arXiv:2601.00641v1 [cs.CL] 2 Jan 2026

降低大语言模型上下文幻觉的概率保证

NilsRautenberg* SvenSchippkus 德国精算师协会
汉堡大学

2026年1月5日

要点

- 独立重复提示词，极有可能得到正确答案。
- 大语言模型评审器可识别正确答案；必要时，多数投票能提升可靠性。
- 总而言之，这些方法提供了明确的概率保证，错误率会呈指数级下降。

摘要

大型语言模型（大语言模型）经常会产生上下文幻觉，即生成的内容与提示词中明确陈述的信息相矛盾或将其忽略。这类错误在确定性自动化工作流中尤为棘手，因为输入是固定的，正确性也没有歧义。我们提出了一个简单且模型无关的框架，用于 *explicit probabilistic guarantees* 降低此类场景中的幻觉。

我们形式化定义了 *specific task*，其具有固定输入和确定性正确性标准，并证明，在独立上下文窗口中发出相同的提示词，会使所有模型输出均不正确的概率呈指数级下降。为从多次运行中找出正确答案，我们引入大语言模型评审器，并证明，评审流程失败的概率会以由评判器的真正例率和假正例率决定的速率下降。当评判器并不完美时，我们通过对独立评审器调用进行多数投票来增强其能力，从而使集成级错误率随投票次数增加而呈指数级下降。这为流程选中幻觉答案的概率提供了明确的上界。

在使用合成噪声评审器进行受控抽取任务的实验中，结果与这些预测完全吻合：流程失败会随着重复次数增加而呈指数级下降，幻觉选择会随着集成中的评判器数量增加而呈指数级下降。综合来看，这些结果提供了一种轻量、模块化且有理论依据的方法，可使固定输入大语言模型工作流中的幻觉概率任意趋近于零，而无需修改模型权重、解码策略或提示工程。

1 引言

大型语言模型（LLM）越来越擅长解决复杂任务、遵循指令，并在长链条推理过程中综合信息。然而，即便是最先进的模型，仍会以一种出人意料地类似人类的方式出错：它们偶尔会断言一些不真实、并非

*联系邮箱：nils.rautenberg@rub.de

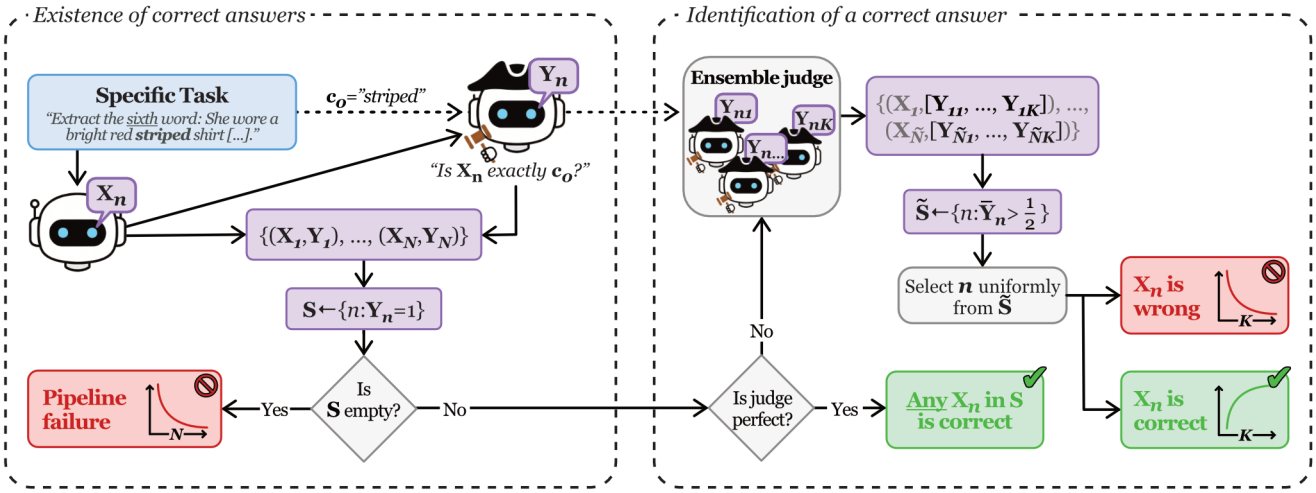

图1：重复判断流水线为减少上下文幻觉提供概率保证。**左：**对于给定的特定任务，通过将任务重复 N 次，并判断每次输出 X_n 的正确性，可以降低所有答案都产生幻觉的可能性。随着 N增加，流水线失败（“所有答案都被判定为不正确”）的概率呈指数级下降。**右：**从被判定为正确的答案中，可以按保证的概率找到真正的正确答案。如果评判器是完美的（$q^{++} = 1,\ q^{-+} = 0$），任何满足 $Y_n = 1$ 的 X_n 都是正确答案。如果评判器存在噪声但优于随机判断（$q^{-+} < 0.5$），集成判别器选出幻觉的概率会随着评判器数量 K的增加而呈指数级下降。

由所提供的信息推导得出，或与这些信息直接矛盾的内容。这些幻觉给要求确定性正确性的自动化流水线带来了挑战。

越来越多的研究表明，幻觉具有多种形式，而非单一形式。综合性综述（Huang 等，2025年；Zhang 等，2025年）梳理了多样的错误类型——从与世界知识相矛盾，到逻辑错误或指令混淆——揭示出幻觉是当今模型的一种结构性属性，而非罕见的故障（Kalai 等，2025年）。摘要研究进一步表明，即使是条件生成系统，也常常会产生无依据的主张（Fabbri 等，2021年；Kryscinski 等，2020年；Maynez 等，2020年），而检索增强模型仍会引入检索到的段落中不存在的内容（Adlakha 等，2024年）。上下文幻觉，即模型违背或忽略提示词中明确提供的信息，已被证明尤其难以消除。既有研究将这些失败归因于长上下文退化（Chang 等，2024年；Kim 等，2024年；Liu 等，2025年）、上下文知识与参数化知识之间的竞争（Longpre 等，2021年），以及在自回归生成中进行可靠证据追踪的固有困难。

目前已有不少工作致力于检测和缓解幻觉。FactCC（Kryscinski 等人，2020年）、AlignScore（Zha 等人，2023年）、REFChecker（Hu 等人，2024）和 Veri- FastScore（Rajendhran 等人，2025年）等技术尝试事后识别缺乏依据的内容；而对比解码策略（Chang 等人，2023年；Shi 等人，2023年）以及受忠实性启发的启发式方法（Wan 等人，2023年）则旨在减少推理过程中的幻觉。多智能体优化和大语言模型评判器系统（Cohen 等人，2023年；Du 等人，2023年；Jain 等人，2023年；Liu 等人，2023年；Luo 等人，2023年；Madaan 等人，2023年；Shinn 等人，2023年；Wang 等人，2024）通过多次调用模型来提升可靠性，但实证研究表明，在具有挑战性的案例中，评判器本身也可能前后不一或存在偏见（Kim 等人，2024；Laban 等人，2023年）。尽管取得了显著进展，现有

缓解策略并未针对残余幻觉发生的可能性提供明确的定量保证。

本文聚焦于一种范围受限但在实践中十分重要的场景：一类我们称为特定任务的任务。特定任务具有以下特点：(i) 输入固定；(ii) 每次运行只发出一次提示词；(iii) 正确性可通过确定性判定标准进行评估。典型示例包括槽位提取、参考资料查找、格式化任务，或在已知文本中定位特定的词语或数字。这类任务常见于自动化工作流中，并提供了一种正确性明确无歧义的清晰场景。在这种场景下，可靠性自然分解为两个相互独立的问题：(i) 模型输出中至少出现一个正确答案的可能性有多大；(ii) 当存在这样的答案时，评判器能多可靠地识别出它。这两个阶段——答案的出现与识别——可以分别通过重复和评判式选择加以改进。

以下两个简单观察构成了我们框架的动机：

1. **重复可以提升正确性**。即使模型有时会给出错误答案，独立地多次运行同一任务（例如，每次调用都使用新的上下文窗口），也会大大提高至少有一个答案正确的可能性。这是独立性带来的直接概率结果，我们将其形式化为*重复引理*。

2. **识别需要评判器**。 仅靠重复并不能告诉我们*哪一个* 回答是正确的。从重复运行中均匀随机选择一个答案，其正确概率与单次运行相同。要从所有出现的答案中选出正确答案，我们需要一种辅助机制来过滤幻觉：一个*大语言模型评审器*。即使评判器并不完美，也可以通过重复本身提升其可靠性。

图 1 基于这些核心概念展示了重复-评审流水线。重复使得至少有一个正确答案出现在模型输出中的可能性极高，而适当的评判器——必要时通过多数投票加以强化——则会以取决于任务的真阳性率和假阳性率识别出这样的答案。综合来看，这些组件构成了一种轻量级、模型无关的流程，能够在不修改模型权重或依赖复杂解码方案的情况下，在固定输入任务上实现任意低的错误率。

本文其余部分将在一个简单的形式化设定中展开这些思想。第 2 节介绍 *特定任务* ，并展示独立重复如何显著降低所有回答均错误的可能性。第 3 节随后解释，一旦重复使正确答案很可能存在，便可以如何使用评判器识别正确答案，以及即使评判器并不完美，集成裁判如何为幻觉选择概率提供明确的界限。第 4节介绍用于验证这两种效应的受控实验，第5 节作出结论。

2 正确答案的存在性

我们的框架核心在于一个简单却有力的建模假设：在新的上下文窗口中进行的独立大语言模型调用，表现得如同从固定输出分布中独立抽取的样本。该假设是本文所有存在性保证的数学基础。它反映了实际情况：每次模型调用都有自身的采样随机性，而从空白、未受污染的上下文开始，则能消除不同运行之间的干扰。在这一独立性假设下，对同一提示词进行重复执行，就相当于针对固定任务进行统计独立的多次“尝试”。

现在，我们将形式化说明本分析所适用的设定。在此阶段，我们仅关注在对模型的重复调用中，正确答案出现的概率。 识别步骤（“找到正确答案”）将在第 3 节单独展开，届时我们会纳入评判器，并在需要时采用多数投票聚合。

“特定任务 ”是指输入固定、每次运行仅发出一次提示词，且可以通过确定性规则评估正确性的任务。在这种情况下，每次调用大语言模型都相当于从一个固定分布中随机抽取一次。因此，独立地多次运行同一任务会产生一系列独立同分布随机变量。我们只需利用这一结构，就能推导出明确的概率保证，说明在重复执行中至少出现一个正确答案的可能性。

2.1 特定任务

令 Z 表示大语言模型可能的输入和输出构成的集合。一个特定任务包括:

- 一个固定输入 $A \in Z$,
- 一个固定的提示词函数 $B : Z \to Z$，用于生成向大语言模型发出的查询,
- 一个确定性的正确性函数 $C : Z \to Z$,
- 和正确性标准 $c_0 \in Z$。

一次任务执行包括在一个新的上下文窗口 中向大语言模型发送提示词 $B(A)$，并获取输出 $X \in Z$。如果 $C(X) = c_0$，我们就称模型回答正确；否则，我们称该回答为幻觉 。

关键的是，在新的上下文窗口中发起每次调用，会使输出 X_1， X_2，.....表现为独立同分布随机变量。这是我们对模型所作的唯一结构性假设，虽简单却十分有力。由于输入、提示词和正确性标准均固定，每次运行的成功概率

$$p = \mathbb{P}(C(X_n) = c_0),$$

是任务的一项稳定属性。如果 $p > 0$，则称该任务可解 。

2.2 重复引理

我们关注 N 次独立模型调用均未能给出正确答案的概率。由于每次调用以 p的概率成功，以 ($1 - p$) 的概率失败，根据独立性可得

$$\mathbb{P}(C(X_1) \neq c_0, \ldots, C(X_N) \neq c_0) = (1 - p)^N.$$

由此得到以下简单而有力的观察。

定理 2.1 (重复引理)。 对于每次执行的正确性概率为 $p > 0$ 的任何可解决特定任务， N 次独立执行全部产生幻觉式响应的概率为

$$(1 - p)^N.$$

该概率会随着 N 的增加而呈指数级下降。

证明。 由于事件 $\{C(X_n) \neq c_0\}$ 相互独立，且每个事件的概率为 $(1-p)$，因此联合失败概率恰好为 $(1-p)^N$。 □

直观而言，如果模型偶尔会给出正确答案，那么反复执行同一任务，就会使某个回答中出现正确答案的可能性极高。该引理给出了明确的定量保证：只需少量重复，就能使“所有答案都是幻觉”这一事件极不可能发生。因此，重复能够保证重复输出中存在正确答案。接下来将讨论如何识别这样的答案。

3 正确答案的识别

重复能够确保重复输出中很可能出现正确答案。然而，单靠重复并不能确定这些回答哪一个才是正确的。从重复运行结果中均匀随机地选取一个回答，其正确概率与单次调用模型相同。为了从重复输出中识别出正确答案，我们使用大语言模型评审器，并在必要时通过多数投票增强其判断。本节将介绍评判器，推导流程未能识别出任何正确答案的概率，并给出所选答案为幻觉的概率上界。

3.1 评判模型

对于每个模型输出 X_n，我们都会向评判器单独发起一次是非查询，询问 X_n 是否符合固定的判定标准 C，即 $C(X_n) = c_0$。每次查询都会在新的上下文窗口中发送，以确保评判器对不同 n 的评估彼此独立。我们将评判器的响应记为一个伯努利随机变量

$$Y_n \in \{0,1\}, \quad Y_n = 1 \text{ meaning “the judge asserts } X_n \text{ is correct.”}$$

为了描述评判器的特性，我们定义其真阳性率和假阳性率

$$q^{++} = \mathbb{P}(Y_n = 1 \mid C(X_n) = c_0), \quad q^{-+} = \mathbb{P}(Y_n = 1 \mid C(X_n) \neq c_0).$$

这些量取决于任务和评判器提示词，但在重复过程中保持不变。完美评判器满足 $q^{++} = 1$ 和 $q^{-+} = 0$。我们强调，评判器评估本身构成一项具有固定正确性标准的特定任务，因此也继承了与模型调用相同的独立性假设。

3.2 评判存在性的可靠性

当且仅当没有任何模型输出被评判为正确时，评判选择流程才会失败。对于任意一次模型调用，要得到正确且被评判为正确的输出，必须发生两个事件：(i) 模型给出正确答案，其概率为 p；(ii) 评判器接受该答案，其概率为 q^{++}。因此，每次试验生成正确且被评判为正确的答案的概率为 pq^{++}。

同样，一个错误答案仍会以概率 $(1-p)q^{-+}$ 被判定为正确，从而产生一个假阳性候选项。尽管这些假阳性候选项无法帮助流程识别出

正确答案，但它们会影响判定为正确的集合的行为，因此会出现在流程失败的完整推导中。

综合这些观察，可得以下保证。

定理 3.1（重复-评审流程的可靠性）。 对于每次试验正确的概率为 $p > 0$、评审器的评审率为 $q^{++} > 0$ 和 $q^{-+} \geq 0$ 的特定任务，流程失败的概率，即在 N 次重复中没有出现经评判为正确的答案的概率，为

$$\big(1 - (pq^{++} + (1-p)q^{-+})\big)^N.$$

证明。 对于任意固定的 n，输出被评判为正确（无论评判正确与否）的概率为 $pq^{++} + (1-p)q^{-+}$。由于各次重复相互独立，在全部 N 次试验中都没有出现经评判为正确的答案的概率即为所述表达式。特别是，没有出现既正确又被评判为正确的答案这一事件包含在该事件之中。

□

Let

$$S = \{n : Y_n = 1\}$$

表示单个评审器判定为正确的响应集合。定理3.1刻画了 $S = \varnothing$ 的概率，即流程报告失败的概率。从现在起，我们以自然事件 $S \neq \varnothing$ 为条件，这意味着至少有一个模型输出被流程判定为正确。

3.3 通过集成裁判进行识别

假设重复判断流水线成功（$S \neq \varnothing$），我们可以根据评判器质量，界定在 S 中找到正确答案的可能性。接下来，我们将研究评判器优于随机水平的情形（$q^{++} > \frac{1}{2}$， $q^{-+} < \frac{1}{2}$）。

对于简单评估任务，例如，当 c_0 是一个单词，且我们询问评判器“ X_n 是否恰好是c_0? ”，大语言模型评审器在实践中往往表现完美，也就是说 $q^{++} = 1$ 且 $q^{-+} = 0$。在这种情况下，流程失败率（定理 3.1）可简化为重复引理（定理 2.1）；当流程成功时（$S \neq \varnothing$），选择任何答案都能保证得到正确答案。

然而，如果 q^{-+} 不为零，幻觉答案可能会出现在评审集合 S 中，并有可能被选中。如果评判器优于随机水平（$q^{-+} < \frac{1}{2}$），我们可以使用评审器集成，并通过多数投票来降低选中幻觉答案的概率。

对于每个 X_n，令 $Y_{n1}, \ldots, Y_{nK}$ 为相互独立的伯努利变量，其

$$\mathbb{P}(Y_{nk} = 1 \mid C(X_n) = c_0) = q^{++}, \quad \mathbb{P}(Y_{nk} = 1 \mid C(X_n) \neq c_0) = q^{-+}.$$

因此，集成中的每个评判器都与原始基础判别器具有相同的质量。我们在下文推导保证时并不要求这一点，但它简化了实际实现，因为可以在所有评审任务中重复使用同一个大语言模型评审器。

如果多数票中严格过半数为正，则集成判别器接受 X_n，即:

$$\widetilde{S} = \left\{ n : \frac{1}{K} \sum_{k=1}^{K} Y_{nk} > \frac{1}{2} \right\} = \left\{ n : \overline{Y}_n > \frac{1}{2} \right\}.$$

我们将集成的真正率和假阳性率分别记为

$$Q^{++}(K) = \mathbb{P}\Big(\overline{Y}_n > \frac{1}{2} \mid C(X_n) = c_0\Big), \quad Q^{-+}(K) = \mathbb{P}\Big(\overline{Y}_n > \frac{1}{2} \mid C(X_n) \neq c_0\Big).$$

~在选择阶段，最终答案从 S 中均匀选出。因此，我们以~自然事件 $S\neq\varnothing$ 为条件；只要集成判别器至少接受一个候选项，该事件就成立。现在计算选出的答案为幻觉的概率。

定理 3.2（集成判别器下的**幻觉选择**。 设

- p 为每次试验的模型正确概率，
- q^{++} 且 q^{-+} 分别为基础判别器的真阳性率和假阳性率，并且
- $Q^{++}(K)$</style}和 (K)分别为由 K 次独立投票形成的集成判别器对应的真阳性率和假阳性率。

由于集成判别器仅用于基础判别器首先接受的输出，因此只有 $S = \{n : Y_n = 1\}$ 中的候选项会被送交集成判别器；于是，正确的模型答案被集成判别器接受的概率为

$$pq^{++}Q^{++}(K).$$

同样，错误答案被接受的概率为

$$(1-p)q^{-+}Q^{-+}(K).$$

~ ~在给定 $S\neq\varnothing$的条件下，从 S 中均匀选择的元素是幻觉的概率为

$$\mathbb{P}(C(X_n) \neq c_0 \mid \widetilde{S}\neq 0) = \frac{(1-p)\,q^{-+}\,Q^{-+}(K)}{p\,q^{++}\,Q^{++}(K) + (1-p)\,q^{-+}\,Q^{-+}(K)}.$$

~证明。 答案进入 S 当且仅当：(i) 模型生成该答案；(ii) 基础判别器将其判定为正确；(iii) 集成判别器中的多数票将其判定为正确。因此，正确答案以~的概率进入 S ，该概率为

$$p \times q^{++} \times Q^{++}(K),$$

而幻觉答案进入的概率为

$$(1-p) \times q^{-+} \times Q^{-+}(K).$$

~ ~在以 $S\neq\varnothing$ 为条件并从 S中均匀选择时，选择概率与这些进入率成正比，归一化后即可得到所述表达式。 □

3.4 集成判别器准确率

集成假阳性率可以写成封闭形式：

$$Q^{-+}(K) = \sum_{i=\lceil K/2\rceil}^{K} \binom{K}{i} (q^{-+})^i (1-q^{-+})^{K-i},$$

对于 $Q^{++}(K)$，也有类似的表达式。标准的 Hoeffding界表明，对于任意 $q^{-+} < \frac{1}{2}$,–

$$Q^{-+}(K) \leq e^{-2K\left(\frac{1}{2}-q^{-+}\right)^2},$$

随着评判器数量 K 增加而呈指数衰减。将这些速率代入定理3.2即可得到呈指数下降的幻觉选择概率。

3.5 算法化表述

以下流程概述了针对完美评判器的重复-评审流水线。

算法 1重复–评判器流程（完美评判器情形）

要求： 具有固定输入 A 的特定任务及正确性标准 C；重复次数 N
确保： 选定一个答案或 FAILURE

对 $n = 1 \ldots N$ **执行**
    $X_n \leftarrow$ 带任务提示的大语言模型查询
    $Y_n \leftarrow$ 在新的上下文窗口中进行评判器查询：“ X_n 是否完全正确？”
</style}结束循环
$S \leftarrow \{n : Y_n = 1\}$ ▷ 所有答案均由（完美的）评判器判定为正确
若 $S = \varnothing$ **则**
    返回 FAILURE
否则
    从 S 中均匀随机选择 n
    返回 X_n
结束条件判断

4 实验

我们开展了受控实验，以验证定理 3.1 （“所有答案产生幻觉的可能性有多大？”）和定理 3.2 （“集成判别器选中幻觉答案的可能性有多大？”）中建立的两个定量保证，所使用的是三项固定输入抽取任务（附录 A）。所有模型输出均使用`Qwen3-4B-Instruct-2507` （Yang 等，2025年）生成。我们在线提供了实现（Schippkus2026）。

每项任务都要求模型从简短提示词中返回一个特定词语，因此正确性是明确的。对于每项任务，我们根据 10,000 个独立样本，估计模型每次试验的正确概率 p。为控制评判器质量，我们使用合成评判器，以 0.25 的概率独立翻转分配的标签 Y_n ，由此得到基础判别器质量率 $q^{++} \approx 0.75$ 和$q^{-+} \approx 0.25$。对于集成裁判，我们为每个输出生成 $K \in \{1, 3, \ldots, 17\}$ 次独立的评判器评估，并采用多数投票。

我们完整执行流程 10，000 次，并记录 (i) 流程是否失败（事件$S = \varnothing$）以及 (ii) 是否从经集成评判器评判的集合中选出幻觉 S。在所有任务中，经验性流程失败率均随 N 下降，这与定理 3.1 的预测完全一致（图 2a– c）。同样，图 2d–f 展示了幻觉选择率，其随 K 下降，这与定理 3.2的预测完全一致。

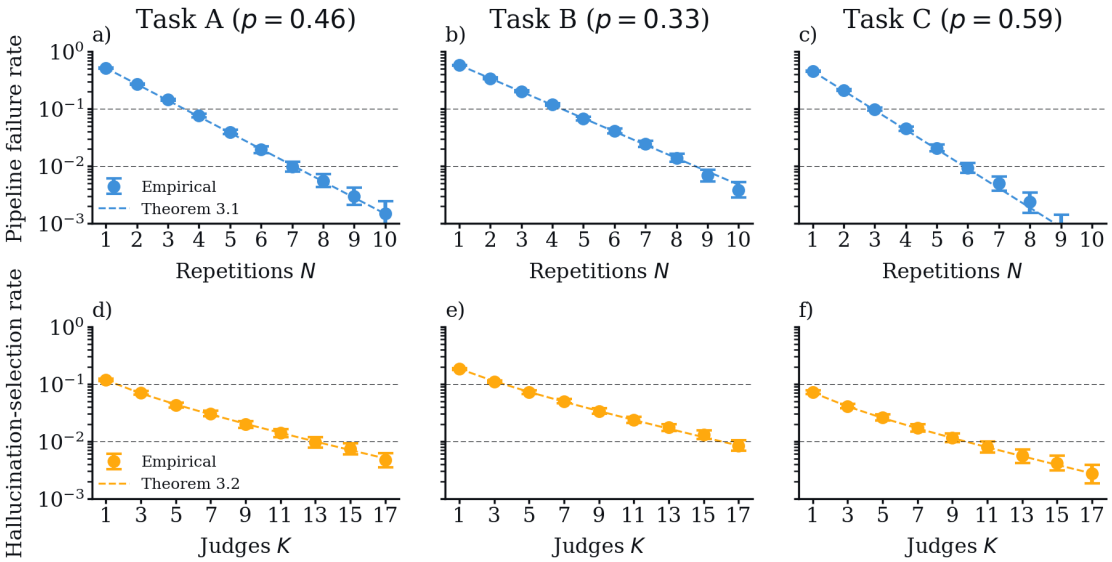

图 2：在 10,000 次流程运行中估算的重复-评审流水线和集成判别器的经验表现。我们展示了三个具有不同 p 的示例任务，并使用一个不完美的评判器（$q^{++} = 0.75$， $q^{-+} = 0.25$）。**a–c**）流程失败率（“所有答案都被判为错误”）随重复次数 N呈指数级下降。定理 3.1 预示了观测到的衰减，虚线表示该趋势。**d–f**）幻觉选择率（“集成判别器选择了一个幻觉答案”）随集成规模 K呈指数级下降。定理 3.2 预示了观测到的衰减，虚线表示该趋势。所有误差线均表示威尔逊置信区间。综合来看，这些结果表明，重复判断流水线为将上下文幻觉降低到预期比率奠定了基础，即使评判器并不完美也是如此。

5 结论

本文提出了一个概率框架，为固定输入大语言模型工作流提供明确的可靠性保证。我们聚焦于特定任务，其中提示词和正确性标准是固定的，并表明可靠性自然分为两个相互独立的组成部分：确保出现正确答案，以及可靠地识别它。我们证明，在上下文窗口相互独立的假设下，这两个问题都可以解决。

对于存在性问题，在新的上下文窗口中对同一提示词进行独立重复，可以使所有模型输出都不正确的概率以指数速率下降（定理 2.1）。对于识别问题，我们引入大语言模型评审器。经评判的流程的可靠性仅取决于评判器的真阳性率和假阳性率（定理 3.1）。当评判器并不完美时，对独立评审器调用进行多数票表决，可使集成级错误率同样呈指数级下降（定理 3.2）。这两个组成部分——重复和评判式选择——提供了明确的端到端保证，且有两个相互独立、可调节的参数：模型重复次数 N 和评审器调用次数 K。

在采用合成噪声评审器的受控抽取任务上进行的实验表明，实测行为与理论预测高度吻合：流程失败随N 呈指数级下降，而幻觉选择随 K呈指数级下降。这证明，无需更改模型权重、提示词或解码策略，也能构建高度可靠的固定输入流程。至关重要的是，这也证明了我们的独立性假设是恰当的，至少对于本文研究的简单任务而言如此。我们鼓励在

要求较不严格。

未来的研究方向包括分析评判器在不同领域中的鲁棒性，优化 N 与 K 之间的计算权衡，以及将该框架扩展到具有分级正确性或更长多步骤工作流的任务。尽管如此，研究结果表明，任何成功概率非零的特定任务，原则上都可以通过结合独立重复与评判式选择，以任意高的可靠性加以解决，从而为可靠的智能体大语言模型系统提供一个简单且具有理论依据的基础

致谢

本研究部分由德国联邦教育与研究部（BMBF）以及汉堡自由汉萨市根据联邦政府和各州的卓越战略资助。

作者贡献（CRediT）

概念化：NR、SS；形式分析：NR、SS；经费获取：SS；调查：SS；方法论：NR、SS；资源：SS；软件：SS；可视化：SS；撰写——初稿：NR、SS；撰写——审阅与编辑：NR、SS。

参考文献

Adlakha, V., P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy（2024）。评估指令遵循模型在问答中的正确性与忠实性。 DOI: 10 .48550 / arXiv.2307.16877。Chang, C.-C., D. Reitter, R. Aksitov, and Y.-H. Sung（2023）。KL散度引导的温度采样。 DOI: 10.48550/arXiv.2306.01286。Chang, Y., K. Lo, T. Goyal, and M. Iyyer（2024）。BooookScore：大语言模型时代的书籍长度摘要系统性探索。 DOI: 10.48550/arXiv.2310.00785。Cohen, R., M. Hamri, M. Geva, and A. Globerson（2023）。LM vs LM：通过交叉质询检测事实性错误。 DOI: 10.48550/arXiv.2305.13281。Du, Y., S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch（2023）。通过多智能体辩论提升语言模型的事实性与推理能力。 DOI: 10.48550/arXiv.2305.14325。Fabbri, A. R., W. Kry´sci´nski, B. McCann, C. Xiong, R. Socher, and D. Radev（2021）。“SummEval：重新评估摘要评估”。载于：《计算语言学协会汇刊》9，页391–409。 ISSN: 2307-387X。 DOI: 10.1162/tacl_a_00373。Hu, X., D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Y. Zhang, and Z. Zhang（2024）。RefChecker：用于大型语言模型的基于参考资料的细粒度幻觉检测器与基准测试。 DOI: 10.48550/arXiv.2405.14486。Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, and B. Qin（2025）。“大型语言模型中的幻觉综述：原理、分类、挑战与开放问题”。载于：ACM信息系统汇刊43.2，页1–55。 ISSN: 1558-2868。 DOI: 10.1145/3703155。

Jain, S., V. Keshava, S. M. Sathyendra, P. Fernandes, P. Liu, G. Neubig, and C. Zhou（2023）。“利用上下文学习对文本摘要进行多维评估”。载于：计算语言学协会论文集：ACL 2023，页8487–8495。 DOI: 10.18653/ v1/2023.findings-acl.537。Kalai, A. T., O. Nachum, S. S. Vempala, and E. Zhang（2025）。语言模型为何会产生幻觉。DOI: 10.48550/arXiv.2509.04664。Kim, Y., Y. Chang, M. Karpinska, A. Garimella, V. Manjunatha, K. Lo, T. Goyal, and M. Iyyer（2024）。FABLES：评估书籍长度摘要的忠实性与内容选择。 DOI: 10.48550/arXiv.2404.01261。Kryscinski, W., B. McCann, C. Xiong, and R. Socher（2020）。“评估抽象式文本摘要的事实一致性”。载于：2020年自然语言处理实证方法会议论文集（EMNLP）。编者：B. Webber, T. Cohn, Y. He, and Y. Liu。线上：计算语言学协会，页9332–9346。 DOI: 10.18653/v1/ 2020.emnlp-main.750。Laban, P., W. Kry´sci´nski, D. Agarwal, A. R. Fabbri, C. Xiong, S. Joty, and C.-S. Wu（2023）。大语言模型作为事实推理器：来自现有基准测试及其他方面的见解。 DOI: 10.48550/arXiv. 2305.14540。Liu, S., K. Halder, Z. Qi, W. Xiao, N. Pappas, P. M. Htut, N. A. John, Y. Benajiba, and D. Roth（2025）。迈向长上下文幻觉检测。 DOI: 10.48550/arXiv.2504.19457。Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu（2023）。G-Eval：使用GPT-4进行自然语言生成评估，提升与人类的一致性。 DOI: 10.48550/arXiv.2303.16634。Longpre, S., K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh（2021）。“问答中的基于实体的知识冲突”。载于：2021年自然语言处理实证方法会议论文集。编者：M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih。线上及多米尼加共和国蓬塔卡纳：计算语言学协会，页7052–7063。 DOI: 10.18653/v1/2021.emnlp-main.565。Luo, Z., Q. Xie, and S. Ananiadou（2023）。ChatGPT作为文本摘要的事实不一致性评估器。 DOI: 10.48550/arXiv.2303.15621。Madaan, A., N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prab-humoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark（2023）。Self-Refine：通过自我反馈进行迭代精炼。 DOI: 10.48550/arXiv. 2303.17651。Maynez, J., S. Narayan, B. Bohnet, and R. McDonald（2020）。论抽象式摘要中的忠实性与事实性。 DOI: 10.48550/arXiv.2005.00661。Rajendhran, R., A. Zadeh, M. Sarte, C. Li, and M. Iyyer（2025）。VeriFastScore：加速长篇内容的事实性评估。 DOI: 10.48550/arXiv.2505.16973。Schippkus, S.（2026年1月）。schipp/hallucination guarantees：预印本。版本：预印本。 DOI: 10.5281/zenodo.18128071。 URL: https://doi.org/10.5281/zenodo.18128071。Shi, W., X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, and S. W.-t. Yih（2023）。信任你的证据：通过上下文感知解码减少幻觉。 DOI: 10.48550/arXiv.2305.14739。Shinn, N., F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao（2023）。Reflexion：结合语言强化学习的语言智能体。 DOI: 10.48550/arXiv.2303.11366。Wan, D., M. Liu, K. McKeown, M. Dreyer, and M. Bansal（2023）。面向抽象式摘要的忠实性意识解码策略。 DOI: 10.48550/arXiv.2303.03278。

-

Wang,Z., S.Mao,W. Wu, T.Ge, F.Wei, and H. Ji (2024). 释放大型语言模型中涌现的认知协同效应：通过多角色自我协作实现的任务求解智能体. DOI: 10.48550/arXiv.2307.05300.Yang, A., A. Li, B.Yang,B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng,D. Liu, F. Zhou, F. Huang,F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang,J. Yang,J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li,M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R.Gao, S. Liu, S. Luo, T. Li, T. Tang,W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y.Su, Y. Zhang, Y.Zhang, Y. Wan,Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu(2025). *Qwen3*技术报告. DOI: 10.48550/arXiv.2505.09388.Zha, Y., Y.Yang, R. Li, and Z. Hu (2023). *AlignScore*：使用统一对齐函数评估事实一致性。 DOI: 10.48550/arXiv.2305.16739.Zhang, Y., Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X.Huang, E. Zhao, Y. Zhang, C. Xu, Y. Chen,L. Wang, A. T. Luu, W. Bi, F.Shi, and S. Shi (2025). 人工智能海洋中的塞壬之歌：大型语言模型幻觉研究综述。 DOI: 10.48550/arXiv.2309.01219。

A 任务提示

	{任务}	**{目标}**	c_0	$\boldsymbol{p}$
A	句子一：她穿着一件鲜红的条纹衬衫去参加派对。 句子二：它非常时尚。	第六	有条纹的	0.46
B	句子一：有关新项目的信息已发送给 所有员工。 句子二：请查看您的电子邮件。	第七	sent	0.33
C	句子一：那三枚小巧的银色圆形硬币被发现了 在泥土里。 句子二：它们看起来像古罗马钱币。	第八	发现	0.59

表1：三个示例任务，其中 **{任务}** 和 **{目标}** 被插入提示模板中。对于每项任务，返回真实值 c_0 的概率为 $\boldsymbol{p}$，该概率根据 10,000 次模型运行估算得出。

我们使用以下提示模板，并将 {任务} 和 {目标} 按表1所列内容插入其中。由此生成的提示词对应于我们通过实证研究的三个特定任务A、B 和 C（图2）。提示模板：

> “下面有两个句子。
>
> 你的任务：从*第一个*句子中找出 {目标} 词，并且只回复那个词。
>
> 注意：第一个句子有很多词，其中一些是干扰项。关键词在句子中间附近，但可能会被干扰项或复杂的措辞所掩盖。如果答案很明确，就给出目标词。如果你不确定或觉得太难理解，就说“我不知道”。
>
> 句子如下：
>
> {任务}
> 只回复第一个句子的 {目标} 词；如果你不确定，就回复“我不知道”。”