练好的分类器进行幻觉检测。目前研究主要从构造高质量和多样化的训练数据集以及改进分类器等维度来提升分

类器的性能。

在构造训练数据集方面，一部分研究关注如何进行高质量的数据标注。Zhou 等人[143]通过合成数据生成和人工标注 2 种方式构建用于训练幻觉检测分类器的数据集。为了提高数据标注的效率，Wojciech 等人[144]提出了一种自动生成标注数据的方法，从源文档中提取句子，并通过语义变换生成事实正确和事实错误的数据。为了检测跨语言摘要任务中生成内容的事实性幻觉，Qiu 等人[145]提出了一种基于多语言信实性度量的幻觉检测方法。具体而言，该方法通过提取“文档-摘要”对的关键事实，利用英语信实性度量工具标注忠实性分数，并将标注后的数据集翻译到目标语言，生成多语言训练数据集。随后，基于多语言 BERT 的分类器被用来检测跨语言摘要的信实性分数，从而量化文本是否包含幻觉，解决了现有幻觉检测方法在低资源语言中标注不足、跨语言生成文本检测难以泛化的问题。

还有一部分研究关注如何生成高质量的幻觉数据和非幻觉数据，以训练更精准的幻觉检测分类器。HaloScope[146]结合了嵌入分解和二分类器，通过无标注的大语言模型生成数据来实现高效的幻觉检测。其核心方法建立在对嵌入空间中幻觉子空间的发现和利用上，为事实一致性幻觉检测提供了一种新思路。Quevedo 等人[147]通过对生成文本概率分布的分析，提取最低词元概率、平均词元概率、最大概率偏差和最小概率扩散 4 个关键特征，结合监督学习对分类器进行训练。这种处理方法具有泛化能力强等优势，适用于多种语言模型和生成任务。Cao 等人[148]针对摘要生成任务中的实体幻觉检测问题，提出了一种基于先验和后验概率的检测方法。该方法利用无条件掩码语言模型计算实体的先验概率，即在不考虑源文档的情况下，实体出现在生成摘要中的可能性；同时通过条件掩码语言模型，结合源文档信息计算实体的后验概率，表示实体在上下文和源文档支持下的生成概率。基于先验概率和后验概率的差异，训练了一个 K 近邻分类器，以区分给定实体的幻觉和事实状态。Santhanam 等人[149]通过数据增强技术，如随机配对、否定、实体替换，生成了包含事实一致和事实不一致的对话响应，用于训练和测试检测模型。

在改进分类器方面，许多研究创新了分类器的形式，从不同维度、粒度进行幻觉检测。Zha 等人[150]提出了一种统一的对齐评估函数，用于评估生成文本与输入文本之间的语义、事实和逻辑一致性。具体而言，将文本对齐视为一个连续对齐分数的计算过程，通过上下文切块对长文本进行分割，逐句评估生成文本中每一句话与输入文本的对齐程度。对齐分数低的句子通常被认为是幻觉内容。为了解决新闻标题生成任务中标题与新闻内容不一致的幻觉问题，Shen 等人[151]提出了一种基于 NLI 和自然语言解释（natural language explanation）的幻觉检测方法——ExHalder。该方法通过将标题与新闻内容之间的关系建模为 NLI 任务，构建一个统一的推理分类器以评估标题是否被新闻内容支持，同时通过生成自然语言解释增强分类器性能。而 Choi 等人[152]使用蒙特卡洛树搜索模拟未来生成路径，计算每个路径的知识一致性得分，通过评估当前与未来的综合一致性引导最优解码策略。同时，通过训练分类器检测生成序列中幻觉的起点，将从拐点开始的所有词元标记为潜在幻觉，从而提供细粒度的词元级知识一致性评分。Qiu 等人[153]提出一种解码方法，结合假设验证，对生成文本的幻觉问题进行了检测。在生成文本的每一步解码时，将当前生成的序列（称为“向后假设”）和可能生成的未来序列（称为“向前假设”）视为假设。使用假设验证模型评估这些假设与输入事实的匹配程度，并根据生成的置信分数对解码候选进行排名。

此外，Himmi 等人[154]针对大多数幻觉检测方法依赖单一类型的检测器的问题，提出了一种无监督的多检测器聚合框架。该框架将多种外部和内部检测器的得分结合起来，充分利用不同检测器的互补优势，以捕捉不同类型幻觉的特征。

综上所述，非零资源幻觉检测方法通过引入外部知识源或辅助工具，能够有效弥补模型内部知识的局限性，从而显著提升检测的可靠性和准确性。然而，此类方法的有效性和适用性在一定程度上受到外部资源质量和计算条件的限制。例如，当知识库存在滞后性、不完整或权威性不足时，可能导致误判；由于外部资源的引入会增加系统的复杂性，检测过程的计算成本和时间开销显著提高。

为了便于读者了解上述幻觉检测方法的检测对象、检测原理及特点，对所综述文献进行分类总结，如表 3 和表 4 所示。

从表 3 中可以看出，不同类型的幻觉需要不同的检测方法。总体而言，白盒模型的检测方法由于能够直接访问模型内部状态，在识别 4 种类型幻觉时表现出较强的适应性和检测能力。然而，现有黑盒模型检测方法由于仅基于输入输出行为推断，受限

Table 3 Classification of Hallucination Detection Methods

表 3 幻觉检测方法分类

适用模型	分类	检测思路	上下文一致性幻觉	语义忠诚性幻觉	事实一致性幻觉	外部依赖性幻觉
白盒模型		基于特征	文献 [92，95，100]	文献 [95，99，103]	文献 [89−90，93−98，101−102，104]	文献 [105]
黑盒模型	零资源	传统方法迁移		文献 [107−108，110]	文献 [109]	
		生成多次响应	文献 [112]	文献 [111−112]	文献 [36，112−113]	
		反向验证			文献 [113−120]	
		基于大语言模型	文献 [125]		文献 [123，125−130]	
	非零资源	基于外部数据库			文献 [132−142]	
		基于分类器		文献 [145，150]	文献 [143−144，146−154]	文献 [72]

Table 4 Summary of Hallucination Detection Methods

表 4 幻觉检测方法总结

适用模型	分类	检测思路	检测原理	优点	不足
白盒模型	单特征	隐藏层激活值	利用模型内部的状态信息，识别生成内容中的幻觉	捕捉深层语义变化，细粒度检测能力强	计算开销较大、模型迁移性与通用性较弱
		logits 值		实现简单、计算效率高	难以捕获深层次语义不一致等问题
		熵值		熵值可结合条件熵等衍生指标，实现细粒度评估与动态调整	过于依赖概率分布的准确性
		注意力权重		能显式衡量生成内容对输入上下文的依赖程度，可解释性较强	注意力机制自身偏差容易引起误判
		梯度		细粒度反映输入特征对输出的影响，检测灵敏	计算开销较大、模型迁移性与通用性较弱
	多特征		结合模型内部状态的多种特征进行幻觉检测	可利用信息丰富、检测精度较高	不同特征之间可能存在信息冗余
黑盒模型	零资源	传统方法迁移	将传统幻觉检测方法的思路迁移至大语言模型上	技术简单、减少了重新设计检测机制的成本	跨领域和跨任务的泛化能力较弱
		生成多次响应	通过多次生成响应，分析样本一致性以评估生成内容的可靠性	无需额外的标注数据、简单易实施、计算成本相对较低	随机采样本身带来的噪声可能影响分析结果，导致检测误差
		反向验证	通过生成的答案重新构造查询，评估生成的查询与原始查询的一致性	无需外部知识库的支持，具有较强的可靠性	反向验证过程中涉及高复杂度的语法或逻辑设计，同时容易产生误差
		基于大模型	利用大语言模型的内部知识对生成的内容进行检测	避免依赖外部数据库或知识库，自动化程度高、易于实现、拓展性强	高度依赖于语言模型本身的推理能力和生成质量
	非零资源	基于外部数据库	将生成内容与外部数据库进行比对验证，以识别生成内容中可能存在的幻觉	具有较高的准确性和可信度	受外部资源质量和计算条件的限制
		基于分类器	构造合适的幻觉数据集对分类器进行训练，利用训练好的分类器进行幻觉检测	具有较高的检测精度和灵活性，自动化检测程度高	数据集构造成本高，且分类器的泛化能力可能受到数据集覆盖范围的限制

于检测原理，往往在识别某些特定类型幻觉时存在一定局限性。

具体而言，目前从反向验证、基于大语言模型和基于外部数据库的思路进行语义忠诚性幻觉检测的研究较为匮乏。反向验证方法的核心是生成—重构—验证的闭环流程，主要关注生成内容的事实准确性。一旦重构内容与输入事实一致，即判定为无幻觉。然而这一机制本质上忽略了用户提示语义与生成内容之间可能存在的偏离，因此在识别语义忠诚性幻觉时能力有限。而基于大语言模型的检测方法在检测语义忠诚性幻觉时，涉及更复杂的语义理解和意图建模，这超出了当前大语言模型的直接能力范围。为了有效检测此类幻觉，需借助语义理解能力更强的模型或设计更有效的语义匹配方法。此外，外部数据库虽然提供了明确的事实知识，但由于知识库通常以事实型三元组或简化文本形式存在，缺乏对复杂语义关系、上下文推理过程的建模能力。因此，基于外部数据库的方法在语义忠诚性幻觉检测中的适用性相对较弱。尽管目前针对外部依赖性幻觉的研究仍较少，但是由于 RAG 技术能够有效支撑大语言模型的垂直应用，针对外部依赖性幻觉的检测方法不容忽视。

结合表 4 可知，根据具体应用场景选择合适的幻觉检测方法至关重要。例如，对于开放式问答或长文本生成任务，需优先考虑能够处理语义忠诚性与上下文一致性幻觉的方法；而在医疗、法律、金融知

识密集型等领域应用中，则需更关注事实一致性与外部依赖性幻觉的检测。

4 幻觉检测基准

随着纵向大语言模型在多领域的广泛部署以及对可靠性需求的不断提升，开发科学而全面的幻觉检测基准已成为一个重要的研究方向。通过定义标准化的数据集、指标和检测流程能够帮助研究人员有效分析并定位大语言模型所输出的不可靠内容。高质量的基准设计需充分体现任务的特定需求，兼顾不同类型的幻觉，以确保检测的准确性与全面性。因此，本节将深入探讨现有的幻觉检测基准。

值得注意的是，文献 [15] 将幻觉基准分为幻觉评估基准和幻觉检测基准。前者侧重于评估大语言模型产生幻觉的程度，而后者主要用于评估现有幻觉检测方法的性能。经调研发现，两者具有类似的数据来源及构建流程，仅在评估对象和评价指标上略有不同。因此，为了便于读者理解与使用，将幻觉评估基准和幻觉检测基准一同综述，并总结各类基准的特点及其适用场景。

基准构建通常涉及数据收集、幻觉生成、幻觉注释等步骤。大多数基准直接使用现有数据集，例如基准 HELM[155] 以高质量的 Wikipedia 文档为主要来源，从中随机抽样 50 000 篇文章，为语言模型的生成任务提供真实且可靠的语料支持。在幻觉生成阶段，通常通过设计提示模板，并利用大语言模型生成包含幻觉的文本。然而，由于无法直接确定模型生成的内容是否包含幻觉，后续往往需要依赖标注环节进行验证。在这一过程中，标注工作通常由人工完成，从而确保结果的准确性。因此，在基准构建阶段，大多数流程采用“人工+自动化”的半自动化形式。除此之外，仅有少量研究采取纯人工方式构建数据集，例如 TruthfulQA[156] 等。

检测细粒度可以分为 6 个层次，包括词元级（token-level）、知识三元组级（knowledge-triplet-level）、句子级（sentence-level）、段落级（dialogue-level）、对话级（passage-level）、语义级（semantic-level）。每个层次对应不同的分析深度和应用场景。目前，大多数研究集中于句子级和段落级的幻觉检测。然而，当句子在语法和逻辑上表面正确，但某些主语或宾语等具体事实错误时，这些基准可能难以精准识别和定位幻觉。因此，部分研究尝试构建更细粒度的幻觉检测基准。例如，基准 HADES[157] 通过分析生成文本中的每个词元是否准确反映了输入数据或知识库中的信息，从而能够定位具体的错误词汇。同样，基准 UHGEval[158] 和 HalOmi[159] 不仅支持句子级的幻觉检测，还能够扩展到词元级，以提高检测的精确度。考虑到生成内容往往包含复杂的、多层次的事实陈述，句子或词元级的检测粒度可能不够明确，且容易导致交叉事实重叠问题。为此，基准 RefChecker[160] 通过使用知识三元组（如主语、谓语、宾语）的结构化表示，提供清晰的边界和语义独立性，从而更有效地识别幻觉内容。此外，传统的句子级和段落级基准在多轮交互场景中难以捕捉上下文依赖性以及对话逻辑的全局一致性。为解决这一问题，基准 DiaHalu[161] 和 HalluDial[162] 考虑对话的全局语境和多轮交互逻辑，从对话级视角评估幻觉检测能力。与此同时，基准 HaluBench[163] 不仅关注文本表面或句法层面的错误，还着重于语义内容的一致性。该基准通过语义扰动构造极难检测的幻觉文本，以测试幻觉检测能力。

在检测语言方面，大部分基准都是基于英文数据集构建。在众多基准中，仅有 UHGEval[158] 和 HalluQA[164] 基准专注于中文数据。这一现象的产生源于模型训练的语言偏好、学术界语言共通性等原因。这导致幻觉检测基准在中文等其他语言上的适用性存在一定局限性。一方面，由于语言结构、表述习惯、实体边界划分等方面存在显著差异，直接迁移英文基准到中文环境可能引发评估偏差[165]。另一方面，幻觉在不同语言中呈现的形式可能不同[166]，例如中文中出现的歧义实体、简繁转换错误等现象在英文中并不常见。此外，训练语料的语言分布不均也导致模型在巴斯克语等低资源语言上的表现更容易出现幻觉[167]。尽管如 UHGEval, HalluQA 等基准针对中文进行了探索，但整体上，当前幻觉检测基准在跨语言迁移与泛化能力方面仍存在明显不足。

在评价指标方面，大多数研究采用了通用的分类指标，如 $F1$ 分数、准确率（accuracy, Acc）、精确度（precision）和召回率（recall）。其他研究提出了新的计算方法，例如 FActScore[133] 等。然而，这些指标仍然存在对生成任务复杂性和多维度错误的覆盖等不足，无法动态适应生成任务的复杂性和实时性等需求。具体而言，大语言模型生成的内容可能涉及多轮对话中的延续生成，或基于有限输入做出合理推断。这使得仅依赖传统的准确率、精确度和召回率等指标，难以全面衡量生成结果的质量。此外，生成任务的错误通常是多维度的。这些指标主要关注语义一致性或事实一致性等单一维度问题，难以有效

捕捉生成过程中多样且复杂的幻觉现象。因此,研究人员越来越倾向于开发覆盖多维度评估标准的基准,以捕捉生成内容中潜在的幻觉问题。

从幻觉类型来看,现阶段绝大多数基准集中于检测事实一致性幻觉。除 RAGTruth[71] 外,其他基准均覆盖了事实一致性幻觉类别。这反映了当前研究对大语言模型生成内容真实性的高度关注。相比之下,针对语义忠诚性幻觉、上下文一致性幻觉及外部依赖性幻觉的检测基准仍处于探索阶段,仅有如 HaluEval[168], AutoHall[169] 等基准在部分任务中涉及这些类型的幻觉。但以上基准的整体覆盖程度较低,且检测粒度多局限于句子级或概念级,难以精确捕捉语义脱离、上下文冲突或检索知识偏移等幻觉现象。

在应用任务方面,当前幻觉检测基准覆盖了多种生成任务类型,主要包括知识问答、摘要生成、多轮对话及机器翻译等主流 NLP 任务。与此同时,HaLoGen[170] 等基准引入了代码生成、科学引用和长文本生成等复杂生成任务,拓展了幻觉检测的应用场景。然而,从整体趋势来看,目前大多数幻觉检测基准仍以通用领域任务为主,缺乏针对医学问答、法律文档生成等专业领域的系统化幻觉检测基准。这一局限主要受专业领域数据隐私保护与高标注成本等因素的限制。

基于以上分析,根据基准类别、检测细粒度和评价指标等维度对幻觉检测基准进行分类,如表 5 所示。

通过统计谷歌学术中 23 种基准所对应文献的引用次数发现,目前最常用的 3 种是 TruthfulQA, FactScore, HaluEval, 如图 7 所示。结合表 5 可知,这 3 种基准的广泛使用反映了当前研究对生成内容事实一致性和语义忠诚性的高度关注。具体而言,TruthfulQA 主要聚焦于语言模型回答问题时的真实性和一致性,特别是模型在对真实问题回答中的准确性; FactScore 通过量化生成内容的事实性来评估模型,广泛应用于长文本的事实一致性幻觉检测; HaluEval 则更加注重模型在多种生成任务中的幻觉现象检测、涵盖语义偏差和事实偏离等复杂情境。而针对上下文一致性幻觉或外部依赖性幻觉的基准如 HADES, RAGTruth 等,同样为特定的应用场景提供了更加精准的检测工具。

由此可见,不同基准的选择仍需根据应用场景的特点及幻觉检测的目标类型灵活调整,以确保检测的有效性和针对性。

5 未来研究方向与挑战

大语言模型近年来获得了业界内的广泛关注,取得了许多突破性进展,但关于针对大语言模型的幻觉检测的研究仍处于起步阶段,依然面临许多的挑战。基于本文对幻觉检测研究现状的深入分析,未来该领域的研究需要重点关注 4 个方向:

1)界定大语言模型幻觉的边界,建立更为明确的分类标准和度量体系。幻觉是大语言模型错误的一种特殊类型,但并非所有的错误都是幻觉。不同于简单的语法或逻辑错误,幻觉往往涉及更复杂的知识冲突或信息丢失,尤其在提示词的影响下表现得更加隐蔽且多样化。明确幻觉的边界,找出大语言模型“虚假误导”与“其他错误”的过渡点,是研究大语言模型幻觉问题的基础。此外,现有研究缺乏对幻觉的精细化分类,对幻觉特性的多维度解析尚不完善。因此,厘清幻觉与其他错误的边界,并构建科学的分类标准和度量体系,是未来幻觉检测问题在理论层面值得研究的方向之一。

2)探索复合性幻觉的解耦机制,揭示不同幻觉类型之间的关联和生成模式,提出针对复合幻觉的检测方法。现有的幻觉检测方法通常集中于特定类型的幻觉。然而,在实际应用中,大语言模型生成的幻觉并非单一类型,而是不同幻觉类型之间相互交织、叠加,表现出更加复杂的错误形式。这导致传统的单一类型检测方法难以有效捕捉幻觉。例如,基于知识验证的检测方法可能遗漏逻辑矛盾,而基于逻辑分析的方法又可能忽略事实性错误。通过解耦多维诱因分析幻觉类型之间的关联和生成机制,探索检测复合性幻觉的方法,以实现对复杂幻觉现象的全面覆盖和精准识别,是未来方法层面的研究方向之一。

3)研究跨模态、跨语言、跨领域场景下的幻觉检测方法。随着大语言模型在多语言、多模态、多领域环境中的广泛应用,幻觉问题呈现出更为复杂的特征。一方面,在跨语言场景中,不同语言之间的语义表达差异、知识覆盖不均与资源分布不平衡,增加了幻觉检测的难度。另一方面,在跨模态场景中,如图文生成、视觉问答、音频生成等任务中,文本幻觉往往与其他模态的实际内容冲突,传统仅基于文本推理的方法难以覆盖此类复合型幻觉。因此,亟需构建能够统一处理多语言、多模态输入的检测框

Table 5 Benchmark for Hallucination Detection in Large Language Models

表 5 大语言模型幻觉检测基准

基准	类型	规模	细粒度	语言	幻觉种类				任务类型、(不限于)	评价指标
					SFH	FCH	CCH	EDH		
TruthfulQA[156]	幻觉评估	817	句子级别	英文		√			知识问答、多项选择	Acc
HaluEval[168]	幻觉评估	35 000	句子级别	英文	√	√	√		知识问答、多轮对话、摘要生成	Acc
UHGEval[158]	幻觉评估	5 141	词元级别、句子级别	中文		√			文本生成	Acc, BLEU-4, ROUGE-L 等
HalluQA[164]	幻觉评估	450	句子级别	中文		√			知识问答、多轮对话	Non-Hallucination Rate 等
AutoHall[169]	幻觉检测	2 844	句子级别	英文		√	√		知识问答、摘要生成、多轮对话	Acc, *F*1
DiaHalu[161]	幻觉评估、幻觉检测	1 103	对话级别	英文	√	√			多轮对话	Precission, Recall, *F*1
FactCHD[170]	幻觉评估	58 343	句子级别	英文		√			知识问答、多轮对话	Micro *F*1, Expmatch
FaithBench[171]	幻觉评估	660	句子级别	英文	√	√			摘要生成	Micro *F*1, Balanced Accuracy
HALoGen[172]	幻觉评估	150 000	句子级别	英文	√	√			代码生成、摘要生成、科学引用	Hallucination Score 等
HELM[155]	幻觉检测	3 342	句子级别、段落级别	英文	√	√			知识问答、摘要生成、文本生成	AUC, PCCs
HalOmi[159]	幻觉检测	<3 546	词元级别、句子级别	多语言	√	√			机器翻译	AUC, AOC
HalluDial[162]	幻觉评估	146 856	对话级别	英文	√	√			知识问答	Acc, Macro *F*1
HaluBench[163]	幻觉评估	15 000	句子级别、语义级别	英文		√	√	√	知识问答、摘要生成	Acc, Macro *F*1 等
HaluEval 2.0[173]	幻觉评估	8 770	句子级别、段落级别	英文		√	√		知识问答、摘要生成	Micro hallucination rate 等
RefChecker[160]	幻觉检测	11 000	知识三元组级别	英文		√		√	知识问答、摘要生成	Acc, Macro *F*1
HADES[157]	幻觉检测	12 719	词元级别	英文	√	√	√		知识问答、多轮对话	Precision, Recall, *F*1 等
PHD[114]	幻觉检测	300	段落级别	英文		√			知识问答	Acc, Precision, *F*1, Recall
FELM[174]	幻觉检测	3 948	句子级别、段落级别	英文		√			多任务	Precision, *F*1, Recall 等
REALTIMEQA[175]	幻觉评估		句子级别	英文		√			知识问答	Acc, Exact Match, *F*1
FACTOR[176]	幻觉评估	4 266	句子级别	英文		√			知识问答、摘要生成	Acc
BAMBOO[177]	幻觉评估	3 004	句子级别	英文		√	√		长文本生成	Concordance Index, Acc, *F*1
Poly-FEVER[25]	幻觉检测	77 973	句子级别	多语言		√			事实验证	Acc
RAGTruth[71]	幻觉评估	18 000	句子级别、段落级别	英文				√	知识问答、摘要生成	Precision, Recall, *F*1

注：SFH 表示语义忠诚性幻觉，FCH 表示事实一致性幻觉，CCH 表示上下文一致性幻觉，EDH 表示外部依赖性幻觉。

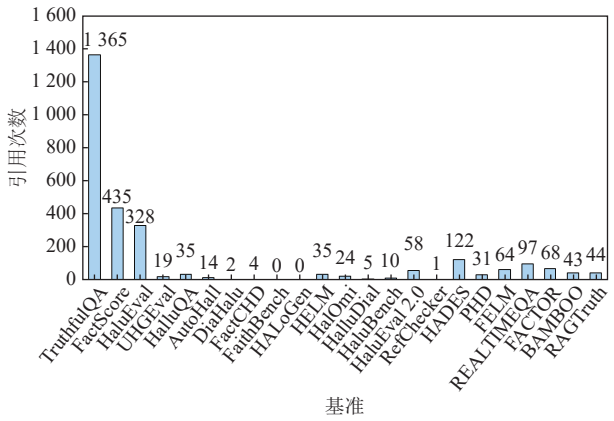


Fig. 7 Citation count of literature corresponding to the hallucination detection benchmarks

图 7 幻觉检测基准对应文献的引用次数

架，兼顾不同领域的知识特点与推理需求，提升模型在跨领域应用中的鲁棒性与泛化性。由此可见，打破当前幻觉检测方法对单一语言、单一模态、单一领域的依赖，建立面向多语言、多模态、多领域统一的幻觉检测体系，是未来幻觉检测问题在方法层面值得研究的方向之一。

4)研究幻觉检测与缓解的协同机制，实现端到端的生成优化。大语言模型的幻觉检测与缓解通常被设计为 2 个独立的处理环节，这将导致幻觉缓解过程难以充分利用检测阶段的细粒度信息，从而限制了缓解策略在提升生成内容质量方面的实际效能。同时，这种处理方式显著增加了整体流程的时间与计算开销，尤其在对实时性要求较高的应用场景中表现出明显的局限性。由此可见，构建幻觉检测与纠正的协同机制以及在生成过程中动态评估并调整输出内容，是未来幻觉检测问题在应用层面值得研究的方向之一。

作者贡献声明: 李自拓负责文献调研、内容设计、论文撰写和最后版本的修订; 孙建彬负责提出指导意见、框架设计和全文修订; 陈广州、方馨悦和崔瑞靖负责论文修改; 田植良和黄震负责论文审核; 杨克巍提出指导意见以及修订论文。其中, 孙建彬和田植良为本文的共同通信作者。

参 考 文 献

[1] Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: A survey[J]. ACM Computing Surveys, 2023, 56(2): 1−40

[2] Fan Lizhou, Li Lingyao, Ma Zihui, et al. A bibliometric review of large language models research from 2017 to 2023[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(5): 1−25

[3] Navigli R, Conia S, Ross B. Biases in large language models: Origins, inventory, and discussion[J]. ACM Journal of Data and Information Quality, 2023, 15(2): 1−21

[4] Zhao Xin Wayne, Zhou Kun, Li Junyi, et al. A survey of large language models[J]. arXiv preprint, arXiv: 2303.18223, 2023

[5] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[6] Ye Wentao, Hu Jiaqi, Wang Haobo, et al. A trusted evaluation system for safe deployment of large language models[J]. Journal of Computer Research and Development, 2025, 62(7): 1668−1684 (in Chinese)
(叶文涛, 胡家齐, 王皓波, 等. 面向大语言模型安全部署的可信评估体系[J]. 计算机研究与发展, 2025, 62(7): 1668−1684)

[7] Kalai A T, Santosh S V. Calibrated language models must hallucinate[C]//Proc of the 56th Annual ACM Symp on Theory of Computing. New York: ACM, 2024: 160−171

[8] Lin, Zichao, Guan Shuyan, Zhang Wending, et al. Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models[J]. Artificial Intelligence Review, 2024, 57(9): 1−50

[9] Hu Songlin, Li Juanzi, Qin Bing, et al. The good and evil big model: A special topic on big models and security[J]. Journal of Computer Research and Development, 2024, 61(5): 1085−1093 (in Chinese)
(虎嵩林, 李涓子, 秦兵, 等. 亦正亦邪大模型——大模型与安全专题导读[J]. 计算机研究与发展, 2024, 61(5): 1085−1093)

[10] Venkit P N, Chakravorti T, Gupta V, et al. An audit on the perspectives and challenges of hallucinations in NLP[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 6528−6548

[11] Li Xu, Zhu Rui, Chen Xiaolei, et al. A survey of hallucinations in large vision-language models: Causes, evaluations and mitigations[J]. Journal of Computer Research and Development, 2025, 62(12):2929−2950 (in Chinese)
(李煦, 朱睿, 陈小磊, 等. 视觉语言大模型的幻觉综述: 成因、评估与治理[J]. 计算机研究与发展, 2025, 62(12):2929−2950)

[12] Dahl M, Varun M, Mirac S, et al. Large legal fictions: Profiling legal hallucinations in large language models[J]. Journal of Legal Analysis, 2024, 16(1): 64−93

[13] Liu Zhuang, Huang Ddegen, Huang Kaiyu, et al. FinBERT: A pre-trained financial language representation model for financial text mining[C]//Proc of the 29th Int Joint Conf on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 2021: 4513−4519

[14] Liu Zeyuan, Wang Pengjiang, Song Xiaobin, et al. Survey on hallucinations in large language models[J]. Journal of Software, 2025, 36(3): 1152−1185 (in Chinese)
(刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025, 36(3): 1152−1185)

[15] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1−38

[16] Zhang Yue, Li Yafu, Cui Leyang, et al. Siren's song in the AI ocean: A survey on hallucination in large language models[J]. arXiv preprint, arXiv: 2309.01219, 2023

[17] Huang Lei, Yu Weijiang, Ma Weitao, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1−55

[18] Abbasi Yadkori Y, Kuzborskij I, György A, et al. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2024: 58077−58117

[19] Wu Junchao, Shu Yang, Zhan Runzhe, et al. A survey on LLM-generated text detection: Necessity, methods, and future directions[J]. Computational Linguistics, 2025, 51(1): 275−338

[20] Black S, Biderman S, Hallahan E, et al. GPT-NeoX−20B: An open-source autoregressive language model[J]. arXiv preprint, arXiv: 2204.06745, 2022

[21] Sun Weiwei, Shi Zhengliang, Gao Shen, et al. Contrastive learning reduces hallucination in conversations[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 13618−13626

[22] Chen Sihao, Zhang Fan, Sone K, et al. Improving faithfulness in abstractive summarization with contrast candidate generation and selection[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2021: 5935−5941

[23] Dziri N, Madotto A, Zaiane O, et al. Neural path hunter: Reducing hallucination in dialogue systems via path grounding[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 2197−2214

[24] Huang Yichong, Feng Xiachong, Feng Xiaocheng. The factual inconsistency problem in abstractive text summarization: A survey[J]. arXiv preprint, arXiv: 2104.14839, 2023

[25] Chen Xinxi, Wang Li, Wu Wei, et al. Honest AI: Fine-tuning “small” language models to say “I Don’t Know”, and reducing hallucination in RAG[J]. arXiv preprint, arXiv: 2410.09699, 2024

[26] Huang Yizheng, Huang J. A survey on retrieval-augmented text generation for large language models[J]. arXiv preprint, arXiv: 2404.10981, 2024

[27] Xie Jinheng, Mao Weijia, Bai Zechen, et al. Show-o: One single transformer to unify multimodal understanding and generation[J]. arXiv preprint, arXiv: 2408.12528, 2024

[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 30th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2017: 5998−6008

[29] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[30] Radford, A, Narasimhan T, Salimans I, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-06-11)[2024-12-21]. https://openai.com/index/language-unsupervised/

[31] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL]. [2024-12-21]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[32] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proc of the 33rd Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 1877−1901

[33] OpenAI. GPT-4 technical report[J]. arXiv preprint, arXiv: 2305.10403, 2023

[34] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. Machine Learning Research, 2020, 21(140): 1−67.

[35] Yang Zhilin, Dai Zihang, Yang Yiming, et al. XLNet: Generalized autoregressive pretraining for language understanding[J]. arXiv preprint, arXiv: 1906.08237, 2019

[36] Elaraby M, Lu Mengyin, Dunn J, et al. HaLo: Estimation and reduction of hallucinations in open-source weak large language models[J]. arXiv preprint, arXiv: 21308.11764, 2023

[37] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[38] Yin Ziqi, Zhang Mingxin, Kawahara D. Harmony: A home agent for responsive management and action optimization with a locally deployed large language model[J]. arXiv preprint, arXiv: 2410.14252, 2024

[39] Li Zuchao, Zhang Shitou, Zhao Hai, et al. BatGPT: A bidirectional autoregessive talker from generative pre-trained transformer[J]. arXiv preprint, arXiv: 2307.00360, 2023

[40] DeRose J F, Wang Jiayao, Berger M. Attention flows: Analyzing and comparing attention mechanisms in language models[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(2): 1160−1170

[41] Saxena A, Bhattacharyya P. Hallucination detection in machine generated text: A survey[EB/OL]. (2025-01-21)[2025-01-22]. https://www.cfilt.iitb.ac.in/resources/surveys/2024/survey_ashita_hallucination_detection_in_machine_generated_text_2024.pdf

[42] Hahn M. Theoretical limitations of selfattention in neural sequence models[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 156−171

[43] Chiang D, Cholak P. Overcoming a theoretical limitation of self-attention[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 7654−7664

[44] Naveed H, Khan A U, Qiu Shi, et al. A comprehensive overview of large language models[J]. arXiv preprint arXiv: 2307.06435, 2023

[45] Annepaka Y, Pakray P. Large language models: A survey of their development, capabilities, and applications[J]. Knowledge and Information Systems, 2025, 67(3): 2967−3022

[46] Xu Weijia, Agrawal S, Briakou E, et al. Understanding and detecting hallucinations in neural machine translation via model introspection[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 546−564

[47] Filippova K. Controlled hallucinations: Learning to generate faithfully from noisy data[C]//Proc of the Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA: ACL, 2020: 864−870

[48] Nicholas C, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//Proc of the 30th USENIX Security Symp. Berkeley, CA: USENIX Association, 2021: 2633−2650

[49] Carlini N, Ippolito D, Jagielski M, et al. Quantifying memorization across neural language models[J]. arXiv preprint, arXiv: 2202.07646, 2022

[50] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1−13

[51] Lee K, Ippolito D, Nystrom A, et al. Deduplicating training data makes language models better[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 8424−8445

[52] Kandpal N, Deng Haikang, Roberts A, et al. Large language models struggle to learn long-tail knowledge[C]//Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 15696−15707

[53] Hernandez D, Brown T, Conerly T, et al. Scaling laws and interpretability of learning from repeated data[J]. arXiv preprint, arXiv: 2205.10487, 2022

[54] Chen Jinyin, Chen Yipeng, Chen Yiming, et al. Fairness research on deep learning[J]. Journal of Computer Research and Development, 2021, 58(2): 264−280 (in Chinese)
(陈晋音, 陈奕芃, 陈一鸣, 等. 面向深度学习的公平性研究综述[J]. 计算机研究与发展, 2021, 58(2): 264−280)

[55] Sheng E, Chang Kaiwei, Natarajan P, et al. Societal biases in language generation: Progress and challenges[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 4275−4293

[56] Wan Yixin, Pu G, Sun Jiao, et al. “Kelly is a warm person, Joseph is a role model”: Gender biases in LLM-generated reference letters[C]//Proc of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA: ACL, 2023: 3730−3748

[57] Karpowicz, M. On the fundamental impossibility of hallucination control in large language models[J]. arXiv preprint, arXiv: 2506.06382, 2025

[58] Liu Yinqiu, Liu Guangyuan, Zhang Ruichen, et al. Hallucination-aware optimization for large language model-empowered communications[J]. arXiv preprint, arXiv: 2412.06007, 2024

[59] Pal A, Umapathi L K, Sankarasubbu M. Med-HALT: Medical domain hallucination test for large language models[C]//Proc of the 27th Conf on Computational Natural Language Learning, Stroudsburg, PA: ACL, 2023: 314−334

[60] Roychowdhury S. Journey of hallucination-minimized generative AI solutions for financial decision makers[C]//Proc of the 17th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2024: 1180−1181

[61] Wang Mengru, Yao Yunzhi, Xu Ziwen, et al. Knowledge mechanisms in large language models: A survey and perspective[C]//Proc of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL, 2024: 7097−7135

[62] Zhang Ningyu, Yao Yunzhi, Tian Bozhong, et al. A comprehensive study of knowledge editing for large language models[J]. arXiv preprint, arXiv: 2401.01286, 2024

[63] Feng Zhangyin, Ma Weitao, Yu Weijiang, et al. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications[J]. arXiv preprint, arXiv: 2311.05876, 2023

[64] Schulman J. Reinforcement learning from human feedback: Progress and challenges[EB/OL]. [2025-01-01]. https://eecs.berkeley.edu/research/colloquium/230419-2/

[65] Pal A, Sankarasubbu M. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations[J]. arXiv preprint, arXiv: 2402.07023, 2024

[66] Wang Haochun, Zhao Sendong, Qiang Zewen, et al. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in Chinese[J]. ACM Transactions on Knowledge Discovery from Data, 2025, 19(2): 1−17

[67] Wei J, Huang Da, Lu Yifeng, et al. Simple synthetic data reduces sycophancy in large language models[J]. arXiv preprint, arXiv: 2308.03958, 2023

[68] Sharma Mrinank, Tong M, Korbak T, et al. Towards understanding sycophancy in language models[J]. arXiv preprint, arXiv: 2310.13548, 2023

[69] Perez E, Ringer S, Lukošiūtė K, et al. Discovering language model behaviors with model-written evaluations[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 13387−13434

[70] Lu Taiming, Shen Lingfeng, Yang Xinyu, et al. It takes two: On the seamlessness between reward and policy model in RLHF[J]. arXiv preprint, arXiv: 2406.07971, 2024

[71] Lee N, Wei Ping, Xu Peng, et al. Factuality enhanced language models for open-ended text generation[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 34586−34599

[72] Renze M, Guven E. The effect of sampling temperature on problem solving in large language models[J]. arXiv preprint, arXiv: 2402.05201, 2024

[73] Chang H S, Peng Nanyun, Bansal M, et al. Real sampling: Boosting factuality and diversity of open-ended generation via asymptotic entropy[J]. arXiv preprint, arXiv: 2406.07735, 2024

[74] Zhang Muru, Press O, Merrill W, et al. How language model hallucinations can snowball[J]. arXiv preprint, arXiv: 2305.13534, 2023

[75] Kang Haoqiang, Ni Juntong, Yao Huaxiu. Ever: Mitigating hallucination in large language models through real-time verification and rectification[J]. arXiv preprint, arXiv: 2311.09114, 2023

[76] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 24824−24837

[77] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 22199−22213

[78] Yee E, Li A, Tang Chenyu, et al. Faithful and unfaithful error recovery in chain of thought[EB/OL]. (2024-07-10)[2024-12-21]. https://openreview.net/forum?id=IPZ28ZqD4I

[79] Agarwal C, Tanneru S H, Lakkaraju H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models[J]. arXiv preprint, arXiv: 2402.04614, 2024

[80] Turpin M, Michael J, Perez E, et al. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2024: 74952−74965

[81] Lanham T, Chen A, Radhakrishnan A, et al. Measuring faithfulness in chain-of-thought reasoning[J]. arXiv preprint, arXiv: 2307.13702, 2023

[82] Chen Qiguang, Chen Libo, Liu Jinhao, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models[J]. arXiv preprint, arXiv: 2503.09567, 2025

[83] Ren Ruiyang, Wang Yuhao, Qu Yingqi, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation[J]. arXiv preprint, arXiv: 2307.11019, 2023

[84] Wen Bingbing, Xu Chenjun, Bin H A N, et al. Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration[C]//Proc of the 37th Advances in Neural Information Processing Systems Workshop. Cambridge, MA: MIT, 2024: 1877−1901

[85] Niu Cheng, Wu Yuanhao, Zhu Juno, et al. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models[J]. arXiv preprint, arXiv: 2401.00396, 2023

[86] Hu Haichuan, Sun Yuhan, Zhang Quanjun. LRP4RAG: Detecting

hallucinations in retrieval-augmented generation via layer-wise relevance propagation[J]. arXiv preprint, arXiv: 2408.15533, 2024

[87] Barnett S, Kurniawan S, Thudumu S, et al. Seven failure points when engineering a retrieval augmented generation system[C]//Proc of the 3rd IEEE/ACM Int Conf on AI Engineering-Software Engineering for AI. Piscataway, NJ: IEEE, 2024: 194−199

[88] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157−173

[89] Rateike M, Cintas C, Wamburu J, et al. Weakly supervised detection of hallucinations in LLM activations[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2023: 1877−1901

[90] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4593−4601

[91] Chuang Y S, Xie Yujia, Luo Hongyin, et al. DoLa: Decoding by contrasting layers improves factuality in large language models. [EB/OL]. (2024-01-16)[2024-12-21]. https://openreview.net/forum?id=Th6NyL07na

[92] Varshney N, Yao Wenlin, Zhang Hongming, et al. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation[J]. arXiv preprint, arXiv: 2307.03987, 2023

[93] Chen Kedi, Chen Qin, Zhou Jie, et al. Enhancing uncertainty modeling with semantic graph for hallucination detection[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 23586−23594

[94] Fu Jinlan, Ng S K, Jiang Zhengbao, et al. Gptscore: Evaluate as you desire[C]//Proc of the 2024 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg, PA: ACL 2024: 6556−6576

[95] Guerreiro N M, Voita E, Martins A F. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation[C]//Proc of the 17th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 1059−1075

[96] Yuan Weizhe, Neubig G, Liu Pengfei. Bartscore: Evaluating generated text as text generation[C]//Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2021: 27263−27277

[97] Xiao Yijun, Wang W Y. On hallucination and predictive uncertainty in conditional language generation[C]//Proc of the 16th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 2734−2744

[98] Su Weihang, Tang Yichen, Ai Qingyao, et al. Mitigating entity-level hallucination in large language models[C]//Proc of the 2024 Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval in the Asia Pacific Region. New York: ACM, 2024: 23−31

[99] Van Der Poel L, Cotterell R, Meister C. Mutual information alleviates hallucinations in abstractive summarization[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 5956−5965

[100] Chuang Y S, Qiu Linlu, Hsieh C Y, et al. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 1419−1436

[101] Sriramanan G, Bharti S, Sadasivan V S, et al. LLM-check: Investigating detection of hallucinations in large language models[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2024: 34188−34216

[102] Zablocki P, Gajewska Z. Assessing hallucination risks in large language models through internal state analysis[EB/OL]. (2024-07-17)[2025-07-11].https://www.authorea.com/doi/full/10.22541/au.172124175.55788724

[103] Hu Xiaomeng, Zhang Yiming, Peng Ru, et al. Embedding and gradient say wrong: A white-box method for hallucination detection[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024, 1950−1959

[104] Snyder B, Moisescu M, Zafar M B. On early detection of hallucinations in factual question answering[C]//Proc of the 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2024: 2721−2732

[105] Sun Zhongxiang, Zang Xiaoxue, Zheng Kai, et al. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability[J]. arXiv preprint, arXiv: 2410.11414, 2024

[106] Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models[J]. arXiv preprint, arXiv: 2312.11805, 2023

[107] Bhamidipati P, Malladi A, Shrivastava M, et al. Zero-shot multi-task hallucination detection[J]. arXiv preprint, arXiv: 2403.12244, 2024

[108] Rashad M, Zahran A, Amin A, et al. FactAlign: Fact-level hallucination detection and classification through knowledge graph alignment[C]//Proc of the 4th Workshop on Trustworthy Natural Language Processing. Stroudsburg, PA: ACL, 2024, 79−84

[109] Sansford H, Richardson N, Maretic H P, et al. GraphEval: A knowledge-graph based LLM hallucination evaluation framework [J]. arXiv preprint, arXiv: 2407.10793, 2024

[110] Durmus E, He H, Diab M. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 5055−5070

[111] Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy[J]. Nature, 2024, 630 (8017): 625−630

[112] Manakul P, Liusie A, Gales M J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 9004−9017

[113] Fang Xinyue, Huang Zhen, Tian Zhiliang, et al. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 23868−23877

[114] Yang Shiping, Sun Renliang, Wan Xiaojun. A new benchmark and

reverse validation method for passage-level hallucination detection[C]//Proc of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA: ACL, 2023: 3898−3908

[115] Honovich O, Choshen L, Aharoni R, et al. Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 7856−7870

[116] Scialom T, Dray P A, Gallinari P, et al. QuestEval: Summarization asks for fact-based evaluation[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 6594−6604

[117] Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 5008−5020

[118] Fabbri A R, Wu C S, Liu Wenhao, et al. QAFactEval: Improved QA-based factual consistency evaluation for summarization[C]//Proc of the 2022 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2022: 2587−2601

[119] Yehuda Y, Malkiel I, Barkan O, et al. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 9333−9347

[120] Cohen R, Hamri M, Geva M, et al. LM vs LM: Detecting factual errors via cross examination[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 12621−12640

[121] Chiang C H, Lee H. Can large language models be an alternative to human evaluations?[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 15607−15631

[122] Liu Yang, Iter D, Xu Yichong, et al. G-EVAL: NLG evaluation using GPT-4 with better human alignment[J]. arXiv preprint, arXiv: 2303.16634, 2023

[123] Adlakha V, BehnamGhader P, Lu Xinghan, et al. Evaluating correctness and faithfulness of instruction-following models for question answering[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 775−793

[124] Gao Mingqi, Ruan Jie, Sun Renliang. Human-like summarization evaluation with ChatGPT[J]. arXiv preprint, arXiv: 2304.02554, 2023

[125] Jain S, Keshava V, Sathyendra S M, et al. Multi-dimensional evaluation of text summarization with in-context learning[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 8487−8495

[126] Luo Zheheng, Xie Qianqian, Ananiadou S. ChatGPT as a factual inconsistency evaluator for text summarization[J]. arXiv preprint, arXiv: 2303.15621, 2023

[127] Dhuliawala S, Komeili M, Xu Jing, et al. Chain-of-verification reduces hallucination in large language models [C] //Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 3563−3578

[128] Luo Jianjun, Xiao Cong, Ma Feng. Zero-resource hallucination prevention for large language models[J]. arXiv preprint, arXiv: 2309.02654, 2023

[129] Agrawal A, Mirac S, Lester M, et al. Do language models know when they're hallucinating references[J]. arXiv preprint, arXiv: 2305.18248, 2023

[130] Das S, Srihari R K. Compos mentis at semeval2024 task6: A multi-faceted role-based large language model ensemble to detect hallucination[C]//Proc of the 18th Int Workshop on Semantic Evaluation (SemEval-2024). Stroudsburg, PA: ACL, 2024: 1449−1454

[131] Zheng D, Lapata M, Pan J Z. Large language models as reliable knowledge bases?[J]. arXiv preprint, arXiv: 2407.13578, 2024

[132] Son S S, Park J, Hwang J I, et al. HaRiM+: Evaluating summary quality with hallucination risk[J]. arXiv preprint, arXiv: 2211.12118, 2022

[133] Min S, Krishna K, Lyu X, et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 12076−12100

[134] Chern I, Chern S, Chen Shiqi, et al. FacTool: Factuality detection in generative AI ——A tool augmented framework for multi-task and multi-domain scenarios[J]. arXiv preprint, arXiv: 2307.13528, 2023

[135] Hu Xiangkun, Ru Dongyu, Qiu Lin, et al. Knowledge-centric hallucination detection[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 6953−6975

[136] Mishra A, Asai A, Balachandran V, et al. Fine-grained hallucination detection and editing for language models[J]. arXiv preprint, arXiv: 2401.06855, 2024

[137] Li Ningke, Li Yuekang, Liu Yi, et al. Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models[C]//Proc of the ACM on Programming Languages. New York: ACM, 2024: 1843−1872

[138] Bayat F F, Qian Kun, Han Benjamin, et al. Fleek: Factual error detection and correction with evidence retrieved from external knowledge[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 124−130

[139] Wang Binjie, Chern S, Chern E, et al. Halu-J: Critique-based hallucination judge[J]. arXiv preprint, arXiv: 2407.12943, 2024

[140] Zhao Xinping, Yu Jindi, Liu Zhenyu, et al. Medico: Towards hallucination detection and correction with multi-source evidence fusion[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 34−45

[141] Zhang Jiawei, Xu Chejian, Gai Yu, et al. KnowHalu: Hallucination detection via multi-form knowledge based factual checking[J]. arXiv preprint, arXiv: 2404.02935, 2024

[142] Wang Xiaohua, Yan Yuliang, Huang Longtao, et al. Hallucination detection for generative large language models by Bayesian sequential estimation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 15361−15371

[143] Zhou Chunting, Neubig G, Gu Jiatao, et al. Detecting hallucinated content in conditional neural sequence generation[J]. arXiv preprint, arXiv: 2011.02593, 2020

[144] Wojciech K, McCann B, Xiong Caiming, et al. Evaluating the factual consistency of abstractive text summarization[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 9332−9346

[145] Qiu Yifu, Ziser Y, Korhonen A, et al. Detecting and mitigating hallucinations in multilingual summarisation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 8914−8932

[146] Du Xuefeng, Xiao Chaowei, Li Yixuan. HaloScope: Harnessing unlabeled LLM generations for hallucination detection[J]. arXiv preprint, arXiv: 2409.17504, 2024

[147] Quevedo E, Yero J, Koerner R, et al. Detecting hallucinations in large language model generation: A token probability approach[J]. arXiv preprint, arXiv: 2405.19648, 2024

[148] Cao Meng, Dong Yue, Cheung J C K. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 3340−3354

[149] Santhanam S, Hedayatnia B, Gella S, et al. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation[J] arXiv preprint, arXiv: 2110.05456, 2021

[150] Zha Yuheng, Yang Yichi, Li Ruichen, et al. AlignScore: Evaluating factual consistency with a unified alignment function[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 11328−11348

[151] Shen Jiaming, Liu Jialu, Finnie D, et al. “Why is this misleading?”: Detecting news headline hallucinations with explanations[C]//Proc of the ACM Web Conf. New York: ACM, 2023: 1662−1672

[152] Choi S, Fang Tianqing, Wang Zhaowei, et al. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 14035−14053

[153] Qiu Yifu, Embar V, Shay B, et al. Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation[C]//Proc of the Findings of the Association for Computational Linguistics: NAACL 2024. Stroudsburg, PA: ACL, 2023: 1628−1644

[154] Himmi A, Staerman G, Picot M, et al. Enhanced hallucination detection in neural machine translation through simple detector aggregation[J]. arXiv preprint, arXiv: 2402.13331, 2024

[155] Su Weihang, Wang Changyue, Ai Qingyao, et al. Unsupervised real-time hallucination detection based on the internal states of large language models[J]. arXiv preprint, arXiv: 2403.06448, 2024

[156] Lin S, Hilton J, Evans O. TruthfulQA: Measuring how models mimic human falsehoods[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 3214−3252

[157] Liu Tianyu, Zhang Yizhe, Brockett C, et al. A token-level reference-free hallucination detection benchmark for free-form text generation[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 6723−6737

[158] Liang Xun, Song Shichao, Niu Simin, et al. UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation[J]. arXiv preprint, arXiv: 2311.15296, 2023

[159] Dale D, Voita E, Lam J, et al. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation[J]. arXiv preprint, arXiv: 2305.11746, 2023

[160] Hu Xiangkun, Ru Dongyu, Qiu Lin. RefChecker: Reference-based fine-grained hallucination checker and benchmark for large language models[J]. arXiv preprint, arXiv: 2405.14486, 2024

[161] Chen Kedi, Chen Qin, Zhou Jie, et al. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models[J]. arXiv preprint, arXiv: 2403.00896, 2024

[162] Luo Wen, Shen Tianshu, Li Wei, et al. HalluDial: A large-scale benchmark for automatic dialogue-level hallucination evaluation[J]. arXiv preprint, arXiv: 2406.07070, 2024

[163] Ravi S S, Mielczarek B, Kannappan A, et al. Lynx: An open source hallucination evaluation model[J]. arXiv preprint arXiv: 2407.08488, 2024

[164] Cheng, Qinyuan, Sun Tianxiang, Zhang Wenwei, et al. Evaluating hallucinations in Chinese large language models[J]. arXiv preprint, arXiv: 2310.03368, 2023

[165] Dale D, Costa-jussà M. Blaser 2.0: A metric for evaluation and quality estimation of massively multilingual speech and text translation[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 16075−16085

[166] Zhang Hanzhi, Anjum S, Fan Heng, et al. Poly-FEVER: A multilingual fact verification benchmark for hallucination detection in large language models[J]. arXiv preprint, arXiv: 2503.16541, 2025

[167] Etxaniz J, Sainz O, Miguel N, et al. Latxa: An open language model and evaluation suite for Basque[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 14952−14972

[168] Li Junyi, Cheng Xiaoxue, Zhao Xin, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models[C]// Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 6449−6464

[169] Cao Zouying, Yang Yifei, Zhao Hai. AutoHall: Automated hallucination dataset generation for large language models[J]. arXiv preprint, arXiv: 2310.00259, 2023

[170] Ravichander A, Ghela S, Wadden D, et al. The HALoGen benchmark: Fantastic LLM hallucinations and where to find Them[EB/OL]. (2024-10-15)[2024-12-30]. https://openreview.net/pdf?id=pQ9QDzckB7

[171] Chen Xiang, Song Duanzheng, Gui Honghao, et al. Factchd: Benchmarking fact-conflicting hallucination detection[C]//Proc of the 33rd Int Joint Conf on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 2024: 6216−6224

[172] Bao F S, Li M, Qu Renyi, et al. FaithBench: A diverse hallucination benchmark for summarization by modern LLMs[J]. arXiv preprint, arXiv: 2410.13210, 2024

[173] Li Junyi, Chen Jie, Ren Ruiyang, et al. The dawn after the dark: An empirical study on factuality hallucination in large language

models[J]. arXiv preprint, arXiv: 2401.03205, 2024

[174] Chen Shiqi, Zhao Yiran, Zhang Jinghan, et al. Felm: Benchmarking factuality evaluation of large language models[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2024: 44502−44523

[175] Kasai J, Sakaguchi K, Takahashi Y, et al. REALTIME QA: What's the answer right now?[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2024: 49025−49043

[176] Muhlgay D, Ram O, Magar I, et al. Generating benchmarks for factuality evaluation of language models[C]//Proc of the 18th Conf of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL 2024: 49−66

[177] Dong Zican, Tang Tianyi, Li Junyi, et al. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models[J]. arXiv preprint, arXiv: 2309.13345, 2023

Li Zituo, born in 1999. PhD candidate. His main research interests include hallucination, security, and evaluation in LLMs.

李自拓, 1999 年生。博士研究生。主要研究方向为大语言模型幻觉、安全和评估。

Sun Jianbin, born in 1989. PhD, associate professor, master supervisor. His main research interests include system test and evaluation, and decision and analysis under uncertainty.

孙建彬, 1989 年生。博士, 副教授, 硕士生导师。主要研究方向为系统试验与评价、不确定性决策分析。

Chen Guangzhou, born in 2003. PhD candidate. His main research interests include LLMs inference optimization, LLMs evaluation, and data stream mining. (chen_gz@nudt.edu.cn)

陈广州, 2003 年生。博士研究生。主要研究方向为大语言模型推理优化、大语言模型评估、数据流挖掘。

Fang Xinyue, born in 2002. Master candidate. Her main research interests include natural language processing, large language models, and deep learning. (fangxinyue@nudt.edu.cn)

方馨悦, 2002 年生。硕士研究生。主要研究方向为自然语言处理、大语言模型、深度学习。

Cui Ruijing, born in 1996. PhD candidate. His main research interests include causal inference, and equipment test and evaluation. (cuiruijing@nudt.edu.cn)

崔瑞靖, 1996 年生。博士研究生。主要研究方向为因果推断、装备试验与鉴定。

Tian Zhiliang, born in 1992. Associate professor. Member of CCF. His main research interests include LLMs and natural language processing. (tianzhiliang@nudt.edu.cn)

田植良, 1992 年生。副研究员。CCF 会员。主要研究方向为大语言模型、自然语言处理。

Huang Zhen, born in 1984. PhD, professor, PhD supervisor. Member of CCF. His main research interests include machine learning, natural language processing, and intelligent systems. (huangzhen@nudt.edu.cn)

黄　震, 1984 年生。博士, 教授, 博士生导师。CCF 会员。主要研究方向为机器学习、自然语言处理、智能系统。

Yang Kewei, born in 1977. PhD, professor, PhD supervisor. His main research interests include equipment test and evaluation, and defense acquisition and systems requirement modeling. (kayyang27@nudt.edu.cn)

杨克巍, 1977 年生。博士, 教授, 博士生导师。主要研究方向为装备试验和评估、国防采办与体系需求建模。