一轮中的信任：基于结构信号的大语言模型置信度估计

杨鹏越 悉尼大学 澳大利亚悉尼
pyan8493
@uni.sydney.edu.au

Jiawen Wen 悉尼大学 澳大利亚悉尼
jwen8784@uni.sydney.edu.au

Haolin Jin 悉尼大学 澳大利亚悉尼
hjin3177
@uni.sydney.edu.au

Linghan Huang 悉尼大学 澳大利亚悉尼
lhua5130
@uni.sydney.edu.au

Huaming Chen*
悉尼大学
悉尼，澳大利亚
huaming.chen@sydney.edu.au

Ling Chen 悉尼科技大学
悉尼，澳大利亚
ling.chen@uts.edu.au

arXiv:2602.00977v1 [cs.CL] 1 Feb 2026

摘要

大语言模型（LLM）正越来越多地应用于错误可能造成重大社会、科学或安全代价的领域。然而，标准置信度估计器（如词元似然、语义相似度和多样本一致性）在分布偏移、领域专用文本和计算量受限的情况下仍然脆弱。本文提出结构置信度（Structural Confidence），这是一种单次遍历、与模型无关的框架，基于从模型最后一层隐藏状态轨迹中提取的多尺度结构信号，提升输出正确性预测能力。通过结合谱、局部变化和全局形状描述符，我们的方法能够捕捉概率和句子嵌入所遗漏的内部稳定性模式。我们在四个异质基准测试上进行了广泛的跨领域评估：FEVER（事实核查）、SciFact（科学主张）、WikiBio-hallucination（传记一致性）和 TruthfulQA（真实性导向问答）。与既有基线方法相比，我们的结构置信度框架在 AUROC 和 AUPR 指标上表现出色。更重要的是，不同于需要多次随机生成并依赖辅助模型的基于采样的一致性方法，我们的方法只需一次确定性前向传播，为在具有社会影响且资源受限的大语言模型应用中开展高效、稳健的事后置信度估计提供了切实可行的基础。

CCS 概念

• **计算方法** → **神经网络**；**知识表示与推理**；**学习潜在表征**。

关键词

大语言模型、不确定性量化、幻觉检测、隐藏状态轨迹

*通讯作者

本作品依据知识共享署名 4.0 国际许可协议授权。*WWW '26*，阿拉伯联合酋长国迪拜© 2026 版权所有者/作者保留版权。ACM ISBN 979-8-4007-2307-0/2026/04 https://doi.org/10.1145/3774904.3793055

ACM 参考格式：
杨鹏越、Jiawen Wen、Haolin Jin、Linghan Huang、Huaming Chen 和 Ling Chen。2026。一轮中的信任：基于结构信号的大语言模型置信度估计。载于*2026* 年 *ACM Web* 会议论文集（*WWW '26*），*2026*年*4*月*13*日至*17*日，阿拉伯联合酋长国迪拜。ACM，纽约，NY，美国，11 页。
https://doi.org/10.1145/3774904.3793055

1 引言

大语言模型（LLM）如今已成为搜索引擎、对话智能体、科学助手和决策支持系统的核心组件，并日益融入面向开发者的工作流程和智能体工具链（例如软件工程）[15]。然而，尽管大语言模型生成流畅度出众，却经常生成错误或虚构的内容，这类现象通常称为幻觉 [14, 32]。生成流畅度与事实忠实度之间的差距，使得事后置信度估计成为安全敏感型应用中的关键需求，例如事实核查、推理和内容审核。

现有置信度估计方法主要分为三类，每类都有其固有局限。（1）基于概率的估计器，例如平均对数概率、词元熵和基于校准的分数，在分布偏移和校准不充分的情况下并不可靠 [12, 30]。（2）语义相似度估计器依赖嵌入几何（例如 Sentence-BERT [33]、基于 DeBERTa 的分类器），当输入文本偏离模型的预训练分布时，其性能可能会严重下降 [1, 20]。（3）多样本一致性方法（如 SelfCheckGPT） [25] 需要反复进行随机采样，并使用辅助自然语言推理（NLI）模型，因此会造成高昂的延迟和计算成本，不适用于高吞吐量或实时应用。此外，另一种新兴研究方向探讨隐藏状态几何信号，考察词元级激活的变化轨迹，而非输出概率或嵌入空间相似度 [8, 40]。

这三种范式并非临时拼凑，而是各自针对置信度—成本权衡的不同维度进行优化。基于概率的评分之所以经济高效，是因为它们直接从模型自身的输出分布中读取不确定性，无需额外组件，且相较于常规解码，只会带来较小的固定倍数开销 [12, 30]。语义方法利用预训练编码器或句子表示引入高层次的相似性结构；当测试分布与预训练语料库

高度匹配时，这类方法能实现较强的域内判别能力 [20, 33]。SelfCheckGPT 等多样本一致性方法 [25]则有意以延迟和计算量换取鲁棒性：它们通过汇总多个随机生成结果之间的一致性模式，并借助辅助的自然语言推理或问答模型，在无需标注数据的情况下，近似获得一种经参考信息增强的验证信号。

然而，这些设计选择为 Web 规模部署带来了结构性矛盾。基于概率的估计器会继承模型的校准偏差，在分布偏移下变得脆弱；语义估计器依赖嵌入流形的稳定性，而该流形在领域专用或长文本中可能发生坍塌；基于采样的一致性方法（如 SelfCheckGPT）则天生成本高昂，因为它们需要结合多次生成与辅助模型。此外，针对隐藏状态几何信号的现有研究仍然稀少且零散，既未提供统一框架，也未能稳健地刻画可跨领域或模型家族泛化的特征。因此，要实现高吞吐量部署，就需要一种置信度信号，它既要（i）遵循严格的单次遍历和模型无关约束，又要（ii）在领域偏移下保持稳健——这促使我们从概率空间和语义空间转向隐藏状态轨迹的结构稳定性。

这些方法共有的一个关键局限是依赖表层信号，例如词元似然、嵌入相似度或跨样本一致性，却忽略了生成模型预测的内部计算过程。在解码过程中，大语言模型会生成一系列隐藏状态，反映其不断演变的推理动态。近期研究表明，这些隐藏状态轨迹会呈现与不确定性相关的独特结构模式，包括激活方差、不稳定的注意力转移[27], 以及潜在空间中的几何或拓扑异常 [7, 37]。然而，尽管有这些发现，轨迹级结构作为一种独立的主要置信度信号，仍未得到充分探索。

本文提出**结构置信度**，这是一种框架，将隐藏状态结构稳定性 建模为置信度指标，旨在我们研究的事实性和幻觉类场景中具备跨分布的稳健性。我们的核心假设是，置信预测对应于平滑、低频且全局几何结构紧凑的轨迹，而不确定性则表现为谱不规则性、明显的局部波动，以及分散或碎片化的全局结构。通过捕捉这些多尺度动力学扰动，结构置信度旨在揭示一类置信度线索：当基于概率和语义的信号退化时，这些线索往往仍能提供信息，尤其是在领域偏移的情况下。我们的贡献如下：

一种基于隐藏状态结构稳定性的新型置信度模态。我们提出结构置信度框架，这是一种基于隐藏状态结构稳定性的新型置信度模态。我们定义轨迹级结构描述符，涵盖词元级隐藏状态的谱稳定性、局部变化和全局形状连贯性，并将其作为一种具有理论依据且经实证验证的置信度信号，能够稳健地补充现有的基于似然、嵌入和采样的估计器。

一种模型无关的单次前向置信度估计器。核心估计器基于从单次确定性前向传播中提取的结构特征运行，无需采样或辅助模型，也不依赖架构特定内部机制。我们进一步表明，将这些结构信号与句子级语义表示相融合，可以得到一种轻量且具有良好泛化能力的估计器，无需额外调用大语言模型。

大规模跨领域评估。我们在 FEVER [36], 、SciFact [38], 、Wik-iBio [19], 和 TruthfulQA [21] 上进行了大规模跨领域评估。结构置信度在 AUROC 和 AUPR 指标上达到最先进水平，同时展现出应对领域偏移的鲁棒性。具体而言，在混合领域基准中，我们的方法效率显著更高。例如，在置信度估计阶段，基于自然语言推理的 SelfCheckGPT 一致性基线平均需要比我们的方法多 5–6× 倍的浮点运算次数，延迟也高出 4–5× 倍。

总而言之，这些发现表明，在我们研究的事实性和幻觉类基准测试中，隐藏状态结构稳定性是进行事后置信度估计的一种前景广阔但尚未得到充分探索的依据。结构置信度提供了一种高效、与模型无关且在实践中对领域变化具有鲁棒性的解决方案，可用于评估大语言模型可靠性，并指明了可信人工智能的一条研究方向，即关注模型激活的内部动态。

2 相关工作

针对大语言模型的事后置信度估计主要采用三种范式：基于概率的估计器、语义相似度估计器和基于采样的一致性方法。另有一类研究考察隐藏状态的几何与拓扑结构，为本文提供理论基础。

基于概率的置信度估计。 早期方法通过平均对数概率、熵或温度缩放置信度等词元级统计量来度量不确定性 [12, 30]。尽管这些信号简单易用，但已知其校准效果较差、对分布偏移高度敏感，并且常常与正确性不一致 [16]。诸如 Prior Networks 等经典的认知不确定性与偶然不确定性分离方法要求显式访问模型逻辑值，因此与我们的黑盒设定不兼容 [24]。近期研究进一步表明，对数概率可能系统性地为幻觉文本或文风流畅的文本赋予高分 [14]。我们的结果证实了这些局限性：基于概率的基线在 SciFact 的领域漂移下失效。

基于语义嵌入的置信度估计。第二种范式依赖语义相似度或嵌入几何来评估置信度，包括选择性预测 [17], 基于表征的嵌入稳定性分析 [18, 29], 以及句子级嵌入分类器 [33]。这些方法在域内表现良好，但当测试分布偏离预训练语料库时，性能会急剧下降 [1, 20]。

基于采样的一致性方法。多样本一致性旨在通过检验模型在重复的随机生成结果中是否得出一致结论来估计置信度。基于推理的幻觉已得到广泛研究，尤其是在思维链场景中 [23]。最突出的例子是 SelfCheckGPT [25], ，它将多个 GPT 样本与辅助的自然语言推理或基于问题生成的检查相结合。按照最初的设计，SelfCheckGPT 需要重复解码和额外的分类器运行轮次，对于简短或含噪的输出而言，这可能成本高昂且不够稳定。在我们的实验中，我们使用 SelfCheckGPT 库中的官方 SelfCheckNLI

检查器作为单次推理基线：我们使用基于 DeBERTa 的自然语言推理模型为单个确定性生成结果打分（不进行多样本解码）。这种设置将基于自然语言推理的检查组件的成本和行为单独呈现，从而能够与其他单次估计器进行更清晰的比较。

隐藏状态几何结构与结构信号。另一条互补的研究路线关注 Transformer 隐藏状态的内部几何结构与动态特性。逐层探测表明，中间层激活编码了丰富的语言和结构信息 [35], , 而机制研究揭示，特定子空间和前馈“键值记忆”实现了可解释的计算模式 [9]。近来，若干论文将不确定性与激活空间中的典型行为联系起来，包括方差增大、曲率变化以及不稳定的注意力动态 [8, 40]。拓扑分析进一步表明，潜在表示具有持久同调结构，这种结构在分布偏移压力下可能发生碎片化 [7]。另有研究关注 BERT 和 RoBERTa 等仅编码器变换器的几何结构，指出其隐藏状态形成了结构化子空间，并沿着平滑且可解释的轨迹演变 [5, 37]。然而，这些研究主要仍停留在诊断层面：它们刻画了隐藏状态现象，却未能在严格的单次遍历约束下，将轨迹级结构转化为实用的端到端置信度估计器。

本文的定位。 我们的方法建立在先前基于隐藏状态和拓扑结构的分析之上，但在目标上有显著差异。我们并非仅将几何或拓扑结构用于事后解读，而是将轨迹结构作为置信度信号加以操作化： 结构置信度从最后一层轨迹中提取一组紧凑的谱稳定性、局部变化和形状一致性描述符，随后将其输入轻量级监督式估计器。与现有基线方法相比，该方法具有诸多优势。我们将其部署为模型无关、可直接插入的评分层：(i) 仅依赖单次确定性前向传播即可获得的特征；(ii) 不要求访问梯度、注意力图或辅助模型。这项工作首次系统地将隐藏状态的结构稳定性操作化为一种统一、跨领域通用的事后置信度模态，并在多种异质事实性任务上对其与当前最先进方法进行了严格评估。

3 问题定义

大语言模型的置信度估计旨在判断生成的答案是否可能正确。形式上，给定输入 x 以及大语言模型单次遍历的自回归输出 $y = (y_1, \ldots\ldots, y_T)$ 和末层隐藏状态 $h_1, \ldots\ldots, h_T$，目标是估计

$$c(x, y) \approx \Pr(\text{correct}(y) \mid x, y, h_{1:T}).$$

在这项工作中，我们聚焦于一种计算约束严格的部署场景。在这种场景下，置信度估计器必须：(i) 仅依赖大语言模型的一次确定性前向传播；(ii) 不使用自然语言推理重排序器等辅助特定任务模型；(iii) 不执行 SelfCheckGPT式采样等多样本解码。这些约束反映了网络规模的大语言模型部署对延迟、成本和可扩展性的要求。现有置信度估计器主要分为三种范式，但每种范式都与这一预算要求相冲突。基于概率的启发式方法（词元对数概率、预测熵）符合单次遍历约束，但众所周知，它们的校准不佳，且在分布偏移下十分脆弱 [12]。语义嵌入方法根据表示相似度估计置信度，但对于域外（OOD）输入，其几何结构往往会发生坍塌，因为嵌入会偏离预训练流形 [33]。基于采样的一致性方法通过聚合多个样本来提升鲁棒性 [25], , 但会使计算量和延迟成倍增加，因此明确不符合单次遍历部署设计。

我们认为，这些共同的局限源于现有方法要么在概率空间中运作，要么在语义空间中运作，而这两者都处于表层，容易受到词汇或分布不匹配的影响。相比之下，不确定性本质上源自大语言模型的内部计算动态。当模型具有置信度时，其隐藏状态轨迹往往沿着连贯的方向平滑演化；当模型不确定时，轨迹则会表现出局部不规则性、高频振荡和全局形状不一致性。这些结构特征是模型计算固有的，与概率或静态嵌入相比，在经验上跨领域更加稳定。

因此，我们为置信度估计定义第三种模态：

隐藏状态轨迹的结构稳定性。

令 $\tau = (h_1, \ldots, h_T)$ 表示最后一层轨迹。我们要寻找一个函数

$$c_{\text{struct}}(x, y) = g(f(\tau)),$$

其中，f 提取多尺度结构描述符（局部变化、谱稳定性和形状一致性），而 g 则将它们映射为一个标量置信度分数。这一表述将置信度估计重新定义为一个基于大语言模型内部动态的几何特性与稳定性的结构推断问题，并直接引出第4节介绍的结构置信度框架。

4 方法

结构置信度是一种单次遍历、与生成器无关的置信度估计器，专为在线大语言模型（LLM）服务而设计。其核心思想是从由冻结编码器计算出的代理隐藏状态轨迹中提取结构稳定性信号，而无需获取生产环境中的大语言模型的逻辑值、采样结果或内部激活。此设计明确以实际部署约束为依据：现代 Web API（例如 GPT-4o）仅提供文本输出，实施严格的速率限制，且不允许访问词元级概率或隐藏状态。因此，任何事后置信度估计器都必须在单次遍历、黑盒且计算受限的环境下运行。结构置信度正是围绕这些约束构建，旨在提供一种稳定、高效且具有良好泛化能力的机制，用于估计不同 Web 领域中的正确性。

4.1 代理隐藏状态轨迹提取

由于现代 Web 大语言模型 API 不会公开内部隐藏状态，我们使用一个冻结的仅编码器模型构造一条代理轨迹。给定 $(x, \hat{y})$，我们使用固定分隔符将二者拼接为 $t = \text{concat}(x, \hat{y})$，并将 t 输入一个由 bert-base-uncased [4]实例化、隐藏维度为 768 的 12层Transformer编码器。

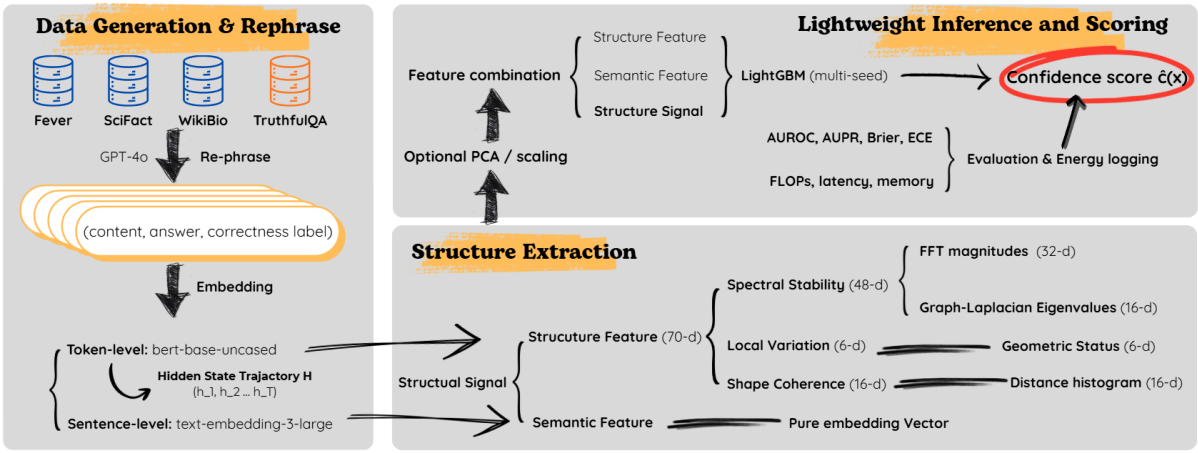


图1：整体结构置信度流程。大语言模型生成一个确定性的答案；编码器将（上下文、答案）映射为隐藏状态轨迹，并从中提取多尺度结构描述符，再由轻量级置信度模型进行评分。

最终层的隐藏状态序列为

$$H = (h_1, \ldots, h_T) \in \mathbb{R}^{T\times 768},$$

其中 T 被截断或填充至最多256个词元。

分词器鲁棒性。 尽管BERT和GPT-4o使用不同的分词器，结构描述符依赖于激活几何，而非词汇对齐。从经验上看，分词方式的微小变化不会改变 H中编码的多尺度模式，而 $T = 256$ 的固定上限则使描述符计算标准化。

在完整的结构信号变体（结构特征 +语义特征）中，我们还从预训练编码器获得句子嵌入 $s \in \mathbb{R}^{d_s}$。编码器均不应用梯度，从而确保严格的生成器无关行为。

4.2 结构描述符族

完整的结构信号由两个部分组成：(i) 从隐藏状态轨迹中提取的结构特征向量，以及 (ii) 由冻结的预训练编码器（OpenAI text-embedding-3-large）生成的固定句子嵌入 $s \in \mathbb{R}^{d_s}$，即语义特征向量。语义特征充当粗粒度的语义锚点，并与结构描述符拼接。

因此，本节仅讨论结构特征部分。给定轨迹 $H = (h_1, \ldots, h_T)$，我们提取三类互补的结构信号。每一类信号都与长度无关：无论 T 如何，都能生成固定大小的描述符，从而支持在不同 Web 输入之间进行一致的推理。

4.2.1 谱稳定性（*48* 维）。该类信号概括隐藏状态轨迹的平滑演化程度。它由两个部分组成：

频率领域平滑度（*32* 维）。我们沿着 H的词元轴应用实值离散傅里叶变换，并保留最低的 $K = 16$ 个非平凡频率的平均功率和最大功率，从而得到 $2K = 32$ 个特征。在进行离散傅里叶变换之前，先在词元空间中进行零填充；由于只保留低频带，填充不会影响保留的谱能量。

图谱扩散（*16* 维）。我们在词元上构建相似图，并计算归一化拉普拉斯矩阵 L 的最小 16 个特征值。这些特征值刻画了轨迹上的全局平滑性和扩散特性。当 T上限设为 256 时，计算最低阶特征值仍然易于处理，而且计算量显著低于多样本生成。

4.2.2 局部变化（*6* 维）。局部变化捕捉隐藏状态轨迹中的短程不稳定性。对于每一对连续词元，我们计算位移 $\Delta_t = \|h_t - h_{t-1}\|_2$，并使用六项统计量总结轨迹：总路径长度、$\Delta_t$ 的均值和方差、起终点距离、逐嵌入方差以及质心范数。这些描述符量化了全局结构度量可能忽略的细粒度波动。

4.2.3 形状一致性（*16* 维）。全局离散度通过将所有成对距离 $\|h_i - h_j\|_2$ 聚合为一个包含 16 个分箱的归一化直方图进行总结，从而生成与长度无关的“形状特征”，反映轨迹的整体一致性。

4.3 统一结构表示与粒度

设 $f_{\text{spec}}(H) \in \mathbb{R}^{48}$、$f_{\text{loc}}(H) \in \mathbb{R}^{6}$，以及 $f_{\text{shape}}(H) \in \mathbb{R}^{16}$。结构置信度将它们拼接为：

$$u(H) = \left[f_{\text{spec}}(H) \,\|\, f_{\text{loc}}(H) \,\|\, f_{\text{shape}}(H)\right] \in \mathbb{R}^{70}.$$

为控制结构分辨率，我们采用三种粒度模式：

全局。 在完整轨迹上计算的所有特征。

局部。 特征在重叠窗口（$w = 5$，步长 $= 2$）上计算，并在各窗口间取平均。较短的窗口能够捕捉全局描述符可能平滑掉的词元级不稳定性。

双尺度。 全局描述符与平均后的局部描述符按元素取平均。尽管这些描述符在性质上有所不同（谱、几何、基于直方图的），但其归一化确保了尺度可比，而取平均则起到类似于基于小波的表示的多分辨率平滑作用。

除非另有说明，否则默认使用双尺度模式。全局独立变体和仅局部变体将在消融实验中进行评估。

4.4 置信度估计器与训练目标

我们训练一个梯度提升决策树模型（LightGBM），将 $u(H)$（仅结构）或 $[u(H) \,\|\, s]$ (Struct+Sent) 映射为正确性概率。模型采用二元逻辑回归目标。树的深度较浅、数量较少，以适应低维输入，并满足快速推理的需求。超参数（学习率、叶节点数、特征比例、最大深度）按数据集固定，并在第 5 节中说明。基于树的模型能够稳健地处理描述符族之间异质的特征尺度和非线性交互。

4.5 计算方面的考量

每个实例的端到端成本为：

$$\mathrm{Cost}(T) = \mathrm{Cost}_{\mathrm{enc}}(T) + O(T \log T + T^2),$$

其中，$\mathrm{Cost}_{\mathrm{enc}}(T)$ 为编码器前向传播。将 T 限制在 256 以内后，$O(T \log T)$ 快速傅里叶变换特征以及 $O(T^2)$ 形状一致性和图谱分量均可在中央处理器上高效计算。实际应用中（第 6.4 节），与基于采样的方法（例如 SelfCheckGPT 的 10 轮采样）相比，结构置信度的延迟和浮点运算次数显著更低。此外，所有计算都是确定性的，无需反复调用生成器，因此该方法适用于实时网页推理。

4.6 为什么代理编码器已经足够

由于 GPT-4o 等专有大语言模型不公开隐藏状态，结构置信度依赖轻量级的仅编码器变换器来获取词元级轨迹。重要的是，我们不 假设代理编码器能够复现生成器的内部激活。相反，我们采用一种以输入为条件的激活几何视角：许多与稳定性和不确定性相关的模式，是由上下文—答案序列的结构以及自回归解码过程所诱发的，而非由模型特定参数决定。

支持这一观点的证据来自先前的研究：几何和结构特征——例如低频谱集中、局部曲率降低以及连贯的流形组织——出现在不同的 Transformer 系列中 [13, 28]。因此，这些模式至少在一定程度上是模型无关的。

我们的实证结果进一步支持了代理编码器的充分性。从 BERT 轨迹计算得到的同一组结构描述符，在 FEVER、SciFact 和 WikiBio-LLM 上始终能够区分正确与错误输出，并能平滑迁移到域外的 TruthfulQA 基准测试。性能是逐渐下降而非灾难性崩溃，这表明代理轨迹保留了与正确性相关的、由文本诱导的稳定性信号。

因此，尽管代理编码器并不能逐层逼近 GPT-4o，但它提供了一条稳定且计算量轻量的替代轨迹，其几何与谱结构保留了事后置信度估计所需的、对正确性敏感的模式。与真实生成器轨迹的直接比较留待未来工作。

4.7 理论动机：为何结构信号能够预测置信度

我们认为，置信度体现在自回归隐藏状态动态的稳定性上：模型有把握时，轨迹平滑且方差较低；而在不确定或产生幻觉时，轨迹会变得不规则、曲率较高，且结构支离破碎。这种行为与 Transformer 解码器在自回归预测过程中更新其隐藏状态的方式一致。

在每个词元步，模型通过 $h_t = f(h_{t-1}$，注意力$_t$，逻辑值$_t)$更新其表征。有把握的预测通常呈现出尖峰状的下一个词元分布和稳定的注意力模式，因此更新会在 t间平滑变化，形成具有以下特征的轨迹：(i) 低频谱能量，(ii) 幅度较小且均匀的增量变化，以及 (iii) 连贯的全局结构。

不确定或幻觉式的预测会破坏这种稳定性。当模型面对模糊证据或相互冲突的上下文线索时，下一词元分布往往变得不那么尖锐（熵更高），并在各步骤之间更加多变，从而放大注意力权重和多层感知机通路的敏感性。由此产生的更新函数在局部变得混沌：逻辑值或注意力的微小变化会导致 h_t 发生不成比例的大幅变化。这会引发 (i) 高频谱分量，(ii) 不规则的局部曲率与步长方差，以及 (iii) 全局轨迹中的碎片化与拓扑噪声。

这些效应已在各类 Transformer 系列中得到实证观察。谱不规则性与不确定性驱动的激活漂移相关 [31], 局部几何不稳定性与不一致的词元级决策边界相一致 [26], 而拓扑连贯性的丧失则与不可靠的内部推理通路相关联 [7]。重要的是，这些现象源于自回归解码的动态过程，而非模型特定参数，因此它们在不同代理编码器和领域中保持稳定。

因此，结构置信度将三个可测量的类别——谱稳定性、局部变化和全局形状连贯性——具体化为轨迹稳定性的模型无关指标：稳定的轨迹意味着较高的置信度，而不稳定的轨迹则表明存在表征不确定性，并可能出现错误。

5 实验设置

我们设计实验来回答三个问题：（Q1）在统一的单次传递协议下，结构置信度信号的竞争力如何？（Q2）这些信号在领域偏移和混合领域训练下有多稳健？（Q3）哪些结构设计选择——信号族、轨迹粒度或语义增强——是实现强置信度估计所必需的？本节介绍数据集、大语言模型生成协议、代理编码器提取、置信度估计器和评估指标。实现严格遵循我们发布的代码库，所有结果均使用由第 4 节和第 6.3 节中的脚本生成的缓存特征文件。

5.1 评估任务

我们在两种设置下评估置信度估计。**（1）句子级预测**：给定一个由输入声明/问题 x 和 GPT-4o 生成的答案 a组成的实例（x， a），置信度估计器会输出一个标量分数，用于预测 a 在事实上是否正确。**（2）段落级排序**：对于来自同一基准测试的一组生成结果，我们评估置信度分数能否将正确输出排在错误输出之前。这种排序视角与近期幻觉和真实性研究中使用的 AUC 式事实性评估完全相同。

5.2 数据集

我们遵循第4节中的 LLM-ese 构建方法。FEVER [36] 和 SciFact [38] 包含与二元支持/反驳标签配对的主张。WikiBio [19] 提供信息框条件生成的传记，并标注其事实正确或错误。我们通过对每个实例提示 GPT-4o 一次，将这三个基准测试都转换为单次评估集，并为每个实例对应一个二元正确性标签。移除格式错误的输出后，评估集分别包含 1,893 个（FEVER）、809 个（SciFact）和 780 个（WikiBio）实例。

对于**跨领域评估**，我们将 LLM-ese 训练集和开发集划分合并为混合领域训练池（mix_train，5,455 个实例；mix_dev，778 个实例）。mix_test 划分（3,131 个实例）包含三个数据集各自留出的部分，以及整个 TruthfulQA 基准测试 [21]，，后者未参与训练。

5.3 大语言模型生成协议

遵循 SelfCheckGPT 的单次处理设计，每个实例仅使用统一的指令模板查询一次 GPT-4o。所有输出均通过贪心解码确定性生成（temperature = 0，top-k = 1）。不使用采样、重新采样或改述。所有置信度估计器都严格基于这些固定输出运行。

5.4 代理编码器与结构特征提取

由于 GPT-4o 不会公开隐藏状态，我们采用第4节所述的轻量级代理编码器。使用 bert-base-uncased 对拼接后的上下文—答案序列进行编码；通过单次前向传播提取末层隐藏状态（h_1, ……， h_T），最大长度为 256 个子词标记。结构描述符按照提取脚本中的实现精确计算：(i) 局部变化 （几何稳定性）、(ii) 谱稳定性 （快速傅里叶变换和图拉普拉斯谱），以及 (iii) 形状一致性 （距离直方图和拓扑数据分析摘要）。拼接后，每个实例得到一个固定的 70 维向量。所有数据集均使用相同的配置。

5.5 基线方法

我们将 Structural Confidence 与四类事后置信度分数进行比较。**（1）基于概率的启发式方法**：使用开放权重模型，根据词元级对数概率计算平均 logp（Mean logp； [11]）和预测熵（logp_系列），因为在我们的黑盒 API 设定下，GPT-4o 不公开词元级对数概率或逻辑值。我们通过对开放权重模型进行教师强制，并提取每个词元的对数概率，在同一个确定性的 GPT-4o 答案上计算这些分数。**（2）语义基线**：采用一种语义特征变体，使用平均池化的 BERT 表征和同一个 LightGBM 分类器，同时根据同一个确定性的 GPT-4o 答案中的蕴含概率，计算语义熵（NLI，单次遍历）分数；不同于原始形式 [6],，该实例化严格运行在单次遍历的黑盒设定下，不进行额外生成，也不使用辅助问答模型。**（3）多样本一致性**：基于 NLI 的 SelfCheckGPT 基线 [25] ，使用官方的"SelfCheckNLI"模型。我们对各段落的矛盾概率取平均，并将其反转以获得正确性分数。我们不纳入更为繁重的多样本推理方法（例如 Stable Explanations [2]），因为这些方法每个查询需要生成 10–50 个样本，并使用外部编码器。**（4）基于几何的离群点基线**：一种 k均值离群点检测器（kmeans_离群点），根据每个实例在表示空间中与最近质心的距离进行评分。该基线采用标准的基于聚类的离群点检测启发式方法；对此，我们引用了近期关于基于聚类的离群点检测方法的比较评估 [34]。我们将比较范围限制在此类单次遍历的黑盒基线，不纳入需要访问生成器隐藏层的白盒内部状态方法。此外，我们在效率比较中加入了 RACE 风格的结构基线，该基线按照原始设计 [39]实例化为一个包含五个模块的代理轨迹流程；该基线使用相同的代理编码器，但以更深层、多阶段的结构分析替代了我们的轻量级描述符。

5.6 监督式训练方案

监督式估计器（结构特征、语义特征、结构信号）使用 LightGBM 和二元逻辑损失。训练严格遵循我们的脚本：200 棵树、学习率 0.05、深度不设上限，并使用固定随机种子以确保可复现性。不进行针对特定数据集的调参，也不使用验证启发式方法。单领域模型使用各数据集的 LLM-ese 训练集/开发集划分进行训练；跨领域模型使用 mix_train 进行训练，并在 mix_dev 上验证。测试集始终严格保持未见状态。

5.7 评估指标

我们按照大语言模型真实性研究中的标准做法 [12, 22, 25]，报告区分能力（AUROC、AUPR）和校准情况（布里尔分数、期望校准误差）。所有指标均基于各方法生成的连续标量置信度分数计算。

5.8 效率方案

我们仅使用经过测量的浮点运算次数和墙钟时间测量来衡量置信度估计阶段的计算成本。对于所有方法，共享的 GPT-4o 前向传播均不计入其中。对于 SelfCheckGPT 等多次遍历基线方法，我们同时报告单次遍历自然语言推理检查的成本，以及针对 S 个假设样本的外推成本（按照 [25] 中所述，随 S 线性缩放）。结构置信度仅使用一次 BERT-base 前向传播和中央处理器端特征提取。

硬件配置。所有实验均在一台工作站上进行，该工作站配备 AMD Ryzen 9 5900HX 中央处理器、32GB 内存，以及一块 NVIDIA GeForce RTX 3080 笔记本电脑 GPU，

显存为 16GB。除非另有说明，所有置信度估计方法均采用此硬件配置。我们仅测量置信度估计阶段的浮点运算次数和实际运行延迟，不包括共享的 GPT-4o 生成成本。

6 结果与分析

我们从四个方面评估结构置信度：句子级预测的主要结果、跨领域鲁棒性、关键设计选择的消融研究，以及效率分析。

6.1 句子级置信度预测

表1比较了 FEVER、SciFact 和 WikiBio 上的所有置信度估计器。我们重点关注三个问题：结构信号与基于概率的启发式方法相比如何、相较于语义基线表现如何，以及融合配置与 SelfCheckGPT 等强基准方法相比表现如何。

结构信号与基于概率的启发式方法。平均对数概率和熵在 FEVER 和 WikiBio 上表现不佳，AUROC 往往接近或低于随机猜测水平。相比之下，结构特征 无需使用词元概率，便能在全部三个数据集上提供稳定的排序信号。在 FEVER 和 WikiBio 上，结构特征 在 AUROC 和 AUPR 上均优于两种基于概率的启发式方法，这支持了这样的观点：轨迹级稳定性能够捕捉简单似然无法提供的正确性信息。在 SciFact 上，平均对数概率表现具有竞争力，但所有基于概率的方法仍然脆弱，区分能力也仅属中等。

结构信号与语义表征。语义特征 在 FEVER 和 WikiBio 上表现良好（AUROC 为 0.682/0.683），但在 SciFact 上则表现骤降（AUROC 为 0.255）；这与先前的观察一致：句子嵌入在领域和风格偏移下会退化 [33]。结构信号则是平缓退化，而非灾难性崩溃：结构特征 在 FEVER 和 WikiBio 上仍具有明显的区分能力，在 SciFact 上也保留了不可忽视的信号（AUROC 为 0.460）；当语义编码器失效时，它能提供更审慎的置信度估计。在无法控制领域和风格的 Web 场景中，这种表现模式更为理想。

融合配置与强参考基线方法的比较。融合配置的结构信号融合了两种来源的优势，在 FEVER 和 WikiBio 上，其 AUROC/AUPR 达到单次遍历基线方法中的最高或接近最高水平，同时避免了 SciFact 上的嵌入崩溃。由于所有监督式变体使用相同的分类器和超参数，性能提升体现的是互补信息，而非模型容量增加。与 SelfCheckGPT（自然语言推理）相比，后者在 FEVER 上取得最高 AUROC（0.700），但计算开销显著更高（第6.4节）；Struct+Sent 在 FEVER 上以微小的准确率差距换取大幅降低的计算开销，同时在 WikiBio 上的 AUPR 与 SelfCheckGPT 持平或更优。因此，对于无法进行重复采样或多次自然语言推理的应用，结构置信度提供了一种有竞争力的单次处理替代方案。

区分能力与校准。除了 AUROC 和 AUPR，校准指标（布里尔分数和 ECE）也呈现相同趋势：在 FEVER 和 WikiBio 上，结构描述符和融合模型 Struct+Sent 的校准效果始终优于仅概率基线，而在 SciFact 上则与其相当。这表明结构稳定性不仅能改善排序，还能使置信度分数更准确地反映经验正确率。

表 1：FEVER、SciFact 和 WikiBio 上的 AUROC / AUPR。粗体条目表示融合的结构信号配置。

方法	FEVER		SciFact		WikiBio	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
谱稳定性	0.555	0.268	0.475	**0.725**	0.504	0.309
形状一致性	0.552	0.263	0.522	0.701	0.540	0.295
局部变化	0.534	0.247	0.515	0.689	0.565	0.320
结构特征	0.573	0.274	0.460	0.670	0.543	0.354
语义特征	0.682	0.349	0.255	0.602	0.683	0.430
结构信号	**0.683**	**0.348**	0.335	0.640	**0.686**	**0.454**
平均 logp [11]	0.438	0.208	0.505	0.677	0.485	0.295
语义熵（自然语言推理，单次遍历） [6]	0.432	0.201	0.512	0.684	0.453	0.264
KMeans 离群点 [34]	0.518	0.234	0.544	0.696	0.594	0.355
SelfCheckGPT [25]	0.700	0.364	0.458	0.650	0.564	0.312

6.2 跨域鲁棒性

现在，我们考察结构和语义置信度信号在领域偏移下的表现。我们在混合领域数据集_训练池上训练单个模型，并在 FEVER-LLM、SciFact-LLM、WikiBio-LLM 以及域外 TruthfulQA 上进行评估。图 2 绘制了 仅结构和语义特征在各领域上的 AUROC。

结构信号能够平稳地泛化。结构特征在 FEVER-LLM、SciFact-LLM、WikiBio-LLM 和 TruthfulQA 上均表现稳定，证实结构稳定性能够捕捉隐藏状态动态中不受领域影响的规律。尽管从训练领域迁移到 TruthfulQA 时性能略有下降，但模型依然稳健，避免了跨领域偏移中常见的性能崩溃。

语义信号波动明显。语义特征因领域而异，在 SciFact-LLM 和 TruthfulQA 上 AUROC 大幅下降。这与单领域实验结果一致，凸显了句子嵌入在风格和主题发生偏移时的脆弱性，尤其是在训练领域与测试领域不一致时。

互补融合。结合使用时，结构信号往往能在每个领域中跟随结构信号或语义信号中表现较强的一方，这表明，当语义编码器未能对齐时，结构稳定性可提供稳健的骨干；而当语义可靠时，语义则能在领域内发挥强大的优化作用。对于 Web 规模部署而言，输入涵盖异质领域，因此这种组合比单独依赖嵌入更为安全。

6.3 消融研究

本节研究消融研究中的主要结构设计选择，包括信号族、语义增强、轨迹粒度和描述符维度。

结构族具有互补性。表 2 将完整的结构描述符分解为三个结构族。每个结构族在至少一个数据集上表现最佳：谱稳定性在 FEVER 上最强，形状一致性在 SciFact 上最强，局部变化在 WikiBio 上最强。

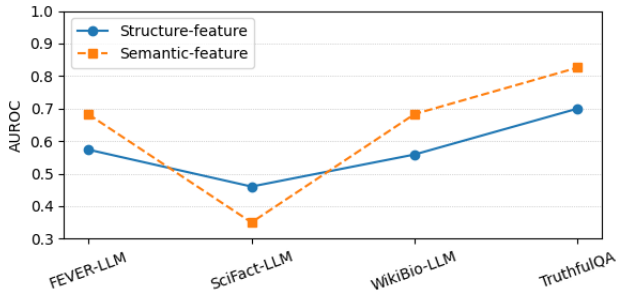


图 2：在 mix_train 上训练并在四个领域上评估的结构特征和语义特征 的跨域 AUROC。

表 2：结构族和语义融合的消融 AUROC。（局部：局部变化，谱：谱稳定性，形状：形状一致性，结构：结构特征，语义：语义特征）

数据集	局部	谱	形状	结构	Sem	**结构信号**
FEVER	0.538	0.579	0.553	0.574	0.682	0.683
SciFact	0.455	0.449	0.499	0.460	0.350	0.347
WikiBio	0.570	0.532	0.520	0.559	0.683	0.685

拼接后的结构特征变体在每个数据集上的表现都接近最佳单一结构族，同时在所有数据集上更加稳定，这表明这三个结构族捕捉了轨迹稳定性的互补方面。

语义增强：何时有帮助？在 FEVER 和 WikiBio 上，语义特征是最强的单一来源信号，而简单拼接（结构信号）带来了虽小但稳定的提升，这表明结构描述符能够进一步完善本已强大的语义基线。然而，在 SciFact 上，结构特征明显优于语义特征，而朴素融合并无帮助，这说明领域偏移下语义信息并非总是有益，必须谨慎地加以结合。

轨迹粒度与维度。改变轨迹粒度（全局、局部、双尺度）后发现，对于较长且噪声较多的主张（FEVER、混合领域），纯局部窗口并不足够，全局和双尺度变体能带来更好的 AUROC。对于主张较短且模板化程度较高的 SciFact，局部结构更有优势。将维度从 70 压缩至 32/16/8 的 PCA压缩，在 FEVER 上仅引起轻微的 AUROC 变化，在 SciFact 上甚至带来小幅提升，这表明该描述符存在冗余，可以存储在一个紧致子空间中。

6.4 效率分析

最后，我们比较结构置信度与更强但成本更高的基线方法在计算量和延迟成本方面的表现。表 3 报告了相对浮点运算次数和延迟，并将单次遍历结构信号的成本标准化为 1。0×近期有关资源受限条件下大语言模型部署的综述充分阐述了最大限度降低推理时计算量的重要性

表 3：置信度估计器（结构置信度 = 1.0）的相对浮点运算次数和延迟。

方法	浮点运算次数	延迟
结构置信度	1.0	1.0
RACE基线	4.0	3.0
SelfCheckGPT（自然语言推理）	6.0	5.0

[10]。所有数字仅反映置信度估计阶段。

单次遍历结构置信度成本最低。结构置信度只需一次BERT-base前向传播，外加轻量的中央处理器侧特征提取与LightGBM推理。在超越原始对数概率和单向量句子嵌入的方法中，该方法实现了最低的浮点运算次数和延迟。

*RACE*式与*SelfCheckGPT*基线方法的成本要高得多。相对于结构信号（结构特征 + 语义特征），RACE式结构流水线大约产生 4×浮点运算次数和 3× 延迟，而SelfCheckGPT（自然语言推理）由于额外的自然语言推理模型和更长的处理流水线，大约产生6× 浮点运算次数和 5× 延迟。在多样本SelfCheckGPT配置中，这些成本会随样本数量线性增长。相比之下，结构置信度始终严格保持单次遍历且基于代理。

7 结论

本文提出了结构置信度，一种面向大语言模型事后置信度估计的单次遍历、模型无关框架。通过分析隐藏状态轨迹的几何与谱结构，并在消融实验变体中结合其拓扑摘要，该方法导出有别于传统方法的基于稳定性的置信度信号，即词元似然、句子级语义和多样本一致性分数，且仅需一次编码器前向传播。以FEVER、SciFact和WikiBio-hallucination作为基准测试，结构置信度在领域偏移下仍保持可靠，优于基于概率的基线，并相对于句子级语义嵌入提供互补增益。在混合领域设置中，模型于FEVER、SciFact与WikiBio-LLM的并集上训练，并在这三个测试集以及域外TruthfulQA基准测试上评估；同一结构模型无需数据集特定调优即可保持有竞争力的AUROC和AUPR，融合的Struct+Sent变体在置信度估计阶段相较于基于自然语言推理的SelfCheckGPT基线，提供了更有利的准确率与效率的权衡。本研究仅限于一小部分事实性与真实性基准、一个专有基础模型，以及特定硬件配置上基于浮点运算次数与延迟的效率估计。未来工作将把结构置信度建模扩展到其他网络任务、开放权重模型与多模态模型，并将其集成到事实核查、检索增强辅助和内容审核等端到端系统中。

参考文献

[1] Rishi Bommasani. 2021. 关于基础模型的机遇与风险。*arXiv* 预印本 *arXiv:2108.07258* (2021)。

[2] Chao Chen, Chenghua Guo, Rufeng Chen, Guixiang Ma, Ming Zeng, Xiangwen Liao, Xi Zhang, and Sihong Xie. 2024. 训练以实现稳定且免费的解释。神经信息处理系统进展 37 (2024), 3421–3457. [3] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. 2005. 持久图的稳定性。载于第二十一届计算几何年度研讨会论文集。263–271. [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT：用于语言理解的深度双向变换器预训练。载于2019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（长篇与短篇论文）。4171–4186. [5] Kawin Ethayarajh. 2019. 上下文化词表示的上下文性如何？BERT、ELMo 和 GPT-2嵌入的几何比较 2 (2019). [6] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. 使用语义熵检测大语言模型中的幻觉。《自然》630, 8017 2024 625 630 ( ), – . [7] Stephen Fitz, Peter Romero, and Jiyan Jonas Schneider. 2024. 隐藏的孔洞：语言模型的拓扑学方面。arXiv 预印本 arXiv:2406.05798 (2024). [8] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2019. 深度集成：损失景观视角。arXiv 预印本 arXiv:1912.02757 (2019). [9] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. 变换器前馈层是键值记忆。载于2021年自然语言处理实证方法会议论文集。5484–5495.[10] Sanjay Surendranath Girija, Lakshit Arora, Shashank Kapoor, Dipen Pradhan, Aman Raj, and Ankit Shetgaonkar. 2025. 面向资源受限环境优化大语言模型：模型压缩技术综述。载于2025年IEEE第49届计算机、软件与应用会议（COMPSAC）。IEEE, 1657–1664.[11] Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023. 大海捞针：神经机器翻译中幻觉的全面研究。载于第17届欧洲计算语言学协会分会会议论文集。1059–1075. [12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. 现代神经网络的校准。载于国际机器学习会议。PMLR，1321–1330.[13] John Hewitt and Christopher D Manning. 2019. 用于在词表示中发现句法的结构探针。载于2019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（长篇与短篇论文）。4129–4138.[14] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. 自然语言生成中的幻觉综述。ACM计算调查 55, 12 (2023), 1–38. [15] Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen. 2024. 从大语言模型到基于大语言模型的智能体：软件工程现状、挑战与未来综述。arXiv 预印本 arXiv:2408.02479 (2024).[16] Saurav Kadavath, Tom Conerly, A manda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran- Johnson, et al. 2022. 语言模型（大多）知道自己知道什么。arXiv 预印本 arXiv:2207.05221 (2022). [17] Amita Kamath, Robin Jia, and Percy Liang. 2020. 领域偏移下的选择性问答。arXiv 预印本 arXiv:2006.09462 (2020). [18] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. 揭示 BERT 的黑暗秘密。arXiv 预印本 arXiv:1908.08593 (2019). [19] Rémi Lebret, David Grangier, and Michael Auli. 2016. 基于结构化数据的神经文本生成：应用于人物传记领域。arXiv 预印本 arXiv:1603.07771 (2016). [20] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michi- hiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Ku mar, et al. 2022. 语言模型的全面评估。arXiv 预印本 arXiv:2211.09110 (2022). [21] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA：测量模型如何模仿人类谬误。载于计算语言学协会第60届年会论文集（第1卷：长篇论文）。3214–3252.[22] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. 2022. 面向自由形式文本生成的词元级无参考幻觉检测基准。载于计算语言学协会第60届年会论文集（第1卷：长篇论文）。6723–6737.[23] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Mar- ianna Apidianaki, and Chris Callison-Burch. 2023. 忠实的思维链推理。载于第13届国际自然语言处理联合会议暨第3届计算语言学协会亚太分会会议（IJCNLP-AACL 2023）。[24] Andrey Malinin and Mark Gales. 2018. 通过 Prior Networks 进行预测不确定性估计。神经信息处理系统进展 31 (2018). [25] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT：用于生成式大语言模型的零资源黑盒幻觉检测。载于*2023*年自然语言处理实证方法会议论文集。9004–9017。[26] R Tho mas McCoy、Junghyun Min 和 Tal Linzen。2020。志趣相投的 BERT 并不会共同泛化：测试集性能相似的模型之间，泛化能力存在显著差异。载于第三届用于分析和解读自然语言处理神经网络的 *BlackboxNLP* 研讨会论文集。217–227。[27] Paul Michel、Omer Levy 和 Graham Neubig。2019。十六个注意力头真的比一个好吗？神经信息处理系统进展 32（2019）。[28] John Morris、Volodymyr Kuleshov、Vitaly Shmatikov 和 Alexander M Rush。2023。文本嵌入所揭示的信息几乎与文本本身一样多。载于*2023*年自然语言处理实证方法会议论文集。12448–12460。[29] Marius Mosbach、Maksym Andriushchenko 和 Dietrich Klakow。2020。论 BERT 微调的稳定性：误解、解释与强有力的基线方法。*arXiv* 预印本 *arXiv:2006.04884*（2020）。[30] Yaniv Ovadia、Emily Fertig、Jie Ren、Zachary Nado、David Sculley、Sebastian Nowozin、Joshua Dillon、Balaji Lakshminarayanan 和 Jasper Snoek。2019。你能相信模型的不确定性吗？评估数据集偏移下的预测不确定性。神经信息处理系统进展 32（2019）。[31] Nasim Rahaman、Aristide Baratin、Devansh Arpit、Felix Draxler、Min Lin、Fred Hamprecht、Yoshua Bengio 和 Aaron Courville。2019。论神经网络的谱偏差。载于国际机器学习会议。PMLR，5301–5310。[32] Vipula Rawte、Amit Sheth 和 Amitava Das。2023。大型基础模型中幻觉问题综述。*arXiv* 预印本 *arXiv:2309.05922*（2023）。[33] Nils Reimers 和 Iryna Gurevych。2019。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。*arXiv* 预印本 *arXiv:1908.10084*（2019）。[34] Braulio V Sánchez Vinces、Erich Schubert、Arthur Zimek 和 Robson LF Cordeiro。2025。基于聚类的离群点检测：比较评估。数据挖掘与知识发现39，2（2025），13。[35] Ian Tenney、Dipanjan Das 和 Ellie Pavlick。2019。BERT 重新发现了经典的自然语言处理流程。*arXiv* 预印本 *arXiv:1905.05950*（2019）。[36] James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018。FEVER：用于事实抽取和验证的大规模数据集。*arXiv* 预印本 *arXiv:1803.05355*（2018）。[37] Lucrezia Valeriani、Diego Doimo、Francesca Cuturello、Alessandro Laio、Alessio Ansuini 和 Alberto Cazzaniga。2023。大型 Transformer 模型隐藏表示的几何结构。神经信息处理系统进展36（2023），51234–51252。[38] David Wadden、Shanchuan Lin、Kyle Lo、Lucy Lu Wang、Madeleine van Zuylen、Arman Cohan 和 Hannaneh Hajishirzi。2020。事实还是虚构：验证科学主张。*arXiv* 预印本 *arXiv:2004.14974*（2020）。[39] Changyue Wang、Weihang Su、Qingyao Ai 和 Yiqun Liu。2025。联合评估答案与推理一致性，以检测大型推理模型中的幻觉。*arXiv* 预印本 *arXiv:2506.04832*（2025）。[40] Daquan Zhou、Zhiding Yu、Enze Xie、Chaowei Xiao、Animashree Anandkumar、Jiashi Feng 和 Jose M Alvarez。2022。理解视觉 Transfor mer 的鲁棒性。载于国际机器学习会议。PMLR，27378–27394。

A 结构描述符的其他属性

本附录对第4节作补充，概述正文中介绍的结构描述符在鲁棒性、复杂度和可移植性方面的特性。我们着重讨论可解释性和复用，而不再赘述算法定义。

A.1 固定维度与长度无关的设计

所有描述符族（局部变化、谱稳定性和形状一致性）均设计为具有固定维度，不受生成长度 T 影响。局部变化将词元间的几何统计量汇总为少量标量。谱描述符将快速傅里叶变换（FFT）的幅值轮廓压缩为低频能量、主导分量和谱熵的摘要。形状一致性通过成对距离构建固定分箱直方图，并可选择结合少量拓扑统计量。

因此，在我们实验采用的全局设置中，最终的结构描述符具有固定维度，能够统一应用于长度差异很大的网络查询。

数据集	仅结构	PCA-32	PCA-16	PCA-8
FEVER	0.574	0.552	0.544	0.534
SciFact	0.460	0.446	0.412	0.395

表 4：结构描述符的PCA压缩影响（双尺度设置，AUROC）。

A.2 对激活与配置变化的鲁棒性

这些描述符直接作用于原始隐藏状态轨迹，无需显式进行逐样本归一化。实验表明，使用能量比、聚合的几何统计量和距离分布，能够增强对激活尺度和偏移适度变化的鲁棒性。

在所有数据集（FEVER、SciFact、WikiBio 和 TruthfulQA）中，结构特征和下游置信度分数在不同随机种子、数据划分及细微配置变化下均保持数值稳定。这一现象符合拓扑数据分析中的稳定性原理：汇总泛函（例如持久图）在底层空间发生扰动时具有可证明的稳定性 [3]。我们并未在此声称提出了新的保证，但同样的直觉也适用：关注全局几何结构的描述符对微观激活噪声不太敏感。

A.3 计算复杂度

与 Transformer 前向传播相比，这些描述符的计算开销很低。局部变化只需计算向量差，并对 T 个步骤进行简单归约。谱描述符对每条轨迹执行一次 $O(T \log T)$ 实数快速傅里叶变换，与模型推理相比可忽略不计。形状一致性采用向量化成对距离计算和固定分箱直方图计算；在窗口化配置中，计算仅限于固定大小的区块，因此对于长序列，其实际开销接近于关于 T 的线性复杂度。

因此，结构置信度的开销主要由单次大语言模型前向传播决定。结构计算仅增加一个很小的常数因子，第6.4节和附录C.2对此进行了实证量化。

A.4 跨模型与任务的可复用性

由于这些描述符仅依赖隐藏状态轨迹，无需词元标识、逻辑值或校准层，因此无需修改即可迁移到不同的仅解码器大语言模型。任何能够提供末层隐藏状态的模型都可以接入结构置信度流程，使这些描述符能够在事实核查、科学辅助或内容审核等网络任务中重复使用。

B 结构置信度的扩展可视化

为补充第 6.1 节和第 6.2 节中的数值结果，我们提供了根据主表所用的相同结果文件生成的更多可视化内容。

B.1 跨数据集AUROC概况

图3展示了结构变体与语义变体在FEVER、SciFact和WikiBio上的AUROC。表1报告的是绝对分数，而条形图则突出了各描述符族之间的相对差距，并说明了结构置信度的多尺度特性。

B.2 SciFact热图

图 4 展示了 SciFact 上相同变体的逐指标热力图（AUROC 和 AUPR），说明了第 6.2 节讨论的领域偏移效应。语义置信度在 FEVER 和 WikiBio 上表现出色，但在 SciFact 上大幅退化；相比之下，结构变体的信号较弱，但在各项指标上更稳定。

C 其他消融实验与鲁棒性检查

我们报告两项补充正文内容的额外消融实验：对结构描述符进行 PCA 压缩，以及对 FEVER 上的能量进行详细分析。

C.1 结构描述符的 PCA 压缩

为评估描述符中的冗余，我们对完整结构向量应用 PCA，并在训练分类器之前保留固定数量的主成分。表 4 报告了 FEVER 和 SciFact 上多个目标维度对应的 AUROC。

适度压缩（PCA-32）仅会使 FEVER 上的表现略有下降，这表明相当一部分信号集中在较低维子空间中。在所有方法都表现不佳的 SciFact 上，激进压缩会导致性能更明显地下降。因此，完整描述符提供了有益的安全余量，尤其是在更具挑战性的领域中。

C.2 FEVER上的详细能耗比较

正文报告了各数据集上的平均相对计算量，并以结构置信度为基准进行归一化。表5对此进行了补充，给出了FEVER上的单数据集视角，并以对每个输入执行多次随机生成的SelfCheckGPT式基线为基准进行归一化。

结果展示了多样本一致性与单次估计器之间的权衡。SelfCheckGPT风格的方法会带来显著的计算量开销，而结构置信度——采用单次贪心生成和轻量级梯度提升分类器——则将浮点运算次数和运行时间减少了一个数量级以上。这支持了以下观点：结构信号能够提供有用的置信度估计，且计算预算适合大规模 Web部署。

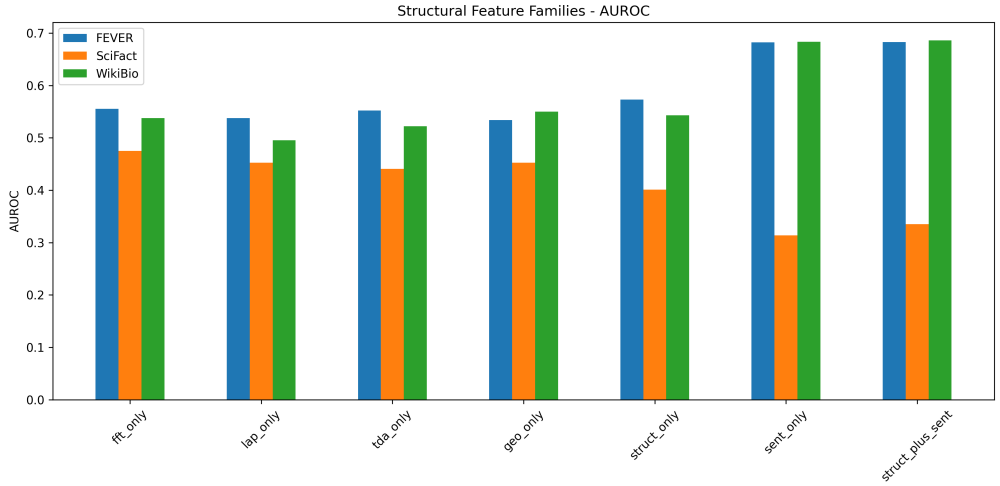


图 3：fft_only、lap_only、tda_only、geo_only、struct_only、sent_only 和 struct_plus_sent 在 FEVER、SciFact 和 WikiBio 上的 AUROC。统一结构描述符和融合后的 Struct+Sent 变体在各数据集上表现最为稳定。

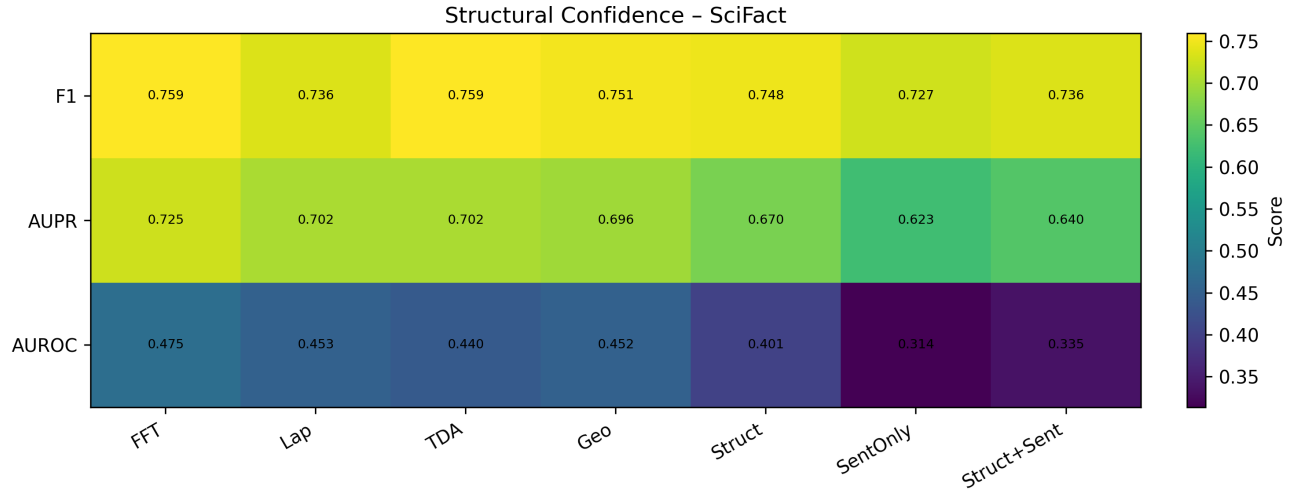


图 4：SciFact 上结构变体和语义变体各项指标（AUROC 和 AUPR）的热力图。与 FEVER 和 WikiBio 相比，语义置信度明显崩塌，而结构轨迹提供的信号较弱，但在各项指标上更为稳定。

方法	#生成次数	相对浮点运算次数	相对运行时间	相对内存
SelfCheckGPT风格基线	5	1.00	1.00	1.00
结构置信度（我们的方法）	1	0.03	0.04	0.07

表 5：FEVER 上的能耗比较。相对成本以每个输入使用 $S = 5$ 个随机生成结果的 SelfCheckGPT风格基线为基准进行归一化。所有方法都依赖于相同的底层大语言模型；差异仅来自推理时生成结果的数量和附加模块。