大语言模型事实性生成的鲁棒不确定性量化

Yuhao Zhang、Zhongliang Yang*、Linna Zhou网络空间安全学院，北京邮电大学，中国北京{yuhaozhang, yangzl, zhoulinna}@bupt.edu.cn

arXiv:2601.00348v1 [cs.CL] 1 Jan 2026

摘要—大语言模型（LLM）技术的迅猛发展，推动了其在专业领域和日常生活中的广泛应用。然而，大语言模型幻觉这一持续存在的挑战已成为关键限制因素，严重损害了人工智能生成内容的可靠性与可信度。这一挑战引起了科学界的广泛关注，并促使研究人员在幻觉检测和缓解策略方面开展了大量研究。当前的方法框架揭示了一项关键局限：传统的不确定性量化方法主要适用于常规问答范式，但在面对非典型或对抗性提问策略时则表现出明显不足。这种性能差距引发了人们对大语言模型在需要具备强大批判性思维能力的实际应用中，其回答是否可靠的重大担忧。本研究旨在弥补这一差距，提出一种适用于多事实生成任务的不确定性量化场景。我们精心构建了一组包含虚假姓名的陷阱问题。在此场景基础上，我们创新性地提出了一种新颖的鲁棒不确定性量化方法（RU）。我们开展了一系列实验来验证该方法的有效性。结果表明，所构建的陷阱问题集表现出色。此外，在四种不同模型上与基线方法进行比较时，我们提出的不确定性量化方法展现出优异性能，其ROC AUC值平均比表现最佳的基线方法高出0.1–0.2，为解决大语言模型幻觉问题提供了新的思路和方法。

索引词—大语言模型，虚假人物传记生成，鲁棒不确定性量化。

I. 引言

大语言模型（LLM）在自然语言生成（NLG）领域的广泛应用，使人们在日常生活中越来越依赖这些模型。人们日益借助大语言模型来辅助阅读和理解文档 [1]，支持决策 [2]，并利用模型的响应和生成内容完成各种任务。这种依赖的加深反过来也提升了模型输出的可信度和可靠性的重要性。然而，大语言模型不可避免地容易出现“幻觉”问题 [3]。在这一现象中，模型可能会生成晦涩或捏造的内容，从而对输出的可信度和可靠性构成重大挑战。

大语言模型的幻觉可分为事实性幻觉和忠实性幻觉 [4]。忠实性幻觉主要评估输出是否忠实于输入，而事实性幻觉主要评估生成内容是否符合现实。只需评估输出与输入之间的相关性，就能识别忠实性幻觉。事实性幻觉具有细粒度特性且分布零散，因此较难凭直觉发现 [5]。模型

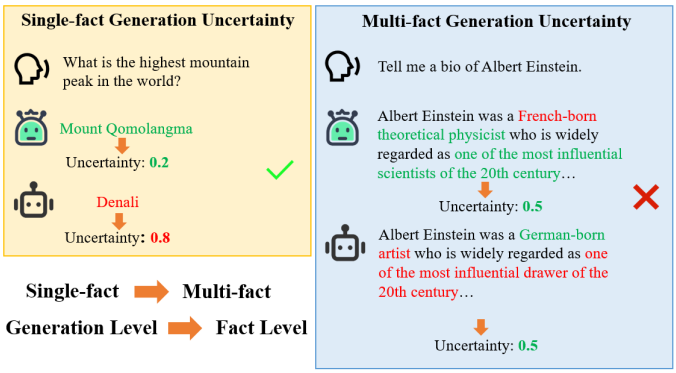


图 1. 单事实生成与多事实生成在不确定性量化方面的差异。

可能会生成表面上连贯且具有说服力的内容。例如，在生成真实人物传记的任务中，它们可能会生成包含错误或虚假事实的输出。或者，当用户无意中提示模型生成虚构人物的传记时，模型可能会像处理常规任务一样继续执行。鉴于消除大语言模型中的幻觉难以实现，我们可以通过从外部测量模型生成的输出的不确定性来解决这一问题。通过突出显示不确定性较高的答案，我们可以提醒用户注意潜在的不准确之处。

目前，已有若干方法用于量化大语言模型生成内容的不确定性。然而，这些方法通常在整段生成文本的层面上考量不确定性。它们依据生成文本的内容和生成词元的逻辑值信息进行计算，且主要用于核查单一事实 [6]。然而，当生成内容涉及多个事实时，现有方法仍难以准确衡量此类情形中的不确定性。如图 1 所示，即使存在事实性错误，测得的不确定性仍可能较低。

*Zhongliang Yang 是本文的通讯作者。我们的代码可在 https://github.com/EdwardChang5467/robust uncertainty 获取。

在自然语言处理（NLP）中，多事实生成是指生成包含多个相互独立事实的文本。评估多事实生成的不确定性比评估单事实生成更复杂。模型需要将生成文本拆解为各个事实，量化每项事实的不确定性，并在复杂语境中同时核查这些事实。此外，生成过程中出现的幻觉更难凭直觉发现，这进一步增加了评估的复杂性。以往研究主要关注基于真实世界数据的多事实生成，并对生成结果进行事实核查 [7]–[11]。然而，在实际应用中，大语言模型的用户可能会有意或无意地向模型提出错误的陷阱问题，例如要求模型生成虚构人物的传记。在这种情况下，模型可能会拒绝生成，也可能以一定概率为虚构人物编造传记，或将真实传记中的事实移植到虚构人物身上；因此，将生成内容拆解为相应的单项事实并不容易。我们认为，以往方法无法充分满足现实用户的需求，因为它们缺乏鲁棒性，面对用户的错误查询时也无法量化生成内容的不确定性。这使得多事实生成中的不确定性量化和事实核查面临更严峻的挑战。

我们的研究旨在基于陷阱问题量化大语言模型中多事实生成的不确定性。为此，我们利用陷阱问题构建了一个多事实生成场景。为了实现这一场景，我们提出了一套用于构建虚构传记生成任务数据集的流程。具体而言，我们创建了一个包含77个陷阱问题以及从四个大语言模型中采样得到的385个生成结果的数据集。此外，我们开发了一种稳健的不确定性估计方法 RU。通过与四个模型及多种基线方法进行比较，我们证明了所提出方法的优越性能。据我们所知，我们的工作是首项基于陷阱问题开展不确定性估计的研究。我们的方法通过细粒度分类与量化，有效解决了陷阱问题的不确定性量化问题。我们希望本研究能够推动多事实生成任务中的不确定性估计研究。

我们的贡献可总结如下：

- 我们提出了一套基于大语言模型构建陷阱问题的流程，并基于该流程构建了一个数据集$MulFactTrap$，其中包含77个陷阱问题，以及基于 $MulFactTrap$纳入的四个大语言模型的385个生成结果。
- 我们提出了一种新颖且鲁棒的大语言模型多事实生成白盒不确定性估计方法。
- 我们对所提出的陷阱问题构造方法和不确定性度量方法进行了实验验证。结果表明，这些方法在四种大语言模型上均表现出色。

II. 相关工作

大语言模型输出的不确定性量化可分为五类 [12]：基于Logit的方法 [13], [14],基于内部状态的方法 [15]–[17]，基于语言表达的方法 [18], [19],基于一致性估计的方法 [20], [21], 和基于代理模型的方法 [22], [23]。基于Logit的方法实现简单，输出直观。然而，在处理复杂的非线性问题时，这类方法可能无法准确捕捉数据中的复杂关系，从而导致欠拟合。基于内部状态的方法能够深入探究模型的内部空间，但计算复杂度较高，且高度依赖模型架构。基于语言表达的方法易于理解，但量化结果准确性较低。基于一致性估计的方法适用于黑盒和白盒模型，但计算成本较高。基于代理模型的方法计算效率高，但精度有限。

上述研究主要聚焦于单事实生成。相比之下，衡量多事实生成的不确定性更具挑战性，因为这需要更细的粒度和更高的准确率。过去两年间，研究人员提出了多种衡量多事实生成不确定性的方法。Fadeeva 等人 [11] 提出了主张条件概率（CCP）方法，通过调整并汇总模型生成词元的概率，得到主张级别的不确定性。Jiang 等人[24]引入了 CORE 组件，该组件依据模型生成主张的独特性和信息量对其进行筛选，并仅对筛选出的事实进行事实核查。这种方法在一定程度上增强了不确定性方法的鲁棒性。Vazhentsev 等人 [25] 提出了 TAD 方法，通过训练一个回归模型来调整当前生成词元的不确定性，该模型以条件生成置信度与无条件生成置信度之间的差值为目标变量。这种方法在长文本生成中尤为有效，但需要构建额外的训练数据，因此计算成本较高。然而，据我们所知，以往的方法尚未涉及陷阱问题的不确定性量化，也未研究多事实生成任务中对陷阱问题的答案的可信度。这令人担忧：现有不确定性量化方法应用于陷阱问题时，是否具备足够的鲁棒性。

III. 鲁棒不确定性量化与场景构建

本节将介绍第 III-A 节提出的不确定性测量场景，同时介绍第 III-B 节中为该场景构建不确定性量化所需数据集的方法。

A. 鲁棒不确定性量化场景描述

用于不确定性量化的典型数据集 [10][11], [26] 仅包含事实正确的问题，缺少棘手或误导性问题，因此在实际应用中用于衡量大语言模型的不确定性并据此评估其输出可信度时，鲁棒性不足。在现实场景中，人们可能有意（为了进行鲁棒性评估或恶意攻击）或无意（由于疏忽）向大语言模型提出错误或虚假问题。如果模型根据用户意图接受并回答这些问题，答案可能包含事实错误，不确定性量化也应反映出更高程度的不确定性。相反，如果模型拒绝这些问题或识别出其中的错误，不确定性量化则应得出较低的值。在此场景中，所使用的字母和符号及其对应含义如表 I 所示。

表 I 场景中使用的符号及含义

符号	含义
q	向大语言模型提出的问题
g	大语言模型根据问题生成的回复
t_j	大语言模型在第 j 步生成的词元
f_i	第 i 个事实可以分解
ϕ	f_i 与 t_j 之间的关系
U	不确定性量化函数
θ	区分高低不确定性值的阈值
F	由 g 分解得到的事实集合

在基于陷阱问题的大语言模型多事实生成不确定性量化中，设问题为 q，且大语言模型生成的由 m 个词元组成的输出为 $g = \{t_1, t_2, ..., t_m\}$。大语言模型的生成 g 可以通过外部模型或方法分解为 n个事实 F，即 $F = \{f_1, f_2, ..., f_n\}$。因此，每个事实 f_i 与大语言模型生成的词元之间存在映射关系 ϕ，如公式（1）所示，其中第 j步生成的词元记为 t_j。

$$\phi : f_i \mapsto \{t_j \in g \mid \text{token } t_j \text{ is relevant to } f_i\}. \quad (1)$$

假设不确定性分数记为 U。在此场景下，需要找到一个合适的函数 U，使其满足公式（2）所示的关系，其中θ 的取值可以根据实际情况确定。

$$\begin{cases} U(g) > \theta, & \text{if } g \text{ is generated} \\ U(g) \le \theta, & \text{if } g \text{ is refused to generate} \end{cases} \quad (2)$$

针对模型错误生成的问题，我们旨在确定一个合适的函数 U，使得 F 中的事实越多，U 的不确定性量化函数值就越高，如式（3）所示，其中 $|F|$ 表示事实集 F 的大小。

$$\begin{aligned} &U : F \to \mathbb{R}, \\ &s.t. \forall F_1, F_2,\ |F_1| > |F_2| \implies U(F_1) > U(F_2). \end{aligned} \quad (3)$$

B. MulFactTrap 数据集构建

现有的多事实生成数据集在衡量不确定性时，难以反映大语言模型的鲁棒性，也无法全面评估不确定性估计方法的性能。因此，我们构建了一个包含陷阱问题的新多事实生成数据集。该数据集既能根据大语言模型在其上的生成表现反映模型的鲁棒性，也可用于不确定性估计，以衡量不确定性估计方法的鲁棒性。

大语言模型在回答传记类问题时，通常会生成多个事实。例如，陈述“阿尔伯特·爱因斯坦是一位出生于德国的理论物理学家，被广泛认为是20世纪最具影响力的科学家之一”可以拆分为几个不同的事实，例如“阿尔伯特·爱因斯坦是出生于德国的人”和“阿尔伯特·爱因斯坦是一位理论物理学家”。因此，我们主要聚焦于传记类生成场景，以构建我们的陷阱问题多事实生成数据集。

首先，如图 2 所示，我们介绍一种使用大语言模型构建陷阱问题的流程。该流程由两个主要组件组成：LLM生成器（LLM_G）和LLM验证器（LLM_V），它们是用于生成和区分事实的大语言模型。LLM_G 负责创建潜在的虚假姓名，而 LLM_V则评估LLM生成器所生成姓名的真实性。

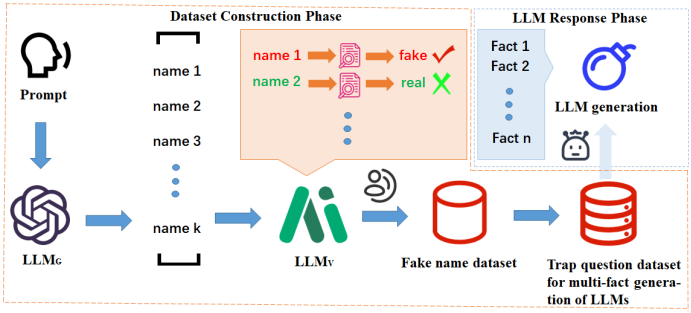


图 2. 我们数据集构建的整体框架。

大语言模型已通过实证证明具备生成传记的能力。这种能力主要源于模型在训练阶段接触了大量关于真实个体的信息。然而，在推理阶段，当提示模型生成一位身份存疑人士的传记时，模型产生幻觉的可能性会显著增加。为了全面评估大语言模型的鲁棒性，我们旨在构造与真实姓名高度相似的假名。为此，我们采用以下三种启发式策略来生成虚假姓名：排列与重组、词语微调和虚构世界创建。

- **置换与重组（PR）：** 该过程包括拆分姓名中各个词语的片段，并在相同位置将不同姓名中的片段重新组合。例如，经过转换后，姓名“唐纳德·约翰·特朗普”和“约瑟夫·罗比内特·拜登二世”会被重新组合成新的姓名“唐纳德·罗比内特·拜登二世”。

- **词语微调（WF）：** 修改并微调人名中的部分字母，同时确保修改后发音不会发生显著变化。例如，“阿尔伯特·爱因斯坦”。

- **虚构世界创建（FWC）：** 融入电子游戏或动漫等虚拟世界中的角色名称。例如，“吉伊卡哇”。

构建虚假姓名的流程分为三个步骤：虚假姓名生成、模型验证虚假姓名，以及人工微调虚假姓名。首先，我们手动构建提示词，指示 LLM_G 根据一系列真实姓名和扰动策略，生成一份包含 k 个潜在虚假姓名（L_{pf}）的列表。由于虚构世界中的角色名称可能与真实姓名重叠，因此通过 LLM_G生成姓名时，我们只使用前两种启发式策略。鉴于LLM生成器可能存在幻觉问题， L_{pf}中生成的姓名未必都是真正的虚假姓名。因此，我们向 LLM_V 提问“[姓名]是真实人物吗？”，以进行真实性验证。我们选取被验证为“虚假”的姓名，组成一份更精炼的潜在虚假姓名列表（L_{mpf}）。将第j个实例第t次验证的结果记为 $c_t^j \in \mathcal{C}$。经过T次独立验证后，最终判定 c^j 取$\{c_1^j, \ldots, c_T^j\}$的多数表决结果。若出现平票，则进行额外一轮验证，直至形成明确的多数意见。只有最终判定为“false”的实例才会保留，用于陷阱问题构建。最后，考虑到LLM验证器可能产生幻觉，我们会进行人工验证和微调，以确保最终数据集中虚假姓名的质量。我们还会手动排除与原始真实姓名发音差异显著的姓名，并采用虚构世界创建策略补充一些虚构世界中的角色名称。完成上述所有步骤后，我们便得到最终的虚假姓名数据集。

具体而言，我们使用从 FActScore [10] 收集的数据以及 CCP [11]实验中使用的真实传记数据，构建了虚假姓名数据集。我们从这些来源中随机选取了 100 个问题，提取其中的姓名，并提示 Yi-Lightning 生成 100 个可能的虚假姓名。经过筛选，我们得到一个包含 80 个虚假姓名的集合。经人工审查和少量调整后，我们最终确定了 77 个虚假姓名用于实验。

我们对 Yi-Lightning 模型通过单次采样生成的 100 个可能为虚假姓名的名称进行了事实核查。用于姓名核查的模型包括 Yi-Lightning[1], GPT-4o[2]，和 Kimi[3]。核查结果见表 II。

表 II 3 个模型的姓名核查性能

模型	准确率	召回率	F1 分数
Yi-Lightning	0.76	0.9091	0.8537
GPT-4o	0.7879	0.8947	0.8662
Kimi	0.7551	0.8831	0.85

如表 II 所示，Yi-Lightning、GPT-4o 和 Kimi 这三个模型在虚假姓名验证任务中均表现出色。模型所提供标签的平均准确率为 0.77，平均召回率为 0.90，平均 F1 分数为 0.85。虽然 GPT-4o 在三者中表现最佳，但其使用受到令牌数量限制的约束，而且 GPT-4o 和 Kimi 的推理成本都较高。因此，在构建数据集时，我们选择了成本效益高且表现出色的 Yi-Lightning 模型来验证虚假姓名。

随后，我们基于这些名称，在虚构名称前添加短语“Tell me a bio of”或“Tell me a brief introduction of”，以获得陷阱问题。由此得到的陷阱问题数据集可用于评估大语言模型的多事实生成能力并衡量其不确定性。在这种情况下，大语言模型在该数据集上生成幻觉的概率显著高于在典型数据集上的概率。

最后，在大语言模型生成阶段，我们将数据集中的每个问题输入待评估的大语言模型或其应用程序接口，以评估生成不确定性。通过使用构造的陷阱问题诱发大语言模型产生幻觉，再使用 GPT-4 等表现出色的模型将模型生成的输出分解为事实，我们可以得到一个包含多事实及生成文本的大语言模型陷阱问题数据集。MulFactTrap 包含 77 个精心设计的陷阱问题，用于生成虚假传记，其中 35% 采用 PR 策略构建，58% 采用 WF 策略构建，7% 采用 WFC 策略构建。此外，MulFactTrap 还包含从四个大语言模型收集的 385 个模型生成的数据样本。在事实分解过程中，我们采用了双样例提示。第一个样例指导大语言模型如何进行事实分解，第二个样例则要求模型在分解过程中不要改动原始生成内容中的虚假和错误事实。这两个样例见图 3。

C. 鲁棒不确定性量化方法

为更好地适应大语言模型生成在陷阱问题上的不确定性测量场景，我们提出了一种鲁棒不确定性测量方法，称为 RU。该方法

[1]https://platform.lingyiwanwu.com/chat

[2]https://chatgpt.com/ [3]
https://kimi.moonshot.cn/

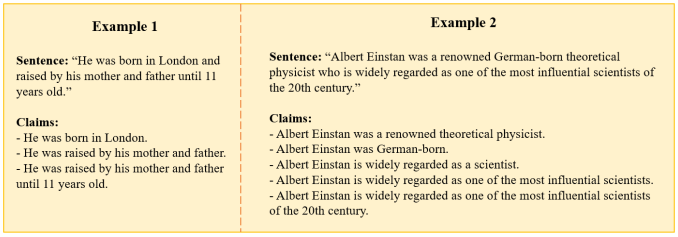


图 3. 事实分解过程的两个示例。

通过根据类别以细粒度方式分解和聚合事实层面的不确定性，实现了更鲁棒、更准确的不确定性量化。 RU 方法的框架如图 5 所示。

我们采用多重采样策略，使用 RU 获取不确定性分数，其目的是实现丰富的生成结果和语义多样性。主要原因在于，多重采样不仅能够全面捕捉模型生成内容中的潜在语义信息，还能让我们采样更多生成数据，即使问题数量有限。这种方法能够更准确地评估不确定性度量的性能。在我们的多重采样方法中，我们结合使用随机采样和 top-k采样策略。k 的值保持模型的默认设置，采样温度则保持不变。我们采用束搜索作为主要解码策略，生成的词元数量上限设为 100。

收集模型生成的内容后，我们将模型输出分为三类：事实识别正确（FR）、事实纠正为真（FT）和事实纠正为假（FF）。FR 表示大语言模型正确识别出问题中的错误或虚假内容，并拒绝作答。FT 表示大语言模型未能识别错误，但在生成过程中将其纠正为真实信息。FF 表示大语言模型未能识别虚假信息，且要么未予纠正，要么将其替换为其他虚假或错误内容。通过构造提示来引导大语言模型（例如 GPT-4 或 Yi-Lightning），即可获得分类标签；这种方法具有较强的通用性。在多事实生成场景中，分类算法的选择应根据生成内容的具体特点进行调整。此外，使用检索器在外部数据集中查找佐证事实，有助于增强对生成信息的验证。例如，在基于虚假姓名陷阱问题对大语言模型生成内容进行不确定性量化时，我们首先使用 DeBERTa 等自然语言推理模型，根据生成文本判断其与表达“很遗憾，我无法提供.....的信息”之间的逻辑关系。如果结果为蕴涵，则将其标记为“FR”。我们使用维基数据作为外部数据，并采用基于 BM25检索算法的检索器搜索维基数据。如果检索到相关人物的描述，则将标签标记为“FT”；否则，将其归类为“FF”。

在将模型生成内容分解为事实并进行词元映射时，我们遵循先前研究的方法，使用提示引导高性能大语言模型 Yi-Lightning 的应用程序接口完成这项任务。与前文介绍的方法类似，事实分解也采用双样本提示策略。在词元映射过程中，我们将完整的生成内容和分解得到的各条事实输入 Yi-Lightning 提示中的大语言模型，让其列出原始生成内容中与每条事实对应的词语。随后，我们使用原始模型的分词器将词语与词元进行匹配，并获取它们在原始生成内容中的位置。最后，我们得到原始生成内容中与每条事实对应的所有词元位置列表。最后，我们根据大语言模型生成内容的逻辑值信息、类别标签，以及事实与生成内容之间的映射关系，计算 RU 结果。

我们继续采用第 III-A 节提出的场景，并假设模型的生成结果表示为 $g = \{t_1, t_2, ..., t_m\}$。由于大语言模型的自回归生成，第 j 步生成的词元的概率表示为 $p(t_j|t_{<j})$ ($i = 1, 2, ... m$)。为便于计算和分析，我们将该概率近似并简化为 $p(t_j)$。对于标记为 FR 的输出，我们使用 (4) 计算生成结果的不确定性。对于标记为 FT 的输出，我们使用 (5) 计算事实级不确定性；对于标记为 FF 的输出，则使用 (6) 计算，其中 $norm(。)$ 表示长度归一化。

$$RU_{FR}(g) = 1 - e^{norm(\sum_{t_j \in g} logp(t_j))}, \tag{4}$$

$$RU_{FT}(f_i) = 1 - e^{norm(\sum_{t_j \in \phi(f_i)} logp(t_j))}, \tag{5}$$

$$RU_{FF}(f_i) = e^{norm(\sum_{t_j \in \phi(f_i)} logp(t_j))}. \tag{6}$$

IV. 实验与结果

本节重点研究涉及虚构人物的传记文本生成的不确定性量化。我们将介绍基于陷阱问题的不确定性量化实验设置（第 IV-A 节），以及实验结果及其分析（第 IV-B 节）。

A. 实验设置

1) 数据： 我们使用第 III-B 节介绍的 MulFact- 陷阱数据集，对大语言模型生成输出的不确定性量化进行了实验。

2）模型：我们使用以下四种大语言模型开展不确定性测量实验：LLaMA3-8B-Instruct [27], Vicuna-13B [28],ChatGLM3-6B-32K [29], 和 Mistral-7B [30]。

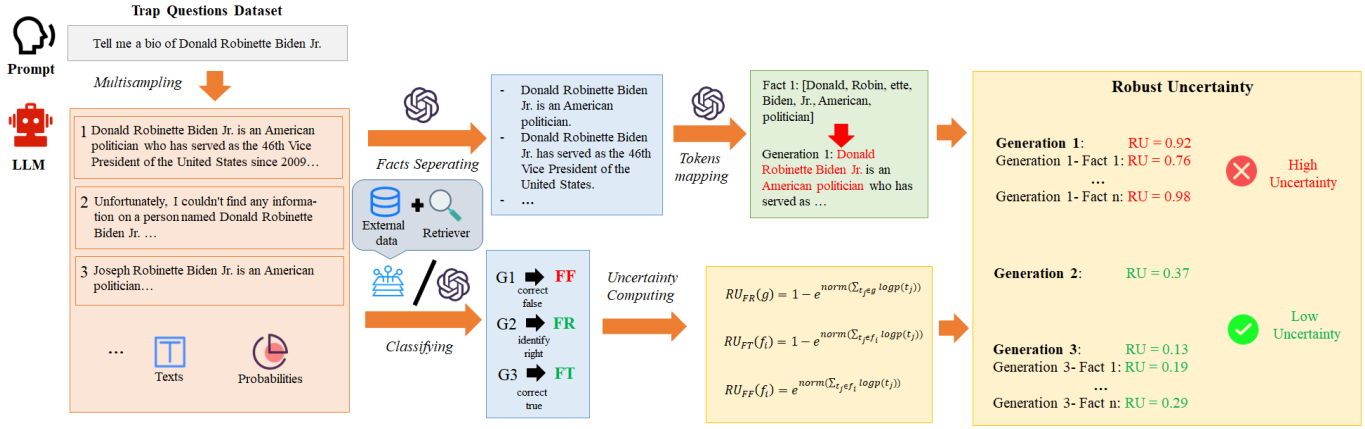


图 4. 所提出 RU 方法的整体框架。

3）基线方法：我们主要采用以下四种方法作为 RU 的基线方法：预测熵 [16],长度归一化预测熵 [31]，最大概率 [11] 和 CCP [11]。由于最大概率和 CCP 方法基于事实粒度[11]，，我们使用在事实粒度上计算的不确定性的平均值和最大值来表示生成粒度上的不确定性。

预测熵： 我们采用传统的预测熵作为基线。具体而言，我们将整个生成过程 g 中生成的第 j个词元的概率记为$p(t_j|t_{<j})$。预测熵定义为整个生成过程的熵值。

$$PE(g) = \sum_{t_j \in g} p(t_j|t_{<j}) logp(t_j|t_{<j}). \tag{7}$$

长度归一化预测熵： 基于生成粒度，按长度对预测熵进行归一化，得到长度归一化预测熵（LNPE）。

$$LN-PE(g) = norm(\sum_{t_j \in g} p(t_j|t_{<j}) logp(t_j|t_{<j})). \tag{8}$$

最大概率： 我们直接将生成概率中最高的值视为置信度分数，并在事实粒度层面进行聚合。

$$MP(f_i) = 1 - \prod_{t_j \in f_i} p(t_j|t_{<j}). \tag{9}$$

主张条件概率：CCP 方法首先使用自然语言推理模型[4] 计算模型词汇表中与原始词语存在蕴涵或矛盾关系的词语的概率。此过程会得到词级主张条件概率分数。随后，计算这些词级主张条件概率分数的乘积，以获得主张级 CCP。

[4]https://huggingface.co/microsoft/deberta-large-mnli

$$CCP_{word}(t_j) = \frac{\sum_{k:\text{NLI}(t_j^k, t_j)=\text{'e'}} P(t_j^k \mid t_{<j})}{\sum_{k:\text{NLI}(t_j^k, t_j) \in \{\text{'e'}, \text{'c'}\}} P(t_j^k \mid t_{<j})}, \tag{10}$$

$$CCP_{claim}(f_i) = 1 - \prod_{t_j \in f_i} CCP_{word}(t_j). \tag{11}$$

4）评估指标：我们采用准确率、召回率和 F1 分数作为模型事实核查步骤的评估指标，其中真实标签通过人工检索获得。与先前研究一致，我们使用 ROC-AUC分数（ROC）、皮尔逊相关系数（PC）和斯皮尔曼相关系数（SC）作为不确定性量化的指标。目前，还没有全面的评估标准来衡量大语言模型的鲁棒性。因此，我们通过提示 Yi-Lightning 判断生成内容中是否存在幻觉，以获取正确性标签。

B. 结果与分析

1）不确定性量化：我们使用四个模型基于构建的数据集进行多重采样，采样规模为 5。生成过程中，束宽也设为 5，采样温度设为 1.0。对于每个生成样本，我们通过特征提取获得相应的标签。我们使用 Yi-Lightning 模型进行事实分解和词元映射。所提出的 RU 方法与基线方法的不确定性评估结果见表 III。具体而言， RU_{gen} 表示将事实对齐到生成粒度，而 RU_{fact} 表示将生成内容对齐到事实粒度。 CCP_{max} 和 $Maxprob_{max}$ 表示基于所有事实中的最大不确定性计算的整个生成内容的不确定性。 CCP_{mean} 和 $Maxprob_{mean}$表示基于所有事实的平均不确定性计算的整个生成内容的不确定性。

为更贴近现实场景中的不确定性度量，我们扩充了 MulFactTrap 数据集

加入了 50 个与生成真实传记相关的额外问题，形成包含 127 个问题的混合数据集，涵盖真实传记和虚假传记生成任务。我们使用该混合数据集，对四个大语言模型进行了不确定性度量实验。各方法的性能如表 IV 所示。

根据表 III 和表 IV 中呈现的结果，我们可以得出以下结论：首先，所提出的 RU 方法在绝大多数情况下优于所有参与比较的基线方法。RU 的平均 ROC-AUC 值比基线方法高 0.1 至 0.2。此外，在皮尔逊相关系数和斯皮尔曼相关系数方面，RU 均显著优于基线方法。其次，在四种基线方法中，CCP 和最大概率通常表现更好。这表明，对事实进行细粒度分解并在事实层级衡量不确定性，能够提升不确定性估计的整体性能和鲁棒性。最后，与 CCP 和最大概率相比，我们的 RU 方法通过针对不同类型的生成引入额外的分类和细粒度计算，进一步提升了性能。这种更精细的分类与计算方法进一步增强了不确定性估计的性能。结果表明，在大多数扰动场景下，生成层级 RU 和事实层级 RU 均保持了良好的校准表现。基线方法的相关性明显较低，有些方法在受扰动的数据集上甚至得出负值，表明它们对干扰和对抗攻击的鲁棒性有限。相比之下，RU 利用事实的细粒度分解和不确定性感知分类，成功检测出模型生成中的陷阱问题和潜在事实性幻觉。

2) 大语言模型对陷阱问题的鲁棒性：我们计算了在 MulFactTrap 中生成“幻觉”（标记为“FF”）的概率。结果如图 5 所示。

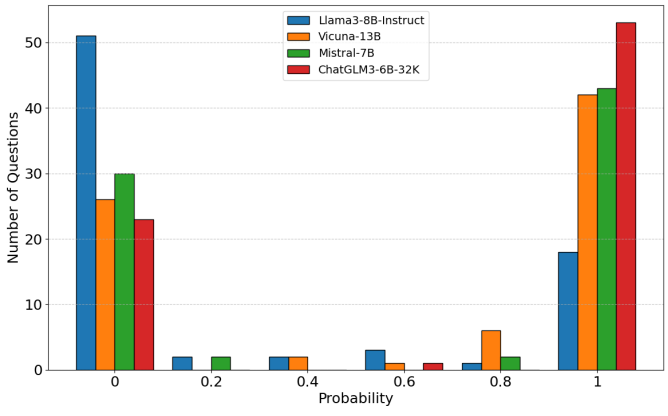


图 5 。模型生成中幻觉的概率分布 结果。

图 5 所示的分布揭示了这些模型面对虚构传记时，产生幻觉的概率呈现双峰趋势。这表明，当面对同一欺骗性任务时，单个模型更可能给出完全正确或完全错误的输出。在不同模型中，Llama3-8b-instruct 模型产生幻觉的平均概率较低，表明其鲁棒性更强。相比之下，ChatGLM3-6B-32K 模型产生幻觉的概率较高，说明其鲁棒性相对较弱。我们的分析得出三项主要结论：面对经过扰动的问题时，受评估的模型极易产生幻觉；指令微调变体的幻觉率低于其基础版本，我们将这一现象归因于对齐流程将训练数据中隐含的人类偏好内化；在受到扰动时，小模型比大模型更容易屈服于幻觉。

C. 消融研究

我们针对虚构名称验证任务开展了消融研究，重点考察采样规模 k 以及在名称验证过程中是否采用思维链（CoT）策略。

1) 采样规模 k：我们研究了 Yi-Lightning 模型中不同采样规模 k 对姓名验证性能的影响，结果如表 IV 所示。对于涉及多个样本的情况，我们采用多数投票策略，选择生成域中出现频率最高的标签作为最终标签。

如表 V 所示，当样本数量相对较少时，人物验证任务的性能会随着样本规模的增加而逐步提升k，但提升速度较慢。性能在 k = 5时达到峰值，进一步增加 k 并不会带来显著改善。鉴于较大的采样规模需要更多时间，我们未对更大的 k值进行实验。

2) CoT 策略：我们进行了实验，以评估样本规模设为 3 时，采用和不采用思维链（CoT）策略的 Yi-Lightning 模型的验证性能。不采用 CoT 策略时，模型的验证准确率从 0.77 降至 0.65，召回率从 0.9221 降至 0.7143，F1 分数也相应地从 0.8606 降至 0.7586。这些结果清楚地表明，CoT 策略能够显著增强模型的验证能力。

V. 结论与未来工作

我们提出了一种基于多事实生成、针对陷阱问题的不确定性量化新任务场景，并对其进行了正式定义。该场景在扩展先前工作的基础上，同时考虑了模型的鲁棒性。我们提出了一套流程，用于构建陷阱问题数据集 MulFactTrap，以实现这一不确定性量化场景，并通过实验验证其有效性。此外，我们提出了一种用于不确定性度量的新型鲁棒方法 RU，并证明了它在四种模型上的优异性能。与当前最先进方法相比，我们的方法在指标上提升了约 0.1–0.2。

表 III RU 和基线方法在 MulFactTrap 的 4 个模型上的性能

模型	LlaMA3-8B-Instruct			Vicuna-13B			ChatGLM3-6B-32K			Mistral-7B		
指标	ROC	PC	SC	ROC	PC	SC	ROC	PC	SC	ROC	PC	SC
PE [16]	0.5353	0.1053	0.0589	0.6798	0.2737	0.2911	0.6511	0.1407	0.1984	0.7469	0.4025	0.4194
$LN-PE$ [31]	0.5022	0.0561	-0.0036	0.5268	-0.0956	-0.0433	0.677	0.2042	0.2324	0.7467	0.4017	0.4191
CCP_{mean} [11]	0.7457	0.4196	0.4203	0.6835	0.3465	0.3178	0.8028	0.3622	0.393	0.7322	0.3879	0.3945
CCP_{max} [11]	0.7301	0.3645	0.3936	0.6491	0.2418	0.2582	0.7608	0.2872	0.3384	0.7811	0.4047	0.4775
$Maxprob_{mean}$ [11]	0.735	0.4289	0.4021	0.7099	0.3459	0.3634	**0.8558**	0.3862	**0.4617**	0.7249	0.3451	0.3822
$Maxprob_{max}$ [11]	0.7521	0.3957	0.4312	0.6606	0.2251	0.2781	0.8191	0.3658	0.4141	0.7106	0.2907	0.3577
RU_{gen}（我们的方法）	0.7442	**0.5432**	0.4075	**0.8660**	**0.7314**	**0.5925**	0.7775	**0.4609**	0.3644	**0.8876**	**0.7868**	**0.6584**
RU_{fact}（我们的方法）	**0.8195**	**0.6156**	**0.5267**	**0.8479**	**0.6998**	**0.5214**	0.7647	**0.4256**	0.3518	**0.898**	**0.7479**	**0.6684**

表 IV RU 与基线方法在混合数据集上 4 个模型中的性能

模型	LlaMA3-8B-Instruct			Vicuna-13B			ChatGLM3-6B-32K			Mistral-7B		
指标	ROC	PC	SC	ROC	PC	SC	ROC	PC	SC	ROC	PC	SC
PE [16]	0.5093	0.019	-0.0141	0.6571	0.2674	0.2702	0.7359	0.3981	0.399	0.7052	0.3036	0.332
$LN-PE$ [31]	0.5141	0.0127	-0.0214	0.6042	0.1476	0.1792	0.7366	0.3997	0.4002	0.7263	0.3608	0.3662
CCP_{mean} [11]	**0.8025**	0.5283	**0.5076**	0.7119	0.3988	0.364	0.6774	0.3108	0.2835	0.6793	0.3061	0.3106
CCP_{max} [11]	0.7907	0.4591	0.4879	0.7183	0.3617	0.375	0.7354	0.3457	0.3762	0.7349	0.3465	0.4068
$Maxprob_{mean}$ [11]	0.7853	**0.5486**	0.4787	0.7467	0.4393	0.4239	0.7229	0.3519	0.3561	0.724	0.3467	0.388
$Maxprob_{max}$ [11]	0.7804	0.5237	0.4705	0.7226	0.3508	0.3825	0.7121	0.3512	0.3389	0.7413	0.3712	0.4179
RU_{gen}（我们的方法）	0.6905	0.4929	0.2908	**0.9001**	**0.7941**	**0.6883**	**0.7636**	**0.4854**	**0.4232**	**0.8562**	**0.7304**	**0.6024**
RU_{fact}（我们的方法）	0.7393	0.5304	0.3495	**0.899**	**0.7874**	**0.6882**	**0.7515**	**0.4347**	0.396	**0.8753**	**0.7145**	**0.6394**

表 V 不同样本规模下的姓名检查结果

样本规模	准确率	召回率	F1 分数
1	0.76	0.9091	0.8537
3	0.77	0.9221	0.8606
5	0.78	0.9351	0.8675
7	0.77	0.9221	0.8606

尽管本文的方法和实验结果令人满意，但仍可能存在一些局限性。首先，本文提出的不确定性度量方法 RU 尚未在其他陷阱问题场景中得到验证，其可迁移性可能受到限制。其次，本文所研究的模型在面对陷阱问题时的鲁棒性评估可能不够全面。我们需要一种更整体的评估标准来衡量模型的鲁棒性。最后，借助基于所提数据集构建的流程，并纳入更多真实姓名，本文的陷阱数据集能够实现良好的可扩展性。未来，数据集的规模可以进一步扩大，以更好地服务于下游自然语言处理任务。考虑这些方面可能会进一步提升不确定性量化的性能。

致谢

本研究部分得到了国家重点研发计划（项目编号：2023YFC3305402）的资助，部分得到了国家自然科学基金（项目编号：62302059 和 62172053）的资助。我们还要感谢研究组的博士生 Qinhong Lin，他为本文初稿提出了宝贵建议，并就实验进行了富有见地的讨论。

参考文献

[1] P. Lewis、E. Perez、A. Piktus、F. Petroni、V. Karpukhin、N. Goyal 等，《面向知识密集型自然语言处理任务的检索增强生成》，神经信息处理系统进展，第33卷，9474，2020年，9459页。[2] D. Chen、Q. Zhang 和 Y. Zhu，《利用大语言模型进行高效序贯决策》。载于《2024年自然语言处理实证方法会议（EMNLP）论文集》，第9157–9170页，迈阿密，佛罗里达州，美国。计算语言学协会（ACL），2024年。[3] Z. Xu、S. Jain 和 M. Kankanhalli，《幻觉不可避免：大语言模型的固有局限》，2024年，arXiv:2401.11817。[在线]。获取地址：https://arxiv.org/abs/2401.11817。

[4] L. Huang、W. Yu、W. Ma、W. Zhong、Z. Feng、H. Wang 等，《大语言模型幻觉综述：原理、分类、挑战与开放问题》，ACM 信息系统汇刊，2023年。[5] B. Snyder、M. Moisescu 和 M.B. Zafar，《事实性问答中幻觉的早期检测》，第30届 ACM SIGKDD 知识发现与数据挖掘大会论文集，2024年。[6] H. Sun、H. Cai、B. Wang、Y. Hou、X. Wei、S. Wang 等，《借助演进式记忆与自我反思实现可验证文本生成》，2023年，arXiv:2312.09075。 [在线]。获取地址：https://arxiv.org/abs/2312.09075。[7] Z. Guo、M. Schlichtkrull 和 A. Vlachos，《自动化事实核查综述》，计算语言学协会汇刊，第10卷，第178–206页，2022年。[8] P. Nakov、G. D. S. Martino、T. Elsayed、A. Barr´on-Cedeno、R. M´ıguez、S. Shaar 等，《CLEF2021 CheckThat! 实验室：检测值得核查的主张、先前已核查事实的主张和假新闻》，载于《信息检索进展：第43届欧洲信息检索研究会议（ECIR 2021），线上活动，2021年3月28日至4月1日，论文集，第二部分43》，第639–649页。施普林格，2021年。[9] D. Wadden、S. Lin、K. Lo、L. L. Wang、M. Zuylen、A. Cohan 和 H. Hajishirzi，《事实还是虚构：科学主张核验》，载于《2020年自然语言处理经验方法会议（EMNLP）论文集》，第7534–7550页，2020年。[10] S. Min、K. Krishna、X. Lyu、M. Lewis、W. Yih、P. W. Koh 等，《FActScore：长篇文本生成中事实精确性的细粒度原子级评估》，《2023年自然语言处理经验方法会议（EMNLP）论文集》，2023年。[11] E. Fadeeva、A. Rubashevskii、A. Shelmanov、S. Petrakov、H. Li、H. Mubarak 等，《通过词元级不确定性量化对大语言模型的输出进行事实核查》，2024年，arXiv:2403.04696。[在线]。获取地址：https://arxiv.org/abs/2403.04696。[12] J. Geng、F. Cai、Y. Wang、H. Koeppl、P. Nakov 和 I. Gurevych，《大语言模型中的置信度估计与校准综述》，《2024年计算语言学协会北美分会会议：人类语言技术论文集》（第1卷：长篇论文），2024年。[13] J. Duan、H. Cheng、S. Wang、A. Zavalny、C. Wang、R. Xu 等，《将注意力转向相关性：迈向大语言模型的不确定性估计》，2023年，arXiv: 2307.01379。 [在线]。获取地址：https://arxiv.org/abs/2307.01379。[14] L. Kuhn、Yarin Gal 和 Sebastian Farquhar，《语义不确定性：用于自然语言生成不确定性估计的语言不变性》，2023年，arXiv: 2302.09664。 [在线]。获取地址：https://arxiv.org/abs/2302.09664。[15] J. Ren、J. Luo、Y. Zhao、K. Krishna、M. Saleh、B. Lakshminarayanan 和 P. J. Liu，《条件语言模型的分布外检测与选择性生成》，NeurIPS 2022 序列建模鲁棒性研讨会，2022年。[16] S. Kadavath、T. Conerly、A. Askell、T. Henighan、D. Drain、E. Perez 等，《语言模型（大多）知道自己知道什么》，2022年，arXiv: 2207.05221。 [在线]。获取地址：https://arxiv.org/abs/2207.05221。[17] C. Burns、H. Ye、D. Klein 和 J. Steinhardt，《无监督地发现语言模型中的潜在知识》，载于《第十一届国际学习表征会议（ICLR）》，2023年。[18] S. J. Mielke、A. Szlam、E. Dinan 和 YL. Boureau，《通过语言校准降低对话代理的过度自信》，计算语言学协会汇刊，第10卷，第857–872页，2022年。[19] M. Xiong、Z. Hu、X. Lu、Y. Li、J. Fu、J. He 和 B. Hooi，《大语言模型能否表达自身的不确定性？大语言模型置信度引出的实证评估》，2023年，arXiv: 2306.13063。 [在线]。获取地址：https://arxiv.org/abs/2306.13063。[20] P. Manakul、A. Liusie 和 M. JF Gales，《SelfCheckGPT：用于生成式大语言模型的零资源黑盒幻觉检测》，2023b，arXiv: 2303.08896。 [在线]。获取地址：https://arxiv.org/abs/2303.08896。[21] Z. Lin、S. Trivedi 和 J. Sun，《自信地生成：黑盒大语言模型的不确定性量化》，2023年，arXiv: 2305.19187。 [在线]。获取地址：https://arxiv.org/abs/2305.19187。[22] V. Shrivastava、P. Liang 和 A. Kumar，《羊驼模型知道 GPT模型没有展示的内容：用于置信度估计的代理模型》，2023年，arXiv: 2311.08877。 [在线]。获取地址：https://arxiv.org/abs/2311.08877。[23] H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix 等，《Llama：开放且高效的基础语言模型》，2023年b，arXiv: 2302.13971。 [在线]。链接：https://arxiv.org/abs/2302.13971。[24] Z. Jiang、J. Zhang、N. Weir、S. Ebner、M. Wanner、K. Sanders 等，《CORE：通过信息丰富的子主张识别实现稳健的事实精确性》，2024年，arXiv:2407.03572。 [在线]。链接：https://arxiv.org/abs/2407.03572。[25] A. Vazhentsev、E. Fadeeva、R. Xing、A. Panchenko、P. Nakov、T. Baldwin 等，《无条件真实性：学习条件依赖关系以量化大语言模型的不确定性》，2024年，arXiv:2408.10692。 [在线]。链接：https://arxiv.org/abs/2408.10692。[26] Y. Feng、P.M. Htut、Z. Qi、W. Xiao、M. Mager、N. Pappas 等，《DiverseAgentEntropy：通过多元视角和多智能体交互量化黑盒大语言模型的不确定性》，2024年，arXiv:2412.09572。[在线]。链接：https://arxiv.org/abs/2412.09572。[27] C. Ackerman 和 N. Panickssery，《检查并控制 Llama3-8b-Instruct 对自生成文本的识别能力》，2024年，arXiv:2410.02064。 [在线]。链接：https://arxiv.org/abs/2410.02064。[28] L. Zheng、WL. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang 等，《使用 MT-Bench 和 Chatbot Arena 评判大语言模型裁判》，神经信息处理系统进展（NIPS）36，46623，2023年，第46595页：-。[29] T. GLM，《ChatGLM：从 GLM-130B 到 GLM-4 全工具的一系列大语言模型》，2024年，arXiv:2406.12793。 [在线]。链接：https://arxiv.org/abs/2406.12793。[30] A. Q. Jiang、A. Sablayrolles、A. Mensch、C. Bamford、D. S. Chaplot、D. de las Casas 等，《Mistral 7B》，2023年，arXiv：2310.06825。 [在线]。链接：https://arxiv.org/abs/2310.06825。[31] A. Malinin 和 M. Gales，《自回归结构化预测中的不确定性估计》，2020年，arXiv:2002.07650。 [在线]。链接：https://arxiv.org/abs/2002.07650。