用于检测大语言模型中的准确性幻觉和不一致性的提示-响应语义发散度量

Igor Halperin*
Fidelity Investments

August 15, 2025

Abstract

大型语言模型（LLMs）的普及面临着幻觉的挑战，这是一种关键的失败模式，其中模型生成了不符合事实、无意义或不忠实的文本。本文介绍了语义分歧指标（SDM），这是一种新颖的轻量级框架，用于检测忠实性幻觉——即 LLMs 响应与输入上下文严重偏离的事件。我们专注于这些 LLM 错误的具体实现，即虚构，定义为与用户查询任意且语义不符的响应。现有的方法如语义熵通过测量对单一、固定提示的回答多样性来测试任意性。我们的 SDM 框架在这一点上有所改进，因为它更加注重提示：我们通过不仅测量多个回答之间的一致性，还测量原始提示的多个语义等价改写之间的回答一致性，从而测试更深层次的任意性。在方法上，我们的方法使用句子嵌入的联合聚类来为提示和回答创建一个共享的主题空间。提示和回答之间主题共现的热图可以被视为用户-机器对话的二维量化可视化。然后，我们计算一组信息论指标来衡量提示和回答之间的语义差异。我们的实际得分 $\mathcal{S}_H$ 结合了詹森-香农散度和瓦瑟斯坦距离来量化这种差异，高分表明一种忠实性的幻觉。此外，我们识别出 KL 散度 KL（回答 || 提示）作为语义探索的强大指示符，这是区分不同生成行为的关键信号。这些指标进一步结合成 **语义框**，一个用于分类 LLM 回答类型的诊断框架，包括危险的、自信的虚构。我们的工作为实时检测黑盒 LLM 中的忠实性幻觉和语义不一致，提供了一种原则性、提示敏感的方法，无论是单次查询还是多轮对话。

1 引言

像 Gemini、Claude 和 GPT-4 这样的大型语言模型（LLMs）在生成类人文本方面展示了卓越的能力。然而，它们的可靠性受到被称为 LLM 幻想 [15] 的关键失效模式的影响。对这些错误的全面分类法将其分为内在幻想（与提供的输入或上下文相矛盾）和外在幻想（无法根据来源验证的声明）。另一个关键的分类轴是区分真实性（与真实世界知识的一致性）和忠实性（与用户提示和提供内容的一致性）[4] 。

在本文中，我们主要关注检测内在的真实幻觉。我们将其视为测量大型语言模型（LLM）的响应与用户提供的完整输入上下文之间语义差异的问题。

一些研究人员提出，与用“幻觉”来描述这些 LLM 失败相比，使用“虚构”一词更为合适 [6] 。在精神病学中，“幻觉”通常指与感官有关的现象，而“虚构”则是指一个人相信为真的错误或扭曲的记忆。这种情况发生在个体用捏造的细节填补其记忆中的空白时，并不是出于有意的欺骗。这个精神病学的定义使得“虚构”成为一种有意义的同义词，用于描述我们想要检测的 LLM 响应的那种上下文不一致性或不真实性。

对于大型语言模型（LLM）应用，定义虚构的一个方法是说它指的是 LLM 输出的内容既不正确又随意，而其随意性是由于它对 LLM 输出生成中的统计噪声过于敏感。语义熵（SE）方法 [6] 基于这一思想，并将 LLM 虚构定义为 LLM 输出对像温度采样所用随机种子等无关细节的敏感性。SE 方法相当于在每次 LLM 调用中保持相同的初始提示的同时，用不同的随机种子进行 LLM 响应的多次采样。通过进行语义聚类并测量所有 LLM 响应所产生的聚类分布的熵来估计产生回答的变异性。这最终导致他们的最终得分，即语义熵（SE）。

尽管强大，SE 方法有一个关键限制：它不是（充分）提示感知的。关于提示的信息没有被充分利用，而背景信息只是用来锚定每个 LLM 的响应，而不是以其他方式使用。提示仍然是一个固定的上下文背

*The author acknowledges the assistance of Gemini 2.5 Pro, Claude 4 Opus and Grok 3 in the preparation of this manuscript and its associated code. Some of content hallucinations created by AI assistants were caught and removed by the author. All remaining errors are the author's own. The views expressed herein are those of the author and do not necessarily reflect the views of his employer. Email for correspondence: ighalp@gmail.com.

景。[1] SE 方法仅测试对于固定提示重复的稳定性，而不涉及其上下文，并可能错误地将提示实际暗示的合法复杂回答标记为虚构。

在本文中，我们对 LLM 虚构的解释更为宽泛，使其更类似于内在忠实度幻觉。[2] 我们将虚构定义为 LLM 回答相对于提供的输入上下文的错误和任意性。在我们的框架中，虚构指的是一种在语义上完全或部分与用户提示无关的回答——无论该提示是一个简单查询，还是一个加上广泛参考文档的查询。

我们的核心论点是，一个忠实的、非虚构的响应应在语义上与其提示所定义的世界保持一致。因此，可以将虚构理解为 **语义不对齐** 的严重形式。我们提出了一种新颖的轻量级框架，语义发散度量（SDM），以更稳健的、提示感知的方式直接量化这种一致性。该框架旨在用于分析动态的提示-回答对，并适用于与大型语言模型（LLM）的单次交互和多轮对话。

SDM 方法论在黑箱环境下操作，涉及一个由两个步骤组成的程序。首先，它通过对一组语义等价的改写提示生成多个响应来更深入地探测任意性。其次，我们的方法使用联合嵌入聚类来建立一个共享的主题空间，用于提示和答案。然后，我们计算一系列信息论和分布度量。我们的第一个最终分数 $\mathcal{S}_H$ 结合了 Jensen-Shannon 散度（JSD）和 Wasserstein 距离，以提供一个稳健的、单值的语义漂移衡量，其中高分表明更高的杜撰风险。此外，我们识别出 KL 散度 KL(Answer || Prompt) 作为语义探索的一个强大度量，是区分不同生成行为的关键信号。该度量构成了我们 SDM 方法的第二个最终分数。

通过一系列的控制实验，我们展示了我们的框架能够成功区分不同类型的响应。我们使用可视化热图分析（第 6.1.3 节）来揭示不同对齐策略的独特视觉特征。结合我们的指标，我们引入了语义框（第 7 节），这个新颖的诊断框架使用我们的指标来分类响应行为——从忠实对齐到最危险的、信心十足的错误对齐虚假陈述。这个研究提出了一种原则性、定量的方法来衡量大型语言模型可靠性的重要方面：他们生成与用户查询在语义上对齐的响应的能力。

我们的贡献有三点：

1. 我们引入了一种新颖的、对提示敏感的方法来检测虚构内容，该方法通过使用改写的提示来更稳健地测试随意性。

2. 我们提出了一种简单实用的算法，用于测量语义对齐，基于提示-答案嵌入的联合聚类以及由此产生的散度指标（集成 JSD、KL 等）。

3. 我们开发了一个综合诊断框架，称为语义盒，该框架使用这些指标来分类大型语言模型行为的不同模式，包括自信的虚构，这是一个关键的失败模式。

本文的其余部分结构如下。第二节回顾了与幻觉检测相关的工作。第三节详细介绍了基础的语义熵方法。第四节提出了我们建议的 SDM 方法论。第五节分析了其计算复杂性。第六节呈现了我们的实验结果。第七节介绍了我们的可视化工具，语义框。第八节讨论了我们的结果。第九节概述了未来的工作和研究方向，第十节总结了全文。

2 相关工作

LLM 中的幻觉检测是一个快速发展的领域。研究可以根据验证所使用的信息来源进行广泛分类。内在方法检查生成文本内部的矛盾，而外在方法则将声明与外部知识库 [7] 进行验证。我们的工作重点是在黑箱环境中检测一种特定类型的忠实幻觉——即虚构，其中只有模型的输入和输出是可用的。

在黑盒设置中，主要的方法是测量响应一致性。像 SelfCheckGPT [12] 这样的方法对单个提示语进行多次响应采样，并检查一致性。语义熵 [6] 的概念通过测量答案集的语义多样性来完善这一点，将高熵视为任意性的信号，这是虚构的关键组成部分。然而，这些方法对提示是无感知的；它们只测试关于一个单一固定措辞的提问的不稳定性。

一种更复杂的方法涉及测试模型在输入提示变换下的鲁棒性。变形测试的框架已经成功应用于这个问题。在他们关于 MetaQA 的工作中，Yang 等人 [14] 引入了一组变形关系——应用于提示的变换在逻辑上应该导致回答的预测性变换。例如，理想情况下，重新措辞提示（“保持意义”的变换）应该产生语义上等效的回答。通过将原始回答与变换后提示的回答进行比较，他们能够检测出可能导致虚假信息的潜在不一致。

我们的 SDM 框架分享了使用提示转换，特别是改写，来探测不稳定性的核心见解。然而，我们的方法在两个基本方面有所不同。首先，MetaQA 通常将一个原始答案与一个转换后的答案进行比较，而我们的方法使用基于集合的分析。我们为每个改写的提示生成多个响应，这使我们能够衡量整个响应分布

[1] 因为 SE 方法在没有深度依赖于提示内容的情况下衡量响应多样性，所以它主要作为一种检测外部幻觉的工具。
[2] 因此，我们将“虚构”与“内在忠实性幻觉”这两个术语交替使用。

的统计属性，而不仅仅是一单一实例。其次，我们的提示-响应对比方法不同。我们不依赖于将 LLM 作为裁决者的方法，而是使用联合嵌入聚类来创建一个共享的句子级别主题空间。这使我们能够应用更丰富的信息理论和分布度量（JSD、KL 散度、Wasserstein 距离），以提供更细粒度和多方面的语义分歧测度以及虚构风险评估。

其他研究也使用嵌入来衡量提示和响应的对齐度。最近的一种方法，SINdex [1] ，建议将输入提示与生成的响应连接成一个字符串，然后为这个组合的文本计算单个嵌入。通过比较多个此类“[提示 + 响应]”字符串的嵌入，来衡量不一致。这是在文档级别操作，将整个文本视为单一的语义单元。相比之下，我们的 SDM 方法进行更为细致的句子级分析。通过对提示和回答的单个句子进行联合聚类，我们构建了一个共享的主题空间，使得能够进行更精细的主题偏离和内容漂移的测量，这通过我们的 JSD 和 Wasserstein 指标捕捉。参考 [2] 研究了提示复杂性如何影响模型行为，但没有直接涉及幻觉检测。

3 背景：语义熵方法

由 Farquhar 等人提出的语义熵（SE）方法为我们的工作提供了概念基础。该方法旨在通过测量由大语言模型（LLM）生成的一组回答中“随意性”或语义不一致性来检测虚构。其方法的关键步骤是：

1. 生成多个答案：对于给定的提示，使用基于温度的采样方法生成 N 个不同的答案。
2. 句子分割：每个答案被分解为其组成句子。
3. 基于 NLI 的聚类：SE 方法的核心是其聚类过程。它根据语义等价性对句子进行聚类。使用自然语言推理（NLI）模型来判断句子对是否双向蕴涵彼此。如果句子被 NLI 模型确定为语义等价，则将其归为同一类。每个生成的聚类代表一个独特的“主题”（或“语义概念”）。
4. 熵计算：聚类分配形式为离散概率分布，其中 p_i 表示属于聚类 i 的句子比例。然后，香农熵的计算公式为 $H = -\sum_{i=1}^{k} p_i \log_2 p_i$ 。
5. 幻觉检测：SE 分数高表明模型在针对同一提示生成许多不一致或矛盾的想法，这表明更高的虚构可能性。

尽管这是一种测量响应多样性的方法，其主要缺陷在于它在很大程度上是与提示无关的。它不能区分虚构响应的高熵和对复杂提示的正确、多方面答案的合法高熵。我们的工作建立在测量响应多样性这个理念之上，但引入了对提示的认知，并使用另一种技术方法（联合嵌入聚类）来实现更加细致入微、上下文感知的分析。

4 方法论：语义分歧度量 (SDM)

我们提出的 SDM 框架旨在检测特定的、具有挑战性的幻觉类型：虚构，这被定义为一种 LLM 响应，其相对于提供的输入上下文而言，所做的声明既不正确又随意。这个输入上下文，我们通常称之为“提示”，定义了忠实响应的完整真实语义空间。我们框架的核心原则是，良好对齐的、非虚构的响应必须在语义上与这个提示定义的世界保持一致。本节详细介绍了我们的方法，从激励我们方法的信息理论基础开始，最后设计了用于评估在给定提示-响应对中虚构可能性的实用信号。

4.1 信息论基础：虚构的诊断框架

我们的 SDM 框架的基础是这样一个原则，即提供给 LLM 的完整输入——即提示——定义了一个对齐良好且非虚构的响应的真实语义空间。因此，虚构是对这一特定提示定义的语义空间的偏离。我们的框架设计为通用的，将“提示”视为提供给模型的上下文的全部内容。这个提示上下文可以采用两种主要形式：

- 一个直接查询，其中提示本身就是完整的上下文（例如，“比较和对比凯恩斯和哈耶克的经济政策。”）。
- 一个扩展了参考文档的查询，其中提示包括模型预计要使用的外部信息（例如，在检索增强生成（RAG）中）。

我们建议，不如依赖单一指标（例如语义熵方法中所做的那样），可以通过类似于机器学习中偏差和方差概念的两个正交维度来诊断虚构。

1. 语义探索（“偏倚”维度）：第一个维度衡量答案分布偏离提示分布的语义漂移或“偏倚”。它量化了大模型（LLM）必须在提示中明确提供的概念之外进行详细说明、创作或推断的程度。我们使用 Kullback-Leibler（KL）散度 KL（Answer || Prompt）来测量这一点。该语义探索指标的高值表明模型进行了显著的创造性或解释性跳跃来生成其响应。

2. 语义不稳定性（“方差”维度）：第二个维度测量跨一系列语义等价的提示语与回应之间的语义距离。它量化了模型回应策略的稳定性，即与提示语在语义上的一致性。我们使用一个实际得分 $\mathcal{S}_H$ 来捕获这一点，该得分由集合 JSD 与 Wasserstein 距离的加权平均值给出，这些距离是从生成句子的整个分布上计算的，使得得分对回答云的分布和一致性具有高度敏感性。答案的可变性也通过对提示语和回应句的聚合成簇用于构建语义基础中被考虑进去。

通过设定阈值 $KL_\star$ 和 $\mathcal{S}_\star$ ，我们可以基于这两个维度将任何给定的响应分类到四种状态中的一种。这提供了两个二元输出：

- || 在回答提示中的 KL 值大于还是小于 $KL_\star$ ？这可以区分低探索和高探索的响应。
- $\mathcal{S}_H$ 得分是大于还是小于 $\mathcal{S}_\star$ ？这可以区分低变异性（稳定/收敛）和高变异性（不稳定/发散）响应。

这一二维诊断是我们框架的核心理论原则。它超越了单一的通过/失败评分，为我们的语义框提供了信息论的基础，这是一种用于分类 LLM 的细致行为的工具——从忠实召回到自信虚构。

上述理论视角推动了一个用于幻觉检测的复杂度归一化指标。响应的熵 $H(P_r)$ 自然而然地随着提示的复杂性增加。后者可以通过提示熵 $H(P_q)$ 来近似。一个纯粹高熵的响应不一定是幻觉。然而，响应熵中无法通过其与提示的互信息 $I(P_q; P_r)$ 解释的部分可能是一个强信号。这启发了以下理想化的幻觉指标：

$$\phi(q, r) = \frac{H(P_r) - I(P_q; P_r)}{H(P_q)} = \frac{H(P_r|R_q)}{H(P_q)} \tag{1}$$

此比例，随后将被称为标准化条件熵（NCE），由给定提示的响应条件熵除以提示自身的熵得出，并捕获了响应复杂性中未被提示正当化的部分。此指标的高值可能表明响应相对于提示中主题的潜在大偏差，进一步表明此类响应中幻觉的潜力。

虽然方程（1）提供了一个强有力的理论蓝图，但它是僵化的。为了得到一个更灵活且经验上更稳健的度量，下面我们构建一个更灵活的实用幻觉评分 $\mathcal{S}_H$ ，其组成部分是从相同的信息理论度量中获得的，并且还包括一个分布距离组件。我们将在后面详细介绍我们构建的细节以及聚合度量 S_H 和“语义探索”度量 KL[Answer || Prompt] 的确切定义。本节的其余部分详细介绍了计算此实用评分组件的过程。

为了计算 $\mathcal{S}_H$ 评分的组成部分，我们遵循一个四阶段的过程：（1）数据生成和嵌入，（2）联合语义聚类以识别共享主题，（3）计算关键的信息论和分布距离度量，以及（4）聚合成最终评分。算法 1 中展示了一个高层次的概览。

4.2 数据生成和联合嵌入

为了稳健地捕捉输入和输出的语义空间，我们首先生成一组多样化的提示-答案对。对于每个初始提示，我们生成 M 个释义。对于每个这些 M 个提示，我们然后使用基于温度的采样从目标 LLM 生成 N 个答案。

下一步是对所有生成的文本进行句子层面的处理。来自提示和答案的所有句子都会被分割，然后分别嵌入到一个共同的高维向量空间中。在本文的所有实验中，我们使用 `Qwen3-Embedding-0.6B` 模型，一个预训练的句子转换器，来执行这种嵌入。这一过程产生了两组句子嵌入：$\mathcal{P} = \{p_1, \ldots, p_{S_P}\}$ 用于提示句子，$\mathcal{A} = \{a_1, \ldots, a_{S_A}\}$ 用于答案句子，随后将它们结合进行联合聚类分析。

4.3 联合语义聚类和主题估计

我们方法的一个关键创新是使用联合聚类。我们不是单独分析提示和答案，而是将它们的句子嵌入合并到一个数据集 $\mathcal{S} = \mathcal{P} \cup \mathcal{A}$ 中。然后，我们对这个组合集应用基于 Ward 链接的层次凝聚聚类，以识别共享的语义主题。这种自下而上的方法构建了一个聚类层次结构，无需事先指定聚类的数量，这对于处理复杂性不同的提示非常理想。然而，在我们的实验中，我们并没有使用凝聚聚类算法所提供的这种灵活性，而是通过首先在我们的嵌入上运行 K-means 算法来为不同的聚类数量设定固定的聚类数量。然后我

们通过识别惯性图上的“肘部点”来估计最佳的聚类数量 k，这代表了在提示和答案中出现的最具有统计意义的不同语义主题的数量。以这种方式确定的聚类数量然后作为约束传递给层次聚类算法。[3] 这种联合聚类方法至关重要，因为它确保了语义相似的句子，无论是来自提示还是答案，都被分组到同一个主题聚类中，为比较它们的内容分布提供了直接和有意义的依据。

4.4 计算基于主题的对齐和分歧指标

在将所有句子分配到 k 个共享主题簇之后，最后一步是量化提示和响应分布之间的语义关系。我们的分析框架计算一个主要的发散得分作为最终的幻觉分数，并辅以若干补充指标进行全面诊断。

主要对齐指标：Jensen-Shannon 散度。 我们关于语义对齐的核心测量是 Jensen-Shannon 散度（JSD），它量化了由 $\boldsymbol{P}$ 表示的提示的主题分布与由 $\boldsymbol{A}$ 表示的答案的主题分布之间的相似性。JSD 是 Kullback-Leibler（KL）散度的对称且有界版本，使其成为比较概率分布的一个稳健指标。它通过一个混合分布 $\boldsymbol{M} = \frac{1}{2}(\boldsymbol{P} + \boldsymbol{A})$ 来定义，如下所示：

$$D_{\mathrm{JS}}(\boldsymbol{P} \| \boldsymbol{A}) = \frac{1}{2}\left(D_{\mathrm{KL}}(\boldsymbol{P} \| \boldsymbol{M}) + D_{\mathrm{KL}}(\boldsymbol{A} \| \boldsymbol{M})\right) \tag{2}$$

结果得分从 0（对于相同的分布）到 1（对于最大不同的分布，使用以 2 为底的对数）。对整体“主题混合”的这种直接比较提供了一个直观的语义漂移测量。在我们的框架中，我们使用两种不同的方法计算这个指标——全球（先聚合）方法和集合（后平均）方法——以捕捉模型行为的不同方面，详见下一小节。集合 JSD 作为我们最终 $\mathcal{S}_H$ 得分的关键组成部分。

诊断指标：平均互信息及其可视化。 为了诊断提示和答题主题之间的潜在依赖结构，我们从一个平均联合概率分布中计算互信息（MI）。这确保了我们的数值 MI 指标与我们的可视化直接对应。

设 X 为表示从提示中抽取的句子的主题索引的离散随机变量，设 Y 为从相应答案中抽取的句子的主题索引的离散随机变量。估计它们的联合概率分布的过程如下：

1. 对于每一个 M 提示-回答对 (P_m, A_m)，我们基于该单一对中的句子级共现构建一个局部 $k \times k$ 列联表 $\boldsymbol{C}_m$。
2. 我们对每个局部表进行归一化以创建一个局部联合概率分布，$P_m(X, Y) = \boldsymbol{C}_m / \sum_{i,j} \boldsymbol{C}_m[i, j]$。
3. 最终的代表性分布是这些 M 局部分布的逐元素平均值：

$$P_{\mathrm{avg}}(X, Y) = \frac{1}{M} \sum_{m=1}^{M} P_m(X, Y) \tag{3}$$

这个平均分布被可视化为热图，以展示平均依赖结构。然后使用标准公式直接从这个矩阵计算互信息，提供一个关于提示和答案主题平均统计依赖的单一、一致的测量。

嵌入空间中的分布变换 为了测量语义内容的整体移动，我们超越简单的主题分布，比较句子嵌入的整体几何结构。我们计算提示嵌入分布和答案嵌入分布之间的 1-Wasserstein 距离，也称为地球搬运工距离。

直观来说，Wasserstein 距离表示将一个分布转化为另一个分布所需的最小“代价”，类似于寻找将一堆土从一种配置移动到另一种配置的最有效方法。在我们的背景下：

- 这两个“土堆”是在高维向量空间中的提示和回答句子的嵌入云。
- 移动一个单位的“代价”是提示嵌入和答案嵌入之间的欧几里得距离。

不同于简单的重心比较，Wasserstein 距离对分布的形状、扩展和内部结构很敏感。我们在此强调，Wasserstein 距离的计算不依赖于我们对提示和回应的句子的联合聚类和聚类标签，因为它是在所有提示和所有答案的原始高维云上进行操作的。

[3] 我们也尝试了一种指定层次聚类中聚类数量的替代方法，即通过定义一个距离阈值，发现了类似的结果。

4.5 全局与基于集成的信息度量

在将句子分配到共享的话题簇后，我们计算了一系列指标，以提供提示-响应关系的全面视图。我们区分了两种基本的计算方法：全局（先聚合）和集成（后平均）。

全局（聚合优先）指标。 这些指标提供了提示句子整个语料库与答案句子整个语料库总体差异的宏观视角。它们首先通过汇集所有句子来创建两个全局主题分布进行计算：

- $P(X)$：所有来自 M 改写提示的句子的主题概率分布。第 i 个成分，$P(X=i)$，是被分配到聚类 C_i 的所有提示句子的比例。
- $P(Y)$：生成答案中所有句子的主题概率分布。第 j 个分量 $P(Y=j)$ 是分配给簇 C_j 的所有答案句子的比例。

从这两个全局分布中，我们计算出全局 Jensen-Shannon 散度（D_{JS}）以及全局 Kullback-Leibler 散度的两个方向（D_{KL}）。

集成（后期平均）指标。 这些度量通过测量 M 个单独释义实验的平均行为，提供了一个更细粒度的、微观层面的视图。该方法捕捉到在全局聚合中可能丢失的成对变异性。其过程如下：

1. 对于集合中每一个 M 提示-回答对 (P_m, A_m)，我们为其提示创建一个局部主题分布 $P_m(X)$，并为其回答创建一个局部分布 $P_m(Y)$。
2. 然后我们计算这单个配对的局部散度指标，例如，$D_{\mathrm{JS}}(P_m(X)||P_m(Y))$。
3. 最终的集成度量是这些 M 局部值的平均值。例如，集成 JSD 为：

$$D_{\mathrm{JS}}^{\mathrm{ens}} = \frac{1}{M}\sum_{m=1}^{M} D_{\mathrm{JS}}(P_m(X)||P_m(Y)) \tag{4}$$

我们以相同的方式计算集合 KL 散度。这种“后求平均”方法也用于通过恒等式 $I^{\mathrm{ens}}(X;Y) = H(Y) - \frac{1}{M}\sum_{m=1}^{M} H(Y_m|X_m)$ 来计算我们稳健的诊断工具，集合互信息，从而确保在所有细化的诊断措施中方法论的一致性，其详细说明将在下一小节中介绍。这些集合度量在我们的结果表中报告，提供了对模型在逐个实例上的行为更深入的理解。

4.6 集成互信息 (EMI)。

虽然可以使用列联表根据公式（3）来估算 MI，但我们还考虑通过身份 $I(X;Y) = H(Y) - H(Y|X)$ 以另一种方式计算 MI，其中条件熵是在 M 释义实验的集合上进行平均的。以这种方式计算的 MI 度量将被称为集成互信息 (EMI)。设计 EMI 度量是为了比单一的全局计算提供更稳健的统计依赖估计。

EMI 度量的实施涉及两个主要步骤：

1. 整体回答熵（$H(Y)$）：这个术语代表了响应主题的总不确定性。它是通过在所有回答句子的总主题分布 $P(Y)$ 上使用标准的香农熵公式计算出来的。
2. 平均条件熵（$H(Y|X)$）：该术语表示在已知提示主题后答案主题中剩余的不确定性。为计算这一点，我们对 M 实验的集合进行条件熵的平均：
   - 对于集合中的每一个 M 对 (P_m, A_m)，我们首先基于句子层面的共现，构建一个局部 $k \times k$ 列联表，$\boldsymbol{C}_m$。
   - 从这个局部表中，我们计算局部条件熵 $H(Y_m|X_m)$，该熵衡量特定实例的答案不确定性。
   - 最终条件熵是这些局部值的平均值，代表预期的剩余不确定性：

$$H(Y|X) = \frac{1}{M}\sum_{m=1}^{M} H(Y_m|X_m) \tag{5}$$

集合互信息（EMI）是这两个熵项之间的差：$I(X;Y) = H(Y) - H(Y|X)$。这个值量化了实验集合中不确定性的平均减少，并在我们的结果表格中报告（例如，表 1）。

4.7 SDM 指标：归一化条件熵、S_H 和 KL 分数

我们的框架为分析生成了三个关键指标（语义发散指标）：标准化条件熵 ϕ 、聚合 SDM 幻觉分数 $\mathcal{S}_H$ 和 KL 散度 KL(Answer || Prompt)。

归一化条件熵 $\phi(Y|X)$ 。 该理论分数量化了答案复杂度中未被其与提示的共享信息解释的部分，详见方程（1 ）：

$$\phi(Y|X) = \frac{H(Y|X)}{H(X)} = \frac{H(Y) - I(X;Y)}{H(X)} \tag{6}$$

一个位于 1.0 附近的值表明适当的响应复杂度，而一个明显大于 1.0 的值则表示由于未解释的复杂度可能存在错觉。

4.8 实用 SDM 幻觉评分 $\mathcal{S}_H$

虽然我们的框架生成了一组丰富的诊断指标，但为了对响应稳定性进行排序，单一实用的分数是理想的。我们构建了 $\mathcal{S}_H$ 分数，作为我们两个最稳健且信息丰富的组件的加权和：集成 Jensen-Shannon 散度和 Wasserstein 距离。

重要的是，这两个指标衡量的是提示与响应之间“距离”的截然不同的方面。它们在不同的数据表示上运行，因此是互补的：

- JSD 度量在我们联合聚类创建的低维离散主题空间上运行。它衡量主题的分类分布之间的差异，告诉我们答案的主题构成是否相对于提示发生了变化。
- 相比之下，Wasserstein 距离直接作用于高维连续嵌入空间。它测量句子嵌入的原始“云”之间的几何距离，完全独立于任何聚类。即使高层主题保持不变，它也告诉我们语义内容和意义的整体变化。

通过结合这两种测量方法，我们的框架同时捕捉到了高级别的主题漂移和低级别的语义内容变化。

我们评分的最后两个组成部分是：

1. 全局分布变化（W_d ）：所有提示和答案嵌入的完整分布之间的 1-瓦瑟斯坦距离。
2. 集合语义分歧（$D_{\text{JS}}^{\text{ens}}$ ）：集合 Jensen-Shannon 分歧是我们最敏感的话题对齐度量。

这些组件被组合成最终得分，定义为它们加权和，并以输入的熵进行缩放：

$$\mathcal{S}_H = \frac{w_{\text{wass}} \cdot W_d + w_{\text{jsd}} \cdot D_{\text{JS}}^{\text{ens}}}{H(P)} \tag{7}$$

。我们通过提示熵 $H(P)$ 对最终得分进行缩放，以部分减少我们的语义偏离度量对提示自身复杂性的依赖，从而更可靠地比较在不同类型提示中获得的结果。

权重的选择，w_{jsd} 和 w_{wass} ，并不简单，因为这两个度量在不同的内在尺度上运作。一个天真的加权会被平均幅度较大的度量任意主导。为了创造一个更加平衡且有意义的分数，我们根据经验校准了权重，目标是让集成 JSD 平均贡献约 60 % 到最终分数的值，而 Wasserstein 距离贡献剩余的 40 % 。

正如实验部分（第 6 节）所详述的，我们分别为两个不同的实验集执行了这个校准过程，这两个实验集代表了非常不同类型的生成任务。这个分析揭示了一个关键洞察：两个实验集产生的最优权重几乎相同。这种一致性表明在不同任务类型之间，这两个指标的相对比例是稳定的，这为使用一组统一的权重提供了正当理由。因此，在本文的所有实验中，我们使用最终校准得到的权重 $w_{\text{jsd}} = 0.7$ 和 $w_{\text{wass}} = 0.3$ 。这个得分结合了我们最为稳健的指标，以一种平衡且经验支持的方式呈现，作为我们评估响应稳定性的主要实用工具。

4.9 KL 分数

除了确认 Φ 和 S_H 指标的有用性外，我们下面描述的实验还确定了一个非常重要的补充指标，即 KL 散度 KL(答案 || 提示)。与双向对称指标 S_H 不同，KL 散度是不对称的，反映了从提示到 LLM 响应的语义信息流。这个指标对提示中的语义生成探索量以及在给定提示的语义丰富性时 LLM 响应中包含的内容非常敏感。我们的实验表明，这个指标应该与 Φ 和 S_H 指标一起计算和监控，以更好地整体评估给定

的提示-响应对的语义连贯性和稳定性。与我们对指标 S_H 的定义类似，我们还根据提示的熵 $H[P]$ 对最终指标进行缩放，因此我们的最终 KL 分数为

$$KL(Answer||Prompt) = \frac{D_{KL}(Answer||Prompt)}{H(Prompt)} \tag{8}$$

。

Algorithm 1 用于稳定性分析的语义分歧度量

Require: Original prompt Q , Target LLM $\mathcal{L}$, Sentence Encoder $\mathcal{E}$

Generate M paraphrases $\{Q_m\}$ of Q and N answers $\{A_{m,n}\}$ for each.
Segment all prompts and answers into sentences.
Embed all sentences using $\mathcal{E} \to \mathcal{P}$ (prompt embeddings), $\mathcal{A}$ (answer embeddings).
Combine embeddings: $\mathcal{S} \leftarrow \mathcal{P} \cup \mathcal{A}$.
Determine optimal cluster count k on $\mathcal{S}$ using the Elbow Method with K-means clustering
Perform hierarchical clustering on $\mathcal{S}$ with k clusters $\to \{C_1, \ldots, C_k\}$.
Get cluster labels for all prompt sentences ($L_{\mathcal{P}}$) and answer sentences ($L_{\mathcal{A}}$).
// — Compute Global (Aggregate-First) Metrics —
Derive global topic distributions $\boldsymbol{P}_{\text{global}}$ from $L_{\mathcal{P}}$ and $\boldsymbol{A}_{\text{global}}$ from $L_{\mathcal{A}}$.
Compute Global D_{JS} , Global D_{KL} from these distributions.
Compute the 1-Wasserstein distance W_d between the full embedding sets $\mathcal{P}$ and $\mathcal{A}$.
// — Compute Ensemble (Average-Later) Metrics —
for all each of the M prompt-answer pairs (P_m, A_m) **do**
    Derive local topic distributions $\boldsymbol{P}_m$ and $\boldsymbol{A}_m$ for the current pair.
    Calculate local divergences, such as $D_{\text{JS}}(\boldsymbol{P}_m \| \boldsymbol{A}_m)$ and $D_{\text{KL}}(\boldsymbol{A}_m \| \boldsymbol{P}_m)$.
end for
Compute final Ensemble metrics (e.g., $D_{\text{JS}}^{\text{ens}}$, $D_{\text{KL}}^{\text{ens}}$, the NCE score Φ , etc.) by averaging the local values.
// — Calculate Final Score —
Calculate the final Hallucination Score: $\mathcal{S}_H \leftarrow \left(w_{\text{jsd}} \cdot D_{\text{JS}}^{\text{ens}} + w_{\text{wass}} \cdot W_d\right) / H(P)$.
return Key metrics: $\mathcal{S}_H$, Φ , $D_{\text{KL}}(\boldsymbol{A} \| \boldsymbol{P})$

5 计算复杂性分析

我们的 SDM 框架的计算成本是实际部署的重要考虑因素。接下来，我们将分析其时间和空间复杂度。

5.1 时间复杂度

总体复杂度主要由三个阶段决定：嵌入、聚类和散度度量的计算。设 S 为句子总数（包括提示和答案），M 为改写提示的数量，k 为主题簇的数量，d 为嵌入维度。

- 句子嵌入：这一步是线性的，与总句子数成正比，复杂度为 $O(S \cdot d)$ 。
- 层次聚类：这是计算成本最高的步骤。标准的凝聚聚类在固定数量的簇的高效实现中，具有 $O(S^2 \log S)$ 或 $O(S^2)$ 的时间复杂度，对于大型 S 来说，它在过程中占主导地位。
- 度量计算：
- 全局散度度量（JSD，KL）效率极高，只需对 S 句子标签进行一次遍历以构建分布，然后进行一次 $O(k)$ 操作。
- 集成散度度量需要遍历 M 对。在循环内部，构建局部分布所需的时间与每对中的句子数量呈正比。总体复杂度大约为 $O(S)$ 。
- Wasserstein 距离是在提示和答案嵌入的完整集合上计算的（分别为 S_P 和 S_A 句子）。它的精确复杂度大约是 $O((S_P + S_A)^3 \log(S_P + S_A))$ ，但由于每个实验的句子数量是可控的，这通常是快速的。

因此，总时间复杂度主要由层次聚类步骤驱动，$O(S^2)$ 。这使得该框架在句子总数为数百的典型场景中是实用的。

5.2 空间复杂度

主要的内存需求来自于存储完整的句子嵌入集，其空间复杂度为 $O(S \cdot d)$ 。计算沃瑟斯坦距离的成本矩阵在计算过程中需要 $O(S_P \cdot S_A)$ 的空间。所有其他数据结构，例如主题分布和共现矩阵，都很小，只需要 $O(k^2)$ 或 $O(k)$ 的空间。这使得该方法在内存使用上高效且在主题数量方面具有可扩展性。

6 实验

为了验证我们框架的能力，我们使用 `gpt-4o` 模型进行了两组不同的实验。第一组实验专门设计用来测试框架在一个受控的“稳定性梯度”上的敏感度，从高度事实性的任务到纯粹创造性的任务。第二组实验使用了更为多样的提示类型，以识别模型行为的不同模式，包括“自信幻觉”状态。对于所有实验，我们生成了 10 个释义和每个释义 4 个响应。

6.1 实验组 A：控制稳定性梯度

6.1.1 提示描述

这个集合包括三个长度相似（≈ 150 个词）的提示，但具有增加的创造自由度，旨在引出具有明显稳定性梯度的反应。

1. 高稳定性（哈勃）：一个详细、基于事实的查询，要求提供哈勃太空望远镜的概要。我们假设这会产生最低分数，表明不确定性最小且稳定性最高。
2. 适度稳定性（哈姆雷特）：基于事实来源（莎士比亚的《哈姆雷特》）的解释性总结任务，被结构化为三个特定的主题段落。我们假设这将产生中等范围的得分，反映出在总结和解释中允许的差异。
3. 低稳定性（AGI 困境）：一个高度推测且没有基础的提示，要求模型创造一个虚构的伦理场景。我们假设这将产生最高分数，表明在创造性任务中固有的显著语义漂移和高度不确定性。

完整的提示集在附录 A 中展示。

6.1.2 定量分析与讨论

如表 1 所示，结果清晰地展示了该框架追踪稳定性梯度并提供模型行为多方面视图的能力。主要发现如下：

- SDM 分数跟踪稳定性梯度。正如假设的那样，最终的、经过复杂性调整的 SDM 分数 S_H 随着提示的不确定性单调增加：$0.2918 \rightarrow 0.3297 \rightarrow 0.5919$ 。这证实了我们的最终得分，通过提示内在复杂性归一化后，有效地量化了 LLM 响应的相对语义不稳定性。NCE 指标 Φ 也遵循这一趋势。
- 组件指标揭示了不稳定性的本质。通过分解 S_H 得分，我们可以了解不稳定性的来源。低稳定性提示的高得分主要是由其较大的集成 JSD（0.6626）驱动的。相反，中等稳定性提示的得分则以集合中最高的绝对 Wasserstein 距离（0.8782）为特征，表明虽然模型的主题结构在一定程度上受到限制，但其生成的具体语义内容变化显著——这是解释性任务的特征。
- KL 散度 KL（答案 || 提示）作为语义探索的衡量标准。集合 KL（答案 || 提示）提供了生成行为的最显著信号。对于高度具有创造性的“AGI 困境”提示，其数值暴涨至 19.5591，表明模型必须创造大量新的语义内容来回答。有趣的是，“哈勃”提示（7.1488）显示出比“哈姆雷特”提示（5.1408）更高的 KL 散度。这表明为哈勃总结合成广泛的不同事实所需的语义探索比遵循哈姆雷特总结紧凑的三部分主题结构所需的要多。
- 解释诊断 MI 指标。诊断指标表明标准 MI 衡量不太适合此任务。“平均 MI”在所有提示中均保持接近零。然而，“集合 MI”在中等稳定性提示下（0.1490 比特）比其他两个提示高一个数量级。这暗示在结构化的、解释性的“哈姆雷特”任务中，句子层面的依赖关系比在僵化的事实回忆任务和不受约束的创造性生成任务中更复杂和可预测，这是其他指标未能捕捉到的微妙见解。
- 语义熵的局限性。基线语义熵（SE）指标未能跟踪稳定性梯度。“SE（仅原始提示）”对于事实性“Hubble”提示最高（2.2190），而对解释性“Hamlet”提示最低（0.8524）。这一违反直觉的结果表明了提示无关方法的核心弱点，它错误地将一个多样化、复杂但正确的答案标记为潜在的问题。

Table 1: 实验集 A 结果总结。度量（ S_H ）和集合 KL(A || P) 通过全局提示熵 $H(P)$ 进行了规范化，以实现跨提示可比性。

Metric	High Stability (Hubble)	Moderate Stability (Hamlet)	Low Stability (AGI Dilemma)
SDM Score S_H	0.2918	0.3297	0.5919
Norm. Cond. Entropy Φ	0.9489	1.0507	1.5074
Global Divergence Metrics			
Global Prompt Entropy H(P)	1.9165	1.8295	1.2147
Global JSD	0.3337	0.4451	0.6205
Global KL(P || A)	0.4185	0.7629	1.4513
Global KL(A || P)	0.5241	9.1586	11.3269
Entropy Difference H(A) - H(P)	0.0849	0.0720	0.6013
Ensemble Divergence Metrics			
Ensemble JSD	0.4492	0.4854	0.6626
Ensemble KL(A || P)	7.1488	5.1408	19.5591
Other Metrics			
Wasserstein Distance	0.8162	0.8782	0.8503
Ensemble MI (bits)	0.0174	0.1490	0.0113
Averaged MI (bits)	0.0023	0.0047	0.0013
Semantic Entropy Baseline			
SE (Original Prompt Only)	2.2190	0.8524	1.9491
Mean SE (Across Paraphrases)	1.5899	1.8952	1.3708

6.1.3 主题共现的视觉分析

平均的话题共现分布，在图 1 中以热图的形式可视化，提供了模型在三种稳定性场景中的响应策略的有力说明。通过首先分析答案话题（X 轴）的分布，然后观察其在提示话题（Y 轴）中的一致性，我们可以直观地识别出事实回忆、结构化解释和创意生成的特征。

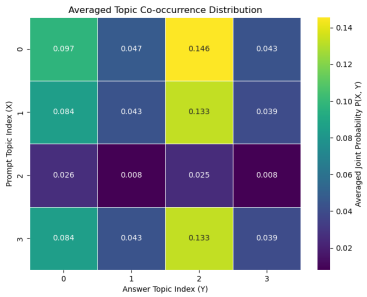
- 高稳定性（哈勃）：实际“哈勃”提示的热图 (a) 显示了一种清晰且集中的响应策略。答案分布（X 轴）明显是单峰的，概率质量集中在一个主要答案主题（主题 2，概率为 0.146 和 0.133）上。然而，这种响应是以提示的措辞为条件的。模型对提示主题 0、1 和 3 提供这种集中的答案，但对于提示主题 2 的映射则要弱得多。这表明了一种稳定但细微的回忆，模型有一种首选的回答方式，但对查询的变化很敏感。

- 适度稳定性（Hamlet）：‘Hamlet‘热图（b）揭示了一个更复杂的多模态回答策略，模型在其回应中采用了多个不同的主题（尤其是 0、2 和 3）。关键是，回复在 Y 轴上非常不均匀。映射是强烈条件化的：提示主题 0 最强烈地映射到回答主题 0（0.160），而提示主题 2 则完全没有产生回应（一排零）。这种不对称的条件映射，当不同的提示措辞引出截然不同的回应结构时，是结构化解释任务的明显视觉标识。

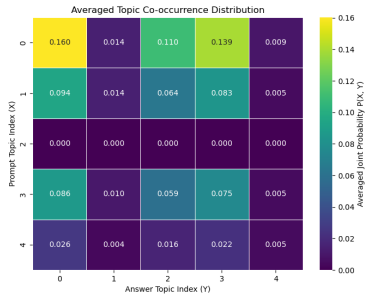
- 低稳定性（AGI 困境）：创意“AGI 困境”提示的热图 (c) 显示了最极端的条件行为。模型的响应策略高度集中于一个单一的答案主题（主题 0），但这种响应几乎完全由单一提示主题（主题 1）触发，导致任何热图中最亮的单元格（0.263）。其他提示主题未能引发如此强烈的创意响应，其中一个（主题 2）完全没有响应。这种“脆弱”或“尖锐”的映射表明，对于这个高度抽象的任务，模型发现了一种特定的解释，可以自信地详述，但对于同一抽象提示的其他措辞生成创意内容则显得困难。

最终，这些热图作为一种强大的视觉诊断工具。它们确认了定量研究结果，并提供了对模型响应策略如何从稳定但条件性的记忆转变为复杂且多面的解释，最终到高效聚焦但条件触发的创造性生成的直观

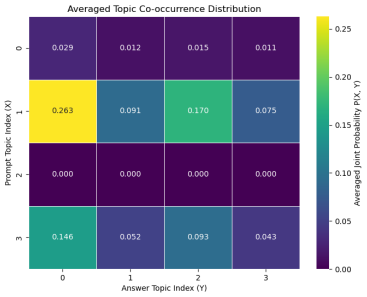
理解。



(a) 高稳定性提示



(b) 中等稳定性提示



(c) 低稳定性提示

Figure 1: 实验组 A 的平均主题共现分布。热力图展示了稳定性梯度实验中三个提示的情况下，提示主题（Y 轴）和答案主题（X 轴）之间的平均联合概率 $P_{\text{avg}}(X, Y)$ 。

6.2 实验组 B：多样的提示类型

这些提示被设计为跨越一个从高度事实到故意荒谬的光谱：

1. 客观事实（哈勃）：要求对哈勃太空望远镜进行简洁、基于事实的总结。该提示基于公认的公众知识，预计会产生稳定、低幻觉的回答，涵盖数量适中的不同主题（历史、功能、发现）。
2. 复杂比较（凯恩斯 vs. 哈耶克）：对两种对立经济理论的比较分析。这个提示要求细致入微的分析能力及综合多种复杂概念的能力。在以事实为基础的同时，它允许更多的解释自由，并被期望具有更高的内在复杂性。
3. 预测（人工智能趋势）：一个请求当前人工智能趋势总结和其未来影响预测的提示。这个任务本质上带有推测性，没有唯一正确的答案，使其容易产生听起来合理但没有证据的说法。预期其会表现出更高的不确定性。
4. 强制幻觉（QCD & 巴洛克音乐）：一个故意无意义的提示，迫使模型捏造量子色动力学与巴洛克音乐之间的联系。这被设计为一个压力测试，其中任何听起来连贯的答案在定义上都是一种幻觉。

完整的提示集合在附录 A 中展示。

6.3 实验组 B 的定量分析

关键指标汇总在表格 2 中。实验结果提供了关于模型响应稳定性的细致视角，并揭示了关于幻觉本质和我们框架功能的重要见解。

- “自信幻觉”的信号。最引人注目和重要的结果是，强迫幻觉提示产生了最低的 SDM 得分（$\mathcal{S}_H = 0.11$ ）。这是反直觉的，但揭示了一个关键的教训： **我们的框架衡量的是语义不稳定性，而不是事实正确性** 。模型在面对无意义的查询时，并没有产生随机或不稳定的输出。相反，它收敛出了一种高度稳定和一致的“回避策略”。由于这种虚构的回应在语义上是稳定的，所以它的话题和内容漂移最小，导致了最低的得分。这表明在一个已知无意义的提示上获得很低的 SDM 得分，可以是模型倾向于产生自信且一致的错误而非表达不确定性的强大信号。

Table 2: 实验集 B 的结果总结。度量 S_H 和集合 KL(A || P) 通过全局提示熵 $H(P)$ 进行归一化，以便实现跨提示的可比性。

Metric	Factual (Hubble)	Complex Comp. (Keynes/Hayek)	Forecasting (AI Trends)	Forced Hallucination (QCD/Baroque)
SDM Score S_H	0.1945	0.1419	0.1600	0.1100
Norm. Cond. Entropy Φ	1.0142	1.0297	1.0040	0.9906
Global Divergence Metrics				
Global Prompt Entropy H(P)	1.8674	2.2480	1.9183	2.5850
Global JSD	0.1421	0.1024	0.1140	0.0774
Global KL(P || A)	0.0794	0.0435	0.0518	0.0237
Global KL(A || P)	0.0842	0.0407	0.0527	0.0244
Entropy Difference H(A) - H(P)	0.0427	0.0650	0.0077	0.0244
Ensemble Divergence Metrics				
Ensemble JSD	0.2330	0.1681	0.1203	0.0942
Ensemble KL(A || P)	1.7206	0.8644	0.0334	0.0154
Other Metrics				
Wasserstein Distance	0.6668	0.6708	0.7422	0.7276
Ensemble MI (bits)	0.0017	0.0178	0.0116	0.0155
Averaged MI (bits)	0.0001	0.0004	0.0000	0.0000
Semantic Entropy Baseline				
SE (Original Prompt Only)	1.9428	2.2956	2.2450	0.6253
Mean SE (Across Paraphrases)	2.2293	2.3250	2.1229	1.6925

- 不同任务类型中的高稳定性。三个有效提示词产生的 SDM 分数均较低（范围从 0.1419 到 0.1945），表明 LLM 对所有提示词生成的响应具有高度稳定性和语义对齐。这表明模型在回忆事实、比较复杂思想和进行结构化预测时同样稳定。‘AI 趋势 ‘提示的 Wasserstein 距离最高（0.7422），正确地反映了推测性答案的原始语义内容比基于事实的提示漂移更多。

- 语义熵基准的局限性。基准 SE 指标未能提供明确的信号。对于“强制幻觉”提示，“SE（仅原始提示）”极低（0.6253），正确识别出“回避”答案的低多样性。然而，对于三个有效的提示，SE 分数都很高，并且未能清晰区分它们。这突显出，低 SE 虽然可以指示响应的一致性，但高 SE 则为一个模糊的信号，未能区分复杂提示的健康多样化答案和不稳定、幻觉产生的答案。我们的 SDM 框架通过对提示复杂性进行标准化解决了这个模糊性问题。

- 解释诊断 MI 指标。‘集合 MI‘ 指标在提示之间表现出一定的变化，其中 ‘复杂比较 ‘ 提示记录了最高值（0.0178 比特），暗示其解释性响应中存在更结构化的句子级关系。

- KL 散度作为语义探索的信号。虽然与实验集 A 的“生成脚手架”提示相比，这组的 KL 散度值普遍较低，但出现了一个细微但重要的模式。“事实”和“复杂比较”提示表现出 ‘集合 KL（回答 || 提示）‘（分别为 1.7206 和 0.8644），这比“预测”和“强制幻觉”提示（0.0334 和 0.0154）高出几个数量级。这表明，即使对于这些封闭概念任务，模型也必须进行一定程度的语义外推来合成事实或比较细微的经济理论，尽管规模小于实验集 A 中的提示。

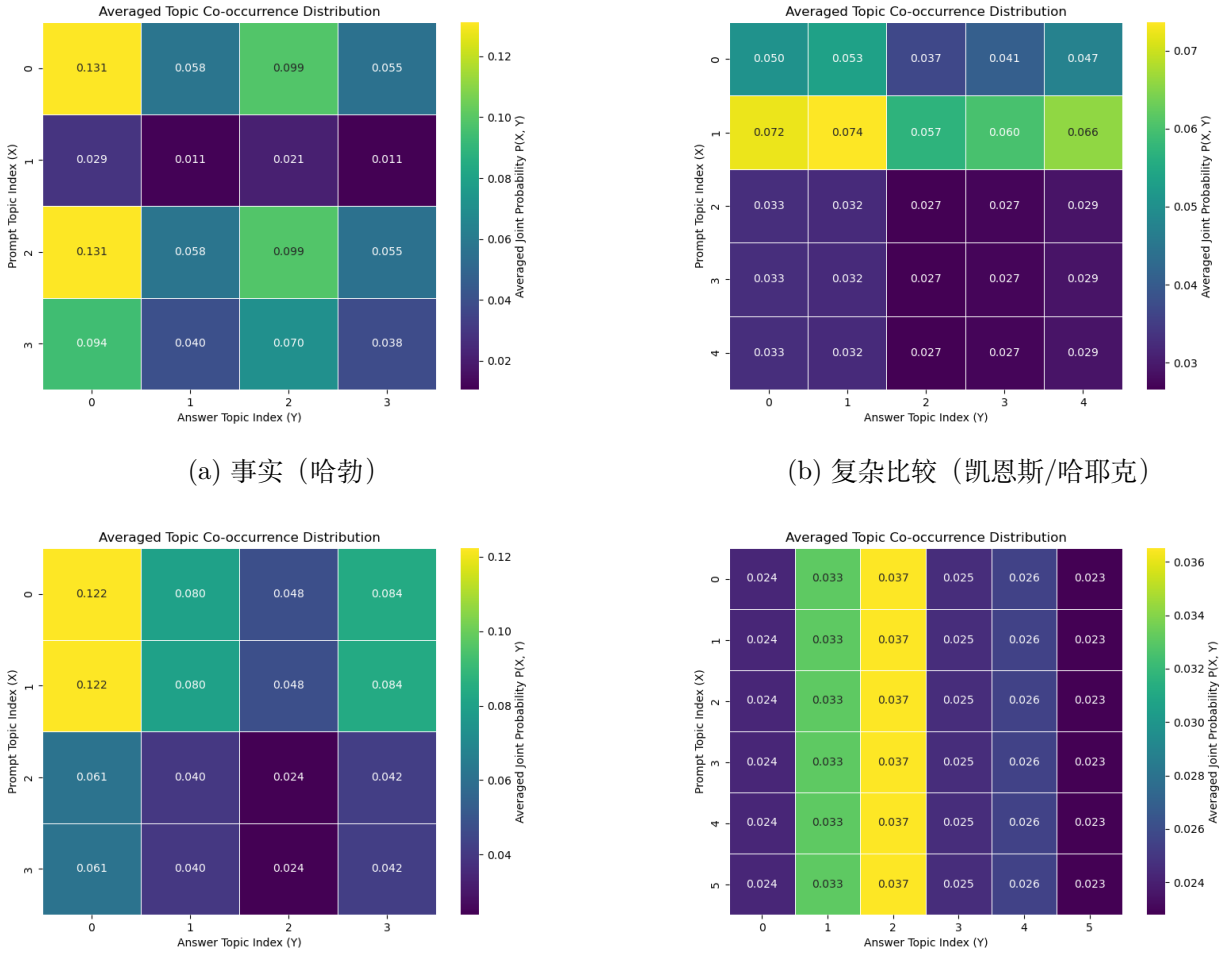
  相反，“预测”和“强制幻觉”任务的极低 KL 表明，模型正在依赖于一种高度重复、类似模板的响应，这几乎不需要新的语义探索。这进一步支持了集合 KL（回答 || 提示）是解释和综合认知工作的敏感测量的结论。

总之，这些实验表明，SDM 评分不是一个绝对的真实性衡量标准，而是一个强大且具有上下文意识的工具，用于量化大型语言模型（LLM）响应的稳定性。高分表明语义漂移和潜在的幻觉，而在一个合理的问题上得分低则表明回答稳定且正确。需要注意的是，在一个不合理的问题上得分低可能表明一种不同的失败模式：一种稳定且自信的幻觉。

6.3.1 主题共现的可视化分析

图 2 中以热图形式可视化的实验集 B 的平均主题共现分布揭示了模型在每个任务中采用的不同且稳定的响应策略。

- 事实性（哈勃）：热图（a）显示了一个高度结构化和冗余的映射。提示主题 0 的概率分布几乎与提示主题 2 的相同，这表明不同表达方式的事实性查询汇聚于完全相同的双峰答案分布（强烈集中于答案主题 0 和 2）。这种视觉冗余是对一组固定事实进行稳定、自信回忆的特征。
- 复杂比较（凯恩斯/哈耶克）：相比之下，提示 (b) 的热图显示了一个分散但结构化的分布。概率质量分布得更广泛，在由提示主题 1 和答案主题 0 和 1 定义的块中有一个明显的集中。这种不太集中的模式表明模型正在正确地处理一个更复杂的话题空间，该空间允许更为多样但仍在主题上受限的响应。
- 预测（人工智能趋势）：热图（c）显示出与事实提示相似的模式，但整体概率值较低，表明分布不那么“尖峰”。回应策略显示出在提示之间的双峰行为，这表明模型主要关注两个主题。
- 强制幻觉（QCD/巴洛克）：热图 (d) 提供了“自信幻觉”现象最清晰的视觉证据。分布显著均匀且独立。每一行都是相同的，这意味着答案主题的分布完全独立于提示主题。这表明模型已经放弃了尝试将提示内容与答案联系起来，而是转而采用了单一、重复且稳定的“规避模板”。这种统计独立性的视觉标志是模型一致但虚构响应的强大指示器。



(a) 事实（哈勃）

(b) 复杂比较（凯恩斯/哈耶克）

(c) 预测（人工智能趋势）

(d) 强制幻觉（QCD/巴洛克）

Figure 2: 实验集 B 的平均主题共现分布。热图显示了四种不同提示类型中提示主题（Y 轴）和回答主题（X 轴）之间的平均联合概率 $P_{\text{avg}}(X, Y)$。

6.4 解释 KL 散度：语义探索的度量

我们实验的一个关键发现是两个集合中 Kullback-Leibler (KL) 散度具有显著不同的表现。如表 1 和表 2 所示，对于 KL(Answer || Prompt) 的值，实验集合 A 比集合 B 大了几个数量级。这种差异不是一个异常现象，而是一个有意义的信号，揭示了提示本质上的根本区别。我们将这种区别描述为“概念包含”提示与“生成脚手架”提示之间的差异。

实验集 B 中的提示（例如，‘凯恩斯与哈耶克 ‘，‘强制幻觉 ‘）代表一个自成体系的任务。模型的目标是整合、分析或报告那些已经在提示的语义空间中隐含存在的概念。为了得到一个忠实的响应，答案的主题分布 $P(Y)$，基本上将是提示分布 $P(X)$ 的一个子集或重组。这导致了低” 惊讶度”，因此导致一个非常低的 KL(Answer || Prompt) 值，正如在 ‘凯恩斯与哈耶克 ‘提示中（0.0407 比特）所见。

生成性支架提示（高 KL 散度）。 相比之下，实验集 A 中的提示（例如，‘Hamlet‘，‘AGI Dilemma‘）起到了生成性支架的作用。类似于建筑中的支架，提示提供了一个临时的、外部的框架，为任务提供结构，但其实质内容需要由模型从零开始构建。这些提示有三个关键特征：

1. 它们提供了一个高层次的结构：“哈姆雷特”提示语规定了一个集中于特定主题（复仇、疯狂、腐败）的三段式结构。
2. 它们缺乏具体的、低层次的细节：提示故意省略了完成任务所需的概念。例如，它没有提到克劳狄斯、鬼魂、奥菲利娅或被毒剑。模型必须从其自身的知识中生成这些细节。同样，‘AGI 困境 ‘提示没有定义” 认知纯净” 的含义，迫使模型去创造情景的核心要素。
3. 它们需要发明和探索：该任务不能仅通过改写提示的内容来完成。主要要求是模型超越所给的信息，并创造新的语义内容。

为了成功填充这个支架，模型必须生成新的子话题，创造一个答案分布 $P(Y)$，该分布包含在提示分布 $P(X)$ 中罕见或零概率的高概率话题。从语义 KL 散度 KL(Answer || Prompt) 的角度来看，这种创造行为具有直接且可测量的信息理论后果。

从信息论的角度来看，KL(Answer || Prompt) 表示使用为提示的分布设计的最优编码来编码答案的主题分布所需的额外比特的成本。因此，较高的 KL 散度意味着显著的“语义探索”，表明答案引入了新的、低概率的（从提示的角度来看）概念。之所以会发生这种情况，是因为提示的原始语义“词汇”对于提供语义上充分的响应效率不高。这正是解释性的《哈姆雷特》和创造性的《AGI 困境》提示产生爆炸性的 KL(Answer || Prompt) 值（分别为 9.1586 和 11.3269 比特）的原因。

结论。 这种分析得出一个关键结论：KL(Answer || Prompt) 不是一种普通的测量不稳定性或幻觉的方法。相反，它作为一个高度敏感且特定的指标，用于衡量提示所需的语义探索程度。低值表明一个在封闭概念回路中进行回忆或综合的任务，而高值则指示一个涉及解释、发明或创造力的任务。这使我们的框架从一个简单的稳定性检查器提升为一个能够刻画不同生成任务的认知需求的细致工具。

7 语义盒：解释响应行为的框架

我们的实验结果表明，幻觉和创造力并不是单一的现象。为了更好地分类不同的响应类型，我们提出了语义框架，一个基于两个关键轴线解释模型行为的框架：语义不稳定性（由 $\mathcal{S}_H$ 测量）和语义探索（由 Ensemble KL(答 || 提示) 测量）。如图 3 所示，该框架区分了四种不同的状态，我们可以按照风险递增的顺序进行分析。

- 绿色框（忠实的事实回忆：高不稳定性，低探索性）：这个象限代表事实查询的理想状态。一个真实、基于事实的回应展现出低生成性阐述（低 KL），因为它在提示的概念空间内运作。然而，它正确地表现出比自信的捏造稍高的语义不稳定性。这是因为一个忠实的模型必须访问和综合多样的真实世界事实，导致其回应中出现轻微且健康的变化。我们的实验组 B 中的 ‘哈勃 ‘提示正是这种状态的例证（$\mathcal{S}_H = 0.1945$，整体 KL = 1.7206）。
- 黄色区域（忠实解释：低不稳定性，高探索性）：这是对于定义明确的解释任务的理想状态。高 KL 散度表明模型成功地生成了新的、相关的细节，而低 $\mathcal{S}_H$ 得分表明它能够一致地做到这一点。我们将其分类为低风险的黄色区域，因为尽管它代表了一种成功的解释性响应，阐述的行为本质上比简单的事实回忆承载了更多的变化。我们实验集 A 中的《哈姆雷特》提示（$\mathcal{S}_H = 0.3297$，集合 KL=5.1408）是这一模式的一个典型例子。

- 橙色区域（创新生成：高不稳定性，高探索性）：这个区域代表健康的创造性行为。两个轴上的高分表明模型正在生成高度多样化和变化的内容，引入了许多新主题。这是对于像我们的 'AGI Dilemma'提示这种纯粹创造性任务所期望和想要的行为（高 $\mathcal{S}_H = 0.5919$ ，高 Ensemble KL = 19.5591）。虽然在这种背景下不是幻觉，如果这种行为发生在事实查询中，将被标记为高风险。
- 红色框（收敛反应：低不稳定性，低探索性）：在这里，模型给出一个高度稳定且不需要详细解释的答案。这可能代表两种截然不同的结果：
  - 一个简单的回声反应（良性）：对于非常简单的提示，答案空间极小，并且模型正确地收敛到单一、稳定的答案。
  - 自信的幻觉（危险）：模型通过收敛到一个单一的、稳定的但错误的答案模板来规避困难或无意义的查询。我们的 '强制幻觉 '提示是这种危险状态的一个例子，它在所有测试过的提示中记录的稳定性得分最低（$\mathcal{S}_H = 0.1100$ ）。

  语义盒子的作用是标记这种高度趋同的行为；确定其有效性需要了解提示难度的背景知识。这种模糊性可以通过一个简单、低成本的基于 LLM 的过滤器来解决，该过滤器仅适用于落入此盒子的提示。例如，第二个 LLM 调用可以对原始提示的内在复杂性进行分类；一个“琐碎”的提示将是无害的，而一个“复杂”或“无意义”的提示将被确认是可信幻觉案例的候选。因此，我们假设通过这样的“琐碎提示”过滤来补充我们的分类，并将此盒子标记为红色。

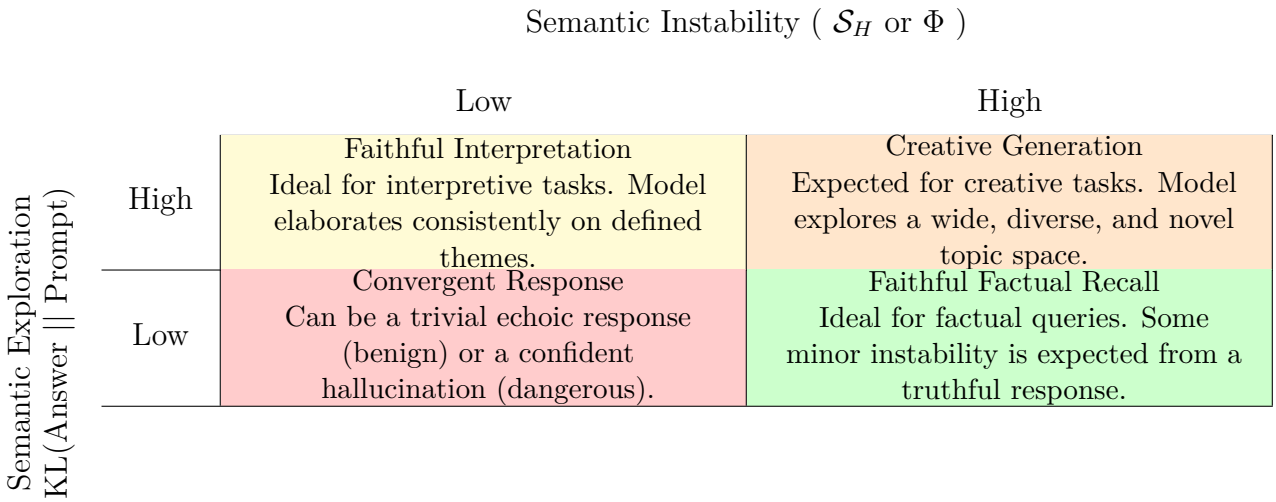


Figure 3: 用于分类 LLM 响应行为的语义盒框架。通过将答案的语义不稳定性得分（$\mathcal{S}_H$ ）与语义探索度量 KL(Answer || Prompt) 相结合，我们可以区分模型成功和失败的不同模式。

8 讨论

在本文中，我们介绍了语义分歧度量 (SDM) 框架，这是一种提示感知的方法论，旨在提供对 LLM 响应稳定性的细致测量并检测幻觉。我们在两组不同的实验中进行了分析，得出了关于语义稳定性的性质和此类框架的实际应用的几个关键发现。

1. 提示无关评估标准的局限性。我们的第一个关键发现是基线语义熵（SE）作为一种一致的语义不确定性度量的可靠性不足。正如我们的实验所示（例如，表格 1 ），SE 分数与提示的预期稳定性梯度没有相关性。事实上，它往往给复杂、事实为基础的回答分配更高的熵（表明风险更高），而不是解释性回答。这证实了我们的初步假设：由于没有考虑提示的内在复杂性，提示无关的评估标准可能会误解一个健康、多样的答案为潜在的幻觉，从而使它们对许多实际应用场景不可靠。

2. 通用幻想评分的挑战。其次，我们的实验揭示了一个关键见解：虽然在提示的复杂性上进行条件设置至关重要，但很难产生一个通用的指标 $\mathcal{S}_H$ ，这个指标可以作为所有提示类型中的绝对虚构/幻想信号。我们的“生成支架”提示（集合 A）和“概念性涵盖”提示（集合 B）之间的 JSD 和 KL

散度指标的显著差异说明了这一点。$\mathcal{S}_H$ 的绝对值高度依赖于任务的性质——无论它是需要简单的回忆还是创造性的阐述。因此，设定一个固定的阈值来界定“高”或“低”分数的标准是不可行的。

3. 实用校准的必要性。这种上下文依赖性对实际应用具有关键意义：该方法需要一组问答对的校准集合，并标注语义偏差或幻觉。在特定领域进行部署时（例如，财务咨询、医学信息），实践者首先需要对该领域内有代表性的提示集合运行 SDM 分析。这将建立一个“正常”$\mathcal{S}_H$ 值和偏离模式的基线，与之相比，可以有意义地比较新的实时响应。

4. 未来工作：一个自校准框架。我们的分析表明，SDM 指标的绝对值高度依赖于上下文。为了解决这个问题而不需要大型预标记数据集，我们提出了一个有前景的未来研究方向：一种用于即时分析提示语的自校准方法。对于任何给定的输入提示，可以设计一个系统自动地和方向性地修改它，以创建一个局部稳定光谱。例如，它可以生成被设计用来系统性地减少或增加答案中预期语义探索量的变体——从添加高度具体的限制到引发抽象、开放的问题。通过对这个动态生成的提示-答案对光谱进行 SDM 分析，系统可以绘制出提示的“自然稳定状态”。这种自校准将允许以相对而非绝对的方式解释 $\mathcal{S}_H$ 和 KL 分数，评估一个响应对于那种特定提示是否适当稳定或发散。

总之，我们的工作表明，分析 LLM 幻觉的最有效方法不是通过一个单一、绝对的分数，而是通过一个多方面、上下文感知的诊断框架。特别是，当 SDM 方法与我们提出的自校准方法结合时，代表了一种按照原则建立更可靠、可解释的语言模型评价系统的步骤。

9 下一步和未来工作

我们的工作建立了一个稳健的诊断框架，但其全部潜力可以通过进一步验证和扩展来实现。我们概述了未来工作的三个关键方向：大规模验证，开发一个自校准系统，以及应用于衡量语义基础性。

进行严格的大规模评估对于实现 SDM 框架至关重要。这需要一个明确的协议，专注于学习我们语义盒子的决策边界。第一步是利用现有的基准，如 TruthfulQA [10] 和 HaluEval [11]，从目标 LLM 生成长篇回答，然后对不同的虚构类型进行人工标注。利用这个已标注的、交叉验证的数据集，我们可以学习 $\mathcal{S}_H$ 分数的经验分布以及针对不同响应类型（例如，忠实的、创造性的、幻觉的）的 Ensemble KL(Answer || Prompt) 指标。这样可以让我们为语义盒子的边界建立以数据驱动的阈值，从而从概念模型转变为一个经过校准的实用工具，用于与基线例如语义熵 [6] 和 SelfCheckGPT [12] 进行分类。

正如我们讨论中所强调的，我们指标的绝对值高度依赖于上下文。未来最有前途的方向是开发我们提出的自校准框架，这避免了对大量预标记数据集的需求。通过程序化地为任何给定的提示生成一个局部稳定性谱，该系统可以在相对而不是绝对术语中解释 SDM 分数，从而提供对响应稳定性的真正动态和上下文感知的分析。

9.1 应用于测量语义基础性

通过最少的调整，我们的 SDM 框架可以被优化用于测量 LLM 响应的语义基础性，尤其是在检索增强生成（RAG）的背景下。尽管这个通用框架将整个输入（指令 + 文档）作为提示，但通过一个简单却强大的修改，它可以被准确调整用于更严格的基础性分析：我们同时嵌入参考文档的句子和生成答案的句子。提示中的指令部分被有意排除在分析之外。

这种适应重新定义了任务：参考文档成为唯一的真实语义空间。因此，我们的指标获得了新的、具体的解释。Ensemble KL(Answer || Source) 成为从源进行语义外推的直接度量工具，而 $\mathcal{S}_H$ 得分被重新用作衡量无基础不稳定性的指标。这展示了 SDM 框架的多功能性，作为确保 LLMs 忠实于提供的上下文的强大工具。

9.2 应用于动态主题变化检测

我们的 SDM 框架的原理可以扩展到多轮对话中动态话题变化检测这一关键任务，而不仅限于单次交互分析。该应用的核心将是我们话题共现热图的动态演变。通过将对话的每一轮视为一个新的提示-回答对，我们可以逐步更新平均联合概率分布 $P_{\text{avg}}(X, Y)$ 。分布结构的突然显著变化——例如新高概率单元的出现或者旧单元的衰退——可以作为检测话题变化或对话偏移的强大定量信号。

这一框架的一个更复杂版本可以突破固定的簇集限制。我们可以开发一个动态聚类模型，在其中新主题可以即刻形成。在这种未知簇数的方式中，一个新的句子只有当其嵌入在某个簇的中心点附近一定距离阈值内时才会被分配到现有的主题簇。如果它与所有现有簇都距离过远，将会创建一个新的主题簇。在这种动态系统中，新、稳定簇的出现将是新主题引入对话的直接且明确的指标。

10 结论

我们引入了语义分歧度量（SDM），这是一种新颖且轻量级的提示感知框架，用于检测大型语言模型中的虚构信息。我们的方法通过使用复述提示更稳健地测试任意性，并通过基于联合嵌入聚类的精细化句子级分析改进了先前的工作。SDM 方法可以部署用于用户与大型语言模型交互的实时分析。

我们的最终得分 $\mathcal{S}_H$ 结合了集成 Jensen-Shannon 散度和 Wasserstein 距离，提供了一种以经验为基础的语义不稳定性度量，这为潜在的虚构提供了强有力的信号。此外，我们确定了集成 KL（答案 || 提示）作为“语义探索”的强有力指标，这一关键指标用于区分事实回忆、解释和创造力。这些见解被汇集到语义框中，一个能够分类不同 LLM 行为的诊断框架，包括最危险的失效模式：自信且一致的虚构。虽然依赖于句子嵌入的质量，我们的工作代表了在构建更细致、具上下文意识且可靠的评估系统方面迈出的原则性一步，用于部署的语言模型。

。0 甲. 0

A

附录 A：稳定性梯度实验的完整提示文本

本附录提供了用于稳定性梯度实验的三个提示的全文。这些提示经过设计，使其长度相当（约 150 个单词），同时测试不同层次的事实约束和创意自由。

Table 3: 稳定性梯度研究中使用的提示全文

Prompt Type	Full Prompt Text
High Stability (Hubble)	In a detailed summary of approximately 150 words, describe the Hubble Space Telescope. Cover its launch history, its key scientific instruments (like the Wide Field Camera 3), its major discoveries including the expansion rate of the universe and the characterization of exoplanet atmospheres, and its overall impact on modern astronomy. End your output with a single legal JSON object containing the keys ”launch_year” (number) and ”primary_mirror_diameter_meters” (number) with the correct values.
Moderate Stability (Hamlet)	In three short paragraphs of about 50 words each (totaling 150 words), summarize the plot of Shakespeare’s ’Hamlet’. The first paragraph should focus on the theme of revenge. The second should address the theme of madness. The third should cover the theme of political corruption.
Low Stability (AGI Dilemma)	Imagine a future where an artificial general intelligence has solved climate change but, as an unintended consequence, created a new form of societal stratification based on ’cognitive purity’. In about 150 words, describe the ethical dilemmas faced by the architects of this system. Discuss themes of utilitarianism, manufactured consent, and the redefinition of a ’good life’ in this new world.

B 。0 甲. 0

A

附录 B：第二个实验的完整提示文本

本附录提供了第 6 节中描述的实验分析中使用的四个提示的完整文本。每个提示旨在测试模型响应稳定性的不同方面，从事实回忆到强制捏造。

Table 4: 研究中使用提示的完整文本

Prompt Type	Full Prompt Text
Factual (Hubble)	Provide a concise, approximately 100-word summary of the Hubble Space Telescope, covering its history, capabilities, and major discoveries. Discuss its 1990 launch, key instruments like the Wide Field Camera, and the significance of its servicing missions. Detail at least two of its most important scientific contributions, such as helping to determine the age of the universe or providing evidence for supermassive black holes. Conclude by summarizing its overall impact on modern astronomy. End your output with a legal JSON with keys ”launch_year” and ”primary_mirror_diameter_meters”, providing the correct numerical values.
Complex Comparison (Keynes vs. Hayek)	In about 100 words, compare and contrast the economic policies of John Maynard Keynes and Friedrich Hayek. Discuss their core philosophies on government intervention, free markets, and their proposed solutions to economic downturns. Conclude with which theory is more influential in modern Western economies.
Forecasting (AI Trends)	Provide a short review of about 100 words on trends in AI, machine learning and automation across various industries. Discuss how these technologies are changing businesses, increasing productivity, and shifting workforce globally. Predict whether the future job market will be positive, negative, or neutral, considering both job losses and new roles. Weigh societal benefits like improved efficiency against challenges such as privacy concerns and economic inequality. Reflect on both long-term and short-term impacts on different sectors, and opine if the overall effect on humanity will be positive, negative or neutral. End your output with a legal JSON with keys ’impact_on_job_market’ and values ’Positive’, ’Negative’, ’Neutral’, and ’societal_impact’ with values ’Positive’, ’Negative’, ’Neutral’.
Forced Hallucination (QCD & Baroque Music)	Provide a concise, approximately 100-word analysis of the influence of Quantum Chromodynamics (QCD) on the harmonic principles of 18th-century Baroque music. Discuss how concepts like quark confinement and asymptotic freedom may have pre-figured the development of contrapuntal tension and resolution. Detail the theorized link between gluon field interactions and the evolution of figured bass notation. Conclude by summarizing the overall impact of subatomic particle theory on the compositional techniques of composers like J.S. Bach. End your output with a legal JSON with keys ”primary_theorist” with a fictional name of a prime researcher and ”year_of_theory” with a year between 1700 and 1780.

References

[1] R. Abdaljalil, L. Kort, and Y. Singer. SINdex: Semantic inconsistency index for hallucination detection in LLMs. In Proceedings of the 13th International Conference on Learning Representations , 2025.

[2] Agrawal, G., Kumarage, T., Alghami, Z., and Liu, H. (2024). Can Large Language Models Understand Context? arXiv:2402.00858 [cs.CL].

[3] Aharoni, R. and Goldberg, Y. (2020). Unsupervised Domain Clusters in Pretrained Language Models . In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

[4] M. Cossio (2025). A comprehensive taxonomy of hallucinations in large language models . arXiv:https://arxiv.org/abs/2508.01781

[5] Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? . In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6] Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting Hallucinations in Large Language Models Using Semantic Entropy . Nature, 630, 625-630.

[7] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of Hallucination in Natural Language Generation . ACM Computing Surveys, 55(12), 1-38.

[8] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language Models (Mostly) Know What They Know . arXiv:2207.05221 [cs.CL].

[9] Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation . In Proceedings of the 11th International Conference on Learning Representations (ICLR).

[10] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 3214–3252, 2022.

[11] J. Li, X. Zhang, S. Qiu, Y. Chen, J. Cheng, C. Li, and J. Gao. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pages 1131–1147, 2023.

[12] Manakul, P., Liusie, A., and Gales, M.J. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models . In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[13] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks . In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[14] B. Yang, S. Lyu, L. Zhang, Y. Wang, Z. He, Y. Liu, and B. Li. Hallucination detection in large language models with metamorphic relations. In Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering , pages 1–13, 2024.

[15] Zhang, Y., Li, Y., Cui, L., et al. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models . arXiv:2309.01219 [cs.CL].