当意义保持不变，但模型漂移：在大型语言模型中评估服务质量时的令牌级行为不稳定性

Xiao Li[1], Joel Kreuzwieser[2], Alan Peters[1]
[1]Electrical and Computer Engineering, Vanderbilt University
[2] Radiology and Radiological Sciences, Vanderbilt University
Correspondence: { xiao.li, joel.p.kreuzwieser, alan.peters } @vanderbilt.edu

Abstract

我们研究大型语言模型如何应对在词级实现上有所不同但保留相同语义意图的提示，这一现象我们称之为提示变异。我们提出了基于提示的语义转移（PBSS），这是一个用于测量在语义等价提示重述下大型语言模型行为漂移的诊断框架。应用于十个受限任务后，PBSS 揭示了模型特定的响应转移——表明与分词和解码相关的统计规律。这些结果突出了在重新措辞下模型评估稳定性的一个被忽视的维度，并建议分词策略和解码动态可能导致训练后服务质量（QoS）不稳定。所有代码、提示和输出可在：我们的 github 链接 获得。

1 引言：提示等效，行为漂移

大型语言模型（LLMs）正在越来越多地被部署在实际环境中——从法律起草和客户服务到临床决策支持。随着这些系统进入高风险领域，对其一致性和可控性的担忧也增加了。一个关键但尚未充分探讨的问题是它们在轻微提示变动下的不稳定性——措辞上的细微变化可能导致输出结果的差异，这会影响到系统的可靠性和用户信任。

考虑一个假设的情景，其中一个商业大型语言模型在临床工作流程中提供帮助。当被提示时：

> 为初级医师简要解释此 MRI 扫描结果。

然后重新表述为：

> “你能不能像我是一名医学实习生一样带我过一遍这个扫描过程?”

同一个模型可能会对冲、投机，或添加原文中没有的警告说明。虽然任务保持不变，但在语调、重点，甚至事实框架上会出现微妙的变化。这些不一致引发了对信任和一致性的担忧，也对安全性产生了影响——尤其是在医疗、法律和代码生成等领域，在这些领域，轻微的措辞修改可能会导致实质上不同的输出。

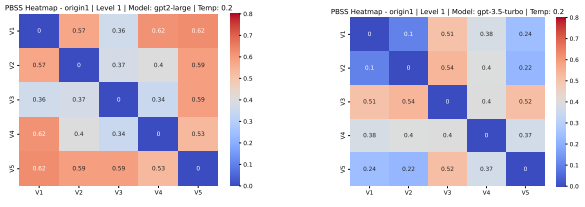


Figure 1: PBSS 热图显示五个语义等价提示的输出差异。红色 = 较少的语义相似性。详见第 5 节。

当被应用于高风险领域时，尤其是临床或法律领域——这种行为漂移不是生成过程中的无害产物。这成为一种潜在的失败模式。一个重新措辞的提示可能会改变诊断语气或事实框架，从而导致模型输出与现实世界产生后果的不一致。我们的框架旨在在这种不稳定性变得危险之前揭示出来。

这引出了一个更广泛的问题：语言模型对其指令的符号形式有多敏感，而不是内容？尽管通常被视为风格上的噪音，我们发现这些变化遵循系统的、可测量的模式——揭示了大语言模型以意想不到的方式内化表面线索的过程。

这项工作有三个贡献：

- 我们将基于提示的语义转移（PBSS）形式化为一种行为诊断框架，用于评估模型对指令中表面级变化的敏感性——通常体现在词汇级实现上。PBSS 揭示了模型系列之间的显著差异，通过使用三个独立的语义编码器进行的 Kruskal-Wallis 测试验证，当编码器组合使用时观察到最显著的效果。

- 我们对提示变异进行系统分析，将其视为一种结构化行为现象——这一角度很少在先前的工作中被涉及，而这些工作通常将输出变异视为偶然的或对抗性的。

- 我们将发布一个轻量级工具包，用于测量意译漂移—提供针对模型、温度和令牌级提示变体的可扩展诊断。据我们所知，这是首次使用语义嵌入模型作为 LLM 在意

义保留的重写下的稳定性行为诊断。本工作不仅识别出模型敏感性的可解释轴线，还旨在鼓励社区更系统地研究语言模型的潜在行为结构。

2 相关工作

最近的研究从多个角度考察了语言模型的行为。我们的工作通过关注释义变化下的语义漂移为这一不断增长的文献作出贡献，这一现象将行为不稳定性与输入实现中的词级差异联系在一起。我们回顾了几个方面的相关研究，这些研究对这种框架提供了信息。

2.1 提示鲁棒性 & 敏感性分析

提示语句的细微变化可能导致大型语言模型（LLM）行为的显著变化（Reynolds and McDonell, 2021; Webson and Pavlick, 2022; Zhou et al., 2024），即使模型输出看似准确。一些研究将提示语框架为潜在能力触发器（Reynolds and McDonell, 2021），而另一些则质疑 LLM 是否真正理解指令语义（Webson and Pavlick, 2022）。通过微调进行对齐可能增加受到注入攻击的风险（Li et al., 2023），而较大的模型，尽管更有能力，但往往变得更不可预测（Zhou et al., 2024）。已显示在分布性转变下，提示语也会加剧虚假的相关性（Setlur et al., 2024）。

这些发现表明，不仅是意图，提示形式也可以引发结构化和模型特定的行为差异。我们的工作基于这一见解，提出了 PBSS 作为一种诊断框架，用于量化表面层次的释义所导致的语义漂移。

2.2 指令微调 & 模型行为塑造

指令微调——在人工监督下对模型进行精心设计的提示进行微调——已经被证明可以在各种任务中提高模型的一致性、真实性和拒绝行为（Ouyang et al., 2022; Chen et al., 2024; Zhu et al., 2025; Bai et al., 2025）。技术范围从函数调用合成和拒绝建模到用于噪声环境的对抗性增强。

通过微调旨在通过再训练来塑造模型行为，而我们的 PBSS 框架提供了一种事后诊断：它不是抑制不稳定性，而是揭示了对语义等效提示的结构化敏感性——无需修改模型。

2.3 幻觉，事实漂移 & 输出一致性

事实不一致——通常称为幻觉——在总结、对话和问答中被广泛研究 (Maynez et al., 2020; Ji et al., 2023)。先前的工作提出了与人类判断一致的评估指标（Maynez et al., 2020），调查了特定领域的模式（Ji et al., 2023），并开发了从样本间一致性到基于归因的解码的检测策略（Manakul et al., 2023; Chen et al., 2025）。

我们将 PBSS 定位为一个上游诊断工具：如果一个模型在语义等价提示下无法保持一致性，它在下游保持事实内容的能力可能会受到影响。行为一致性已成为高风险大语言模型部署中的核心问题，研究突出提示表达、角色转变和越狱攻击带来的风险。最近的工作记录了对抗性提示策略、多轮次越狱以及在 RLHF 下挑战一致性的突现行为。

我们的工作采用了一种互补的视角：PBSS 不通过对抗性触发器来探测失败，而是量化在良性、语义等同的提示下的行为漂移——提供了一种对软对齐不稳定性的中性诊断视角。随着 LLMs 的规模扩大，越来越多无法预测的行为出现——这使得评估变得复杂，并引发了对超越准确性的诊断的兴趣。链式思维提示和新兴推理为能力提升提供了基准，但往往掩盖了结构和风格上的潜在不稳定性。

PBSS 通过追踪微观层面的行为漂移来补充这方面的工作：它不是强调新的能力，而是显露出在语义稳定环境中波动的早期信号—提供了一种轻量级的视角来观察新出现的不可预测性。

2.4 语义嵌入

像 SBERT 和 SimCSE 这样的语义嵌入模型被广泛用于测量文本相似性，通常在检索或评估环境中（Reimers and Gurevych, 2019; Gao et al., 2021）。虽然先前的工作将输出与外部参考进行比较，我们重新利用这些模型作为内部诊断工具—量化在语义等价提示下模型内的变化。

PBSS 将嵌入模型重新定义为稳定性透镜：我们不再对正确性进行评分，而是跟踪词元级表示如何聚合成意义的分布漂移。这使得对文体不一致性和潜在提示敏感性进行轻量级的黑箱分析成为可能。

3 PBSS: 用于标记级别行为漂移的诊断框架

3.1 为什么提示差异需要考虑词元的评估

我们将提示方差定义为由表面层次的意图保持的改写触发的 LLM 输出的行为变化。这些变化不是错误或幻觉，而是由微小词汇或句法变化引起的偏移。

关键的是，这种差异通常源于不同的标记序列——标准的准确度指标忽视的变化。我们认为，提示差异反映了一种独特的行为轴，需要评估方法对分词效果和在复述时的修辞稳定性敏感。

3.2 PBSS: 测量提示变化下的行为敏感性

我们将提示基础语义转移（Prompt-Based Semantic Shift，PBSS）定义为一个轻量级框架，用于衡量大型语言模型（LLMs）如何响应改写过的提示——在保留语义意图的情况下进行的表层级别的词汇变化。PBSS 将这种变化视为一种结构化输入扰动，探测词汇或句法上的小幅编辑是否会导致输出的不成比例变化。

使用句子嵌入，我们计算模型对这些改写的响应之间的成对距离。由此产生的漂移矩阵捕捉不同提示下的行为敏感性，突出显示由于分词和解码导致的不稳定性。与准确性指标不同，PBSS 强调修辞一致性：即使意义保持不变，小的词汇层面的变化是否会导致输出不稳定？

正式定义和实现细节详见附录 8 。

3.3 投射漂移：S-BERT 作为行为透镜

PBSS 通过使用 S-BERT 将模型响应投影到语义嵌入空间中量化输出一致性，从而捕捉超越表面变异的句子级别含义。这让我们可以评估措辞不同但意图相同的提示是否能引出风格上连贯的响应。从这个角度来看，提示措辞相当于对标记序列进行轻微的扰动。如果这种表面级别的标记变化产生了大的嵌入分歧，我们将其解释为行为不稳定。因此，PBSS 将语义嵌入重新定义为漂移检测器，无需真实标签，即可捕捉语气、结构或修辞差异。尽管当前的基准测试侧重于正确性，PBSS 则探查在轻微的标记级别变化下模型行为的内部一致性。使用余弦距离提供了一个简单且可解释的衡量指标，用以衡量跨提示、温度和模型的行为变异性。PBSS 最终提出了一个简单却揭示性的问题：当我们以不同方式表达相同内容时，模型的行为是否一致？

4 实验设置

我们通过在十个任务设置中构建受控的提示变体来评估模型对提示分词的敏感性。每个任务包括一个规范提示和一组改写的同义词变体，这些变体在语义偏移最小的情况下改变了表面形式。这些变体在保持意图的同时引发不同的分词路径，从而能够结构化地分析在词级别上的行为不稳定性。

图 2 展示了流水线的概况，完整的数据集、模型和解码配置详见附录 **??** 。

4.1 评估流程概述

图 2 展示了我们的评估工作流程。我们在 10 个现实任务中应用结构化提示扰动，并在两种解码温度下评估 5 种语言模型的行为漂移。输出通过 S-BERT 变体嵌入，并使用 PBSS 漂移分数、累积分布和可视化指标（例如，t-SNE，z-score 热图）进行分析。

我们的目标是评估小的词汇或句法上的改写，即词级的表面变化，是否会引发有意义的输出漂移，即使语义意图保持不变。

4.2 验证提示语义

为了确保表面变化而不改变意图，每个提示集都经过三个检查层：人工审查、基于规则的语法过滤和使用 `all-mpnet-base-v2` 进行语义阈值设定。图 10 确认了所有任务变体之间紧密的语义聚类。

图 11 展示了来自两个任务的示例嵌入，以 S-BERT 相似度（x 轴）和浅层语法距离（y 轴）绘制。提示风格发生变化，但意义得以保留——这使得能够对由分词引起的漂移进行受控分析。

4.3 分词特定的变异维度

表 6 定义了我们的扰动维度，每个维度都旨在调整标记级别的实现，同时保留核心意义。这包括句法重构、语气/风格的变化以及不良输入的压力测试。

4.4 任务覆盖

表 7 总结了 10 个任务领域。提示集模拟了现实世界的互动方式——医疗指导、财务解释、政策说明——提供了一个广泛的基底，用于探究对标记敏感的行为差异。

5 模式和实证发现

5.1 通过嵌入空间聚类在模型输出中涌现的任务结构

为了探查潜在结构，我们使用 SBERT（`all-mpnet-base-v2`）嵌入所有模型输出，并通过 t-SNE 进行降维。在 3 任务和 10 任务的设置中（图 3 –15），输出按任务形成不同的簇——即使没有访问原始标签。这表明任务框架留下了可检测的语义印记，为我们手动定义的任务提供了无监督的验证。

除了基于来源的聚类之外，我们在任务和模型中观察到一个反复出现的中央聚类——反映出低方差、模板式的响应。这表明，某些改写过的提示会引发语义压缩，即尽管表面上有所变化，模型仍默认为通用的指令跟随模式。关键是，这种行为在不同的解码温度（$T = 0.2$ 和 $T = 1.3$）下仍然存在，指向内在的响应动态，而不是采样噪声。

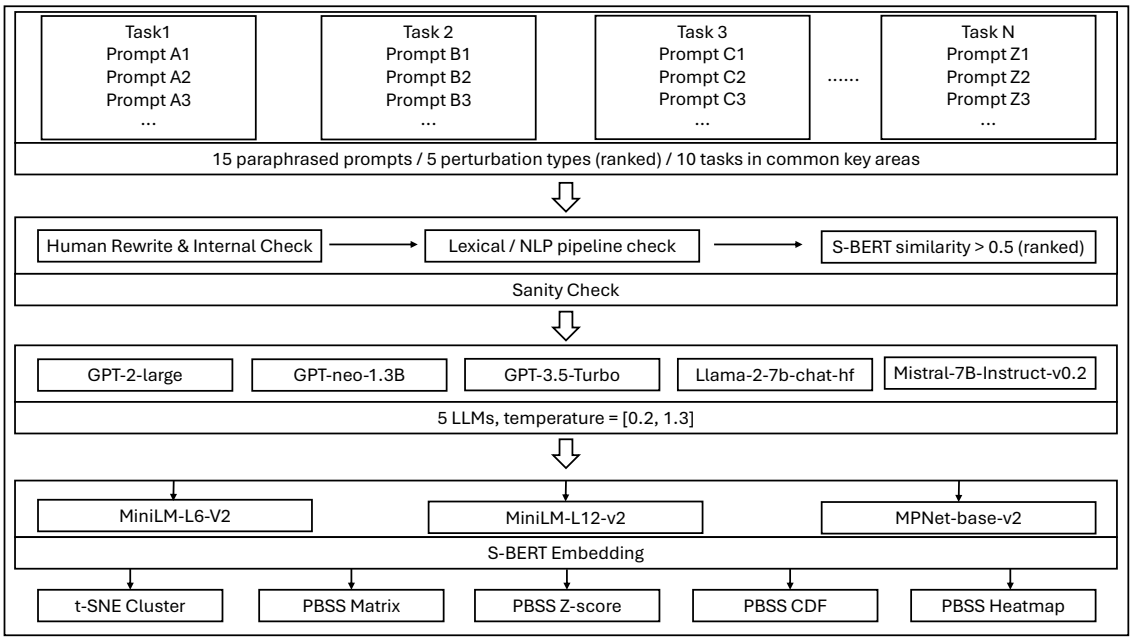


Figure 2: 我们实验框架的概述。结构化提示集在十个任务中构建，经过多层语义和词汇验证，并在两种解码温度下输入到五个大型语言模型（LLMs）。输出通过三个 S-BERT 编码器进行嵌入，并使用 t-SNE、PBSS 矩阵、z-score 漂移检测和累积分布进行分析。该设计使得在释义压力下进行受控的行为漂移分析成为可能。

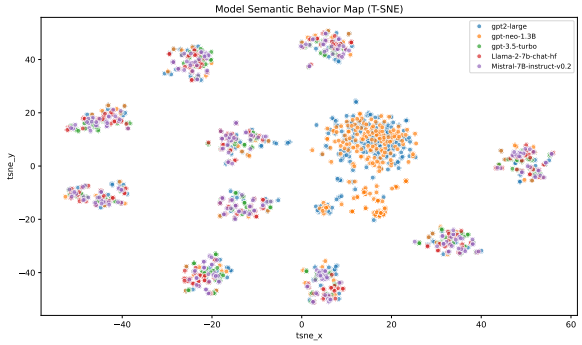


Figure 3: 模型空间 t-SNE（10 个任务，5 个模型）- 1

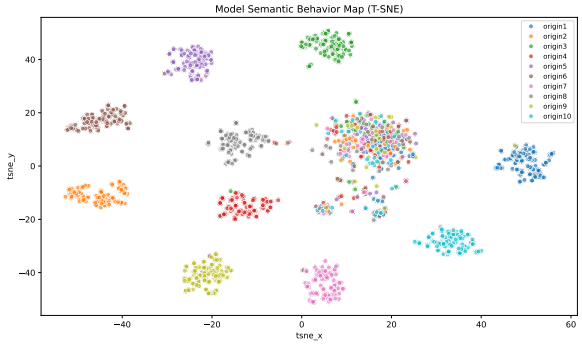


Figure 4: 原始空间 t-SNE（10 个任务，5 个模型）- 1

5.2 SBERT 作为编码器的理由

我们使用 SBERT 不是为了绝对的忠实性，而是因为它对修辞和语义变化的高敏感度——这对于捕捉由提示引起的偏移非常理想。PBSS 依赖于相对的输出结构，而 SBERT 始终如一地揭示了不同变体、模型和解码设置中的模型特定模式。

5.3 通过语义嵌入一致性验证任务有效性

尽管任务类别是手动定义的，但 SBERT+t-SNE 投影在没有监督的情况下恢复了它们，验证了我们提示集的内部一致性。这表明任务的框架不仅塑造了输出内容，还塑造了修辞结构，产生了稳定且可衡量的行为特征。

5.4 跨嵌入架构的稳定漂移模式

为了测试 PBSS 结果是否依赖于句子编码器，我们使用三个 SBERT 变体计算漂移分数的 CDF：`MiniLM-L6`，`MiniLM-L12` 和 `all-mpnet-base-v2`（图 16 - 24）。在不同的编码器之间，我们观察到一致的模型排名和曲线形状：GPT-3.5、LLaMA-2 和 Mistral 显示出稳定、低方差的行为；GPT-2 和 GPT-Neo 显示出更广泛的漂移。

5.5 行为漂移作为相界

CDF 图显示了一个明显的分界：微调过指令的模型（如 GPT-3.5、LLaMA-2）在提示变化下紧

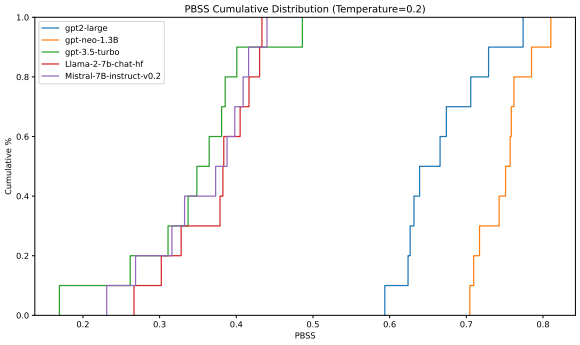


Figure 5: MiniLM-L6 在 $T = 0.2$ 下的 PBSS CDF

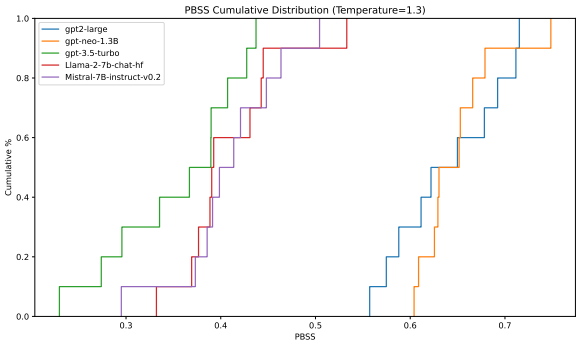


Figure 6: MiniLM-L6 在 $T = 1.3$ 下的 PBSS CDF。

密聚集，而较旧的模型（GPT-2、GPT-Neo）则表现出广泛且不稳定的分散。这不是一个渐进的趋势，而是行为上的相变——较新的模型能够在不发生语义漂移的情况下适应措辞变化，而较旧的模型则陷入重复或分歧。因此，PBSS 揭示了对齐意识架构与传统架构之间的明显界限。

5.6 在变化下的 PBSS：跨模型、温度和任务的鲁棒性

随着任务覆盖范围从 3 个领域扩大到 10 个领域，PBSS 分数保持稳定和一致，CDFs 展现出明确的模型特定模式。在图 5 和 6 中，将不同温度（$T = 0.2, 1.3$）的输出汇聚增加了变动性，但并未破坏漂移结构。这些结果表明 PBSS 捕捉到了持久的行为信号——在提示领域、解码熵和模型家族中均表现出鲁棒性。

在 3 任务和 10 任务的设置中，我们观察到强烈的组间差异，这些差异在所有 SBERT 编码器变体中保持一致，强化了观察到的漂移模式的稳健性和可靠性。

5.7 描述性统计。

为了进一步评估稳健性，我们扩展了模型池，并按照参数规模系统性地对其分类（表 1），涵盖了从传统架构到现代指令调整系统的范围。此分组反映了先前研究中观察到的重要架构和对齐差异。

我们的扩展评估涵盖了 10 个不同的任务，每个任务有 5 个提示集和每个集 15 个语义等效的改写，总计每个模型有 4500 个提示情况。虽然这可能看起来规模很大，但我们的目标并不是详尽地探查每一个响应的极端情况，而是抽取足够的变体以检测新出现的模式。由于 PBSS 的设计是为了揭示在保持语义不变的改写下的结构输出漂移，较高的覆盖范围提升了我们识别重复行为特征的能力，而无需大规模的全面扫描。

重要的是，我们的测试平台是围绕语言变异的受控爆炸而设计的，不是为了引发过拟合，而是为了压力测试相同的语义输入是否在不同的提示形式下导致一致的修辞行为。这种多样性对于揭示潜在的漂移倾向至关重要，尤其是在跨模型层比较编码器输出时。

综合 CDF 图和热图诊断结果适用于完整的评估集，详见附录 **??** 。

我们将模型分为三个以经验为基础的组：

- 传统小模型（例如，GPT-2、GPT-Neo-1.3B、SmolLM）——这类模型在对齐前或者经过轻微微调，参数数量在 20 亿以下，在释义扰动情况下表现出较高的方差和较差的一致性。
- 中型、轻度对齐的模型（例如，LLaMA-2-7B、Phi-2、Mistral-7B、OpenChat）——最近的开源模型，具有对齐目标，但容量较低或微调深度较浅，表现出中间的漂移行为。
- 大容量指令调优模型（例如，GPT-3.5-Turbo，MythoMax-13B）——与商业对齐的系统在重新措辞方面具有强一致性，表明在提示变化下具备抗漂移能力。

该分类法反映了对齐历史和容量层级，并重现了在 PBSS 分布中观察到的明确的行为边界。通过 Kruskal–Wallis 检验确认这些群体之间存在统计显著的差异。这种结构化分组使我们能够比较漂移动态，不仅在模型层面上进行比较，还可以跨理论相关的行为机制进行比较。

所有模型和编码器的综合 CDF 可视化和热图诊断在附录 **??** 中给出。

我们对三个 SBERT 编码器和一个按大小和对齐阶段分组的模型分层集合的 PBSS 分数应用 Kruskal-Wallis 检验。虽然大多数比较产生了极其显著的差异（$p \ll 0.01$），尤其是在传统系统和经过指令微调的系统之间，我们注意到在具有相似对齐和规模的模型之间的比较中，显著性变得边缘化或消失 - 如果行为趋同

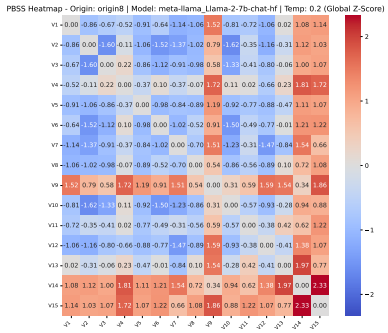


Figure 7: MiniLM-L6：显示出结构上的分歧和较弱的聚类。

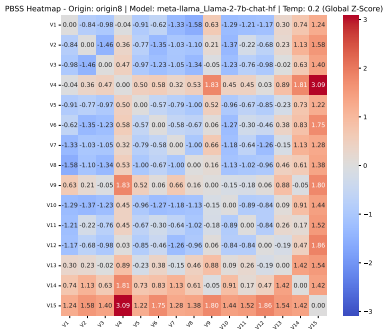


Figure 8: MiniLM-L12：显示更紧密的热图簇。

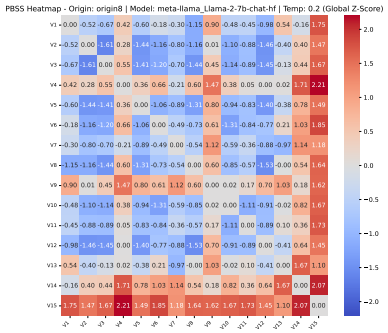


Figure 9: MPNet：在释义之间保持一致的结构。

Table 1: 模型参数大小类别

Model Name	Parameter Count	Size Category
HuggingFaceTB SmolLM-360M	∼ 0.36B	Small
GPT2-Large	∼ 0.76B	Small
EleutherAI GPT-Neo-1.3B	∼ 1.3B	Small-Medium
Microsoft Phi-2	∼ 2.7B	Medium
Meta-LLaMA Llama-2-7b-chat-hf	7B	Medium-Large
MistralAI Mistral-7B-Instruct-v0.2	7B	Medium-Large
OpenChat OpenChat-3.5-1210	∼ 7B	Medium-Large
MythoMax-I2-13B	13B	Large
GPT-3.5-turbo	Undisclosed (API)	Extra-Large

正在发生，这是预期的结果。为了从结构上评估这些模式，我们将我们的模型池分为三个层次：（1）传统小型模型如 GPT-2 和 GPT-Neo，它们在释义变体下表现出高漂移和弱的修辞一致性；（2）中型轻对齐系统，包括 LLaMA-2 和 Phi-2，它们表现出减弱但仍不均匀的漂移模式；以及（3）高容量指令微调模型如 GPT-3.5，它们表现出低 PBSS 方差和有限的提示间不稳定性。重要的是，这项研究并不是为了全面衡量可能的提示或采样条件的空间。相反，它作为一种结构化的合理性检查：如果经过良好策划的提示网格可以在仅使用适度数据量的情况下展示可解释的行为层次，这意味着存在潜在的、内部结构化的模型敏感性。我们报告了最近评估的结果，包括跨越 10 个任务、每个任务 5 个提示集、每个集 15 个语义变体、以及 2 种解码温度——总计超过每个模型 13,000 个提示-响应案例。至关重要的是，PBSS 信号在编码器之间保持稳健性：尽管绝对分数的幅度存在小差异，模型的相对排名依然保持一致。这种编码器不可知的稳定性——我们称之为语义共振——表明 PBSS 捕获了模型的真实行为属性，而非由嵌入或提示结构引起的噪声。这些发现共同确认语义漂移是可以衡量的、结构化的、模型特定的——并且 PBSS，即使在受限制的评估下，也能够为这种大语言模型行为的潜在轴提供可靠的视角。

5.8 令牌级漂移的统计摘要

使用多种方法——Kruskal-Wallis 检验、累积分布和 t-SNE 投影，我们发现表面级提示变化会引发行为漂移的持续证据。这些效应在 SBERT 变体、解码温度和提示领域中持续存在。

尽管每个 PBSS 分数反映了特定的成对提示比较，提示的构建是在不同的任务、释义模板和词汇实现中刻意多样化的。每个模型的 750 个独特指令生成了 300 个不同的提示对，并且不重复使用提示词干——这样最大限度地减少了词汇记忆效应和表面过拟合。

这种一致性——跨编码器结构和温度范围——表明 PBSS 捕捉的是模型内部的规律性，而不是由提示设计引起的虚假相关性。

关键的是，这种漂移不是噪声：它反映了模型特定的指令令牌形式的结构化敏感性。即使在意义保持不变的情况下，不同的表面实现，通常是在令牌级别，也会导致不同的模型行为。PBSS 使这种变化可见，为评估模型对语义等效但令牌不同的提示的响应一致性提供了一种实际的诊断视角。如果 PBSS 仅仅反映提示设计的特异性，我们会预期看到编码器分歧

Table 2: 所有模型的整体 PBSS 描述性统计数据

Model	Count	Mean	Std Dev	25 % Quartile	75 % Quartile
GPT-3.5-Turbo	300	0.422	0.116	0.338	0.509
Mistral-7B	300	0.427	0.103	0.351	0.504
LLaMA-2-7B	300	0.453	0.115	0.362	0.534
MythoMax-13B	300	0.464	0.117	0.384	0.540
OpenChat-3.5	300	0.462	0.115	0.370	0.550
Phi-2	300	0.583	0.129	0.489	0.687
SmolLM-360M	300	0.588	0.090	0.520	0.660
GPT-Neo-1.3B	300	0.650	0.103	0.566	0.736
GPT-2 Large	300	0.635	0.094	0.563	0.705

Table 3: Kruskal-Wallis H 统计量和 p 值用于跨模型规模组和 SBERT 编码器的 PBSS。测试在每个编码器上进行，并在不同的温度设置下汇总。

Group	Encoder	H (All)	p (All)	H ($T = 0.2$)	p ($T = 0.2$)	H ($T = 1.3$)	p ($T = 1.3$)
Small	MiniLM-L6	23.86	6.6×10^{-6}	25.15	3.5×10^{-6}	27.22	1.2×10^{-6}
	MiniLM-L12	18.18	1.1×10^{-4}	37.14	8.6×10^{-9}	20.67	3.2×10^{-5}
	MPNet	25.76	2.6×10^{-6}	50.48	1.1×10^{-11}	15.36	4.6×10^{-4}
	Combined	—	—	103.01	4.3×10^{-23}	31.07	1.8×10^{-7}
Mid	MiniLM-L6	72.36	1.3×10^{-15}	8.52	3.6×10^{-2}	107.11	4.6×10^{-23}
	MiniLM-L12	100.23	1.4×10^{-21}	21.95	6.7×10^{-5}	111.54	5.1×10^{-24}
	MPNet	67.93	1.2×10^{-14}	7.67	5.3×10^{-2}	97.44	5.5×10^{-21}
	Combined	—	—	34.97	1.2×10^{-7}	306.73	3.5×10^{-66}
Large	MiniLM-L6	7.79	5.3×10^{-3}	2.78	9.5×10^{-2}	5.24	2.2×10^{-2}
	MiniLM-L12	3.80	5.1×10^{-2}	1.36	2.4×10^{-1}	2.76	9.7×10^{-2}
	MPNet	7.34	6.7×10^{-3}	2.56	1.1×10^{-1}	4.78	2.9×10^{-2}
	Combined	—	—	6.62	1.0×10^{-2}	12.14	4.9×10^{-4}
All Models	MiniLM-L6	428.89	1.2×10^{-87}	260.94	8.2×10^{-52}	254.38	2.0×10^{-50}
	MiniLM-L12	385.94	1.9×10^{-78}	288.33	1.3×10^{-57}	203.61	1.1×10^{-39}
	MPNet	288.76	1.0×10^{-57}	251.67	7.6×10^{-50}	158.43	3.4×10^{-30}
	Combined	—	—	791.58	1.3×10^{-165}	553.12	2.8×10^{-114}

和温度不稳定，但这两者都没有被观察到。

6 通过 PBSS 进行面向 QoS 的评估

在确定了提示差异反映的是结构化的行为动态而不是偶然的噪声后，我们现在介绍一个实用框架，用于系统地评估模型对语义等价的提示改写的响应。我们的方法，即基于提示的语义转移 (PBSS)，可以解释为对大型语言模型 (LLM) 的服务质量 (QoS) 一致性的一种轻量级诊断。

6.1 动机：从噪声到结构化方差

先前的分析，包括统计测试和 t-SNE 投影，表明模型对改写后的提示的响应并非随机分散，而是形成结构化的簇，这表明存在系统性的行为漂移。PBSS 通过服务质量（QoS）的视角重新审视这些观察，将焦点从正确性转向行为可靠性：在基本语义意图保持不变时，模型的响应有多一致？

6.2 PBSS 作为 QoS 指标

传统的评估框架通常优先考虑任务的准确性，但在提示措辞变化下的一致行为对语言模型的安全和可靠部署同样至关重要。PBSS 指标和热图为评估这种一致性提供了一个实用且可解释的界面，填补了当前大型语言模型诊断中的空白。例如，当模型被要求使用五个稍微不同的提示来描述相同的医学模糊性时，为什么有些回答会收敛成固定的措辞，而其他回答在语气和结构上则差异很大？

PBSS 使这种变化变得可见、可追踪，并且重要的是，可诊断的。这对于像临床自然语言处理这样的领域有重要影响，在这些领域，大型语言模型越来越被用于生成摘要、治疗建议或面向患者的解释。在这些情况下，叙述结构与解释紧密相关：风格或强调上的细微变化可以影响后续的临床决策。通过量化在受控释义变化下的行为一致性，PBSS 为评估医学语言模型的稳健性提供了一个关键的诊断层。

值得注意的是，我们观察到不同的 SBERT 编码器在采用全局 z-score 标准化时生成了不同的 PBSS 热图结构。这突显了 PBSS 对输出变量性以及由选择的嵌入模型引入的语义细粒度的敏感性。这种变化是预期的，因为 SBERT 编码器在建模句子级相似性方面存在不同，这由训练目标（例如，自然语言推理与语义文本相似性）、架构深度和嵌入尺度敏感性的差异所决定。

7 讨论

7.1 影响与界限

本研究不旨在模拟身份、个性或拟人化一致性。所有观察均来源于标准指令跟随框架内的中立、任务导向型提示，并未使用对抗性的或微调的输入。

重要的是，我们观察到的漂移模式不是孤立的异常现象，而是具有一致性和统计可验证的趋势。它们从语义稳定、非对抗性的提示中出现，突显出语言模型中的一种潜在行为维度——这种维度是可测量的、可再现的，并且对开发人员来说越来越重要。

7.2 提示漂移和越狱行为

越狱可以被重新定义为提示诱导漂移的极端案例。与其仅将越狱视为安全失效，我们将其解读为提示空间中高敏感区域的响应——在这些区域中，微小的变化会产生不成比例的偏离输出。

PBSS 可能有助于及早识别这些区域。虽然我们没有直接测试越狱提示，但未来的工作可以探讨高 PBSS 漂移是否与越狱易感性相关，从而允许从行为学角度进行主动的模型压力测试。

7.3 临床相关性和高风险领域

某些下游领域——尤其是临床和诊断环境——对文体的歧义性和修辞的不一致性极为敏感。医学提示往往语义浓厚，结构受限，并且高度依赖于上下文。在这样的环境中，即使是提示措辞的细微变化也可能导致相互矛盾的解释，从而影响安全性和可靠性。

与需要表达变化的开放式创造性任务不同，临床应用需要语义的精确性和确定性回应。在这些情况下，提示变化不仅仅是一个语言异常——它代表了一个潜在的风险因素。

随着将人工智能系统纳入临床工作流程的兴趣不断增加，在输入提示时没有错误的余地。即使是轻微的语义变化也可能导致灾难性后果，即对临床数据的误解造成误诊，从而可能导致患者更糟糕的结果，特别是在需要立即行动的环境，例如急诊室。在这种高风险环境中，沟通的准确性至关重要；模糊性几乎不存在。一个由人工智能生成的电子病历（EMR）出错或推荐被误解可能会延迟治疗、误导干预或掩盖重要的诊断信号。因此，确保构建提示和模型响应的方式的一致性和清晰度不仅仅是一个技术问题——而是一个临床上的迫切需要。在急诊医学、诊断和决策支持中集成人工智能必须伴随着对其在不同措辞下语义稳定性的严格验证，以保护患者结果和维护护理标准。

在这里，PBSS 提供了一种低成本的前端筛选工具：它在实际高风险工作流程部署之前，标记在良性变化下的反应性。随着大型语言模型（LLMs）逐渐接近医疗决策支持，提示差异可能成为一个风险向量，其中表面措辞的差异在保持意图不变的情况下产生不同的输出。

7.4 模式，而非噪声

PBSS 不评估模型所说的内容，而是评估它们的表达方式，为改写变化下的修辞一致性提供一个最小的视角。我们的结果显示出跨重述的结构化、可重复的行为转换。模型展示出独特的修辞反应模式，在模型家族内保持一致，但对措辞变化敏感。

我们不做拟人化的声明。但我们确实观察到了模型特定的行为景观——在语义表面上，释义引发了可预测的修辞分歧。理解这一结构对于可解释性、部署和界面的稳健性至关重要。

8 未来方向

PBSS 作为行为透镜 我们将 PBSS 不是作为一个基准，而是作为一个最小的探针，用于研究在语义固定、风格各异的提示下的大型语言模型一致性。我们的结果表明，这种变化反映了结构化的行为漂移——是系统性的，而不是偶然的。

词级漂移和潜在惯性 漂移通常源于措辞或语调的差异——这表明由预训练塑造的持久风格先验。这种惯性可能反映出更深层次的提示处理偏见。在多轮或角色自适应环境中研究它，可能会揭示出模型随着时间的推移如何编码修辞上下文。

漂移作为越狱信号 PBSS 提供了一种非对抗性的视角来审视越狱现象。在提示空间中的高漂移区域可能与模型利用或误解的不稳定性阈值对齐。这为越狱风险诊断开辟了新途径，而无需依赖剥削工程。

在医学或政策等领域，修辞不一致是一种风险因素。PBSS 可以作为一个轻量级的前端过滤器，在模型部署于高风险场景之前，标记提示中的不稳定性。

走向可解释性 提示重述激活了模型的标记化和解码景观中的不同路径。PBSS 可以帮助将表面行为与内部动态连接起来，从而获得关于大型语言模型如何构建意义的认知和解释性见解。

开放框架 我们发布了 PBSS 作为一个简化、可扩展的诊断框架——邀请社区将其拓展为研究由标记引起的行为差异的基础，不仅限于准确性，还包括一致性和修辞对齐。

References

Shengyuan Bai et al. 2025. Enhancing nlu in large language models using adversarial noisy instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Mingyang Chen et al. 2024. Facilitating multi-turn function calling for llms via compositional instruction tuning. *arXiv preprint arXiv:2410.12952*.

Yuyan Chen et al. 2025. Attributive reasoning for hallucination diagnosis of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

J. Gao, D. Tang, Y. Xu, X. Feng, Z. Zhang, L. Shou, B. Qin, D. Jiang, and T. Liu. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–707. https://arxiv.org/abs/2104.08821.

Ziwei Ji et al. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zekun Li et al. 2023. Evaluating the instruction-following robustness of large language models to prompt injection. *arXiv preprint arXiv:2308.10819*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez et al. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.

Amrith Setlur et al. 2024. Prompting is a double-edged sword: Improving worst-group robustness of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lexin Zhou et al. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

Runchuan Zhu et al. 2025. Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

A PBSS 框架的正式定义

令 $\mathcal{P}$ 表示一个固定任务的一组释义提示。令 f 为一个语言模型，使得 $f : \mathcal{P} \to \mathcal{Y}$ 将输入映射到文本输出。对于提示 $p_i, p_j \in \mathcal{P}$ ，令 $y_i = f(p_i)$ 和 $y_j = f(p_j)$ 为生成的响应。

令 $s : \mathcal{Y} \to \mathbb{R}^d$ 代表一个嵌入函数（例如，S-BERT）。我们定义漂移指示器 D 为输出之间的余弦距离：

$$D(p_i, p_j) = 1 - \cos(s(y_i), s(y_j)) = 1 - \frac{s(y_i) \cdot s(y_j)}{\|s(y_i)\| \cdot \|s(y_j)\|}$$

漂移评分量化了当模型面对表面形式不同但语义意图一致的提示时，输出在语气、结构或风格上变化的程度。在 PBSS 框架中，我们计算所有重述提示对之间的得分 D ，将结果汇总成一个诊断漂移矩阵，该矩阵捕捉模型在变异中的敏感性特征。

PBSS 累积分布（CDF）。 给定一组 n 改写的提示 $\{p_1, ..., p_n\}$ ，我们使用语义嵌入函数 $s(\cdot)$ 计算所有成对的 PBSS 漂移分数：

$$D_{i,j} = 1 - \cos(s(f(p_i)), s(f(p_j))), \quad \forall i \neq j$$

这导致了在提示集上的一组 $\binom{n}{2}$ 漂移评分。我们定义 PBSS 累计分布函数（CDF）为：

$$F(\delta) = \frac{1}{|S|} \sum_{(i,j) \in S} \mathbb{I}[D_{i,j} \leq \delta]$$

其中 $S = \{(i, j) \mid i < j\}$ 和 $\delta \in [0, 2]$ （最大余弦距离范围）。$F(\delta)$ 表示行为差异小于或等于阈值 δ 的提示对的比例。

这种分布反映了模型对表层提示变化的总体稳定性。累积分布函数（CDF）在接近 $\delta = 0$ 处陡峭上升，表示高行为一致性，而较平缓或右移的曲线则反映了较大的变异性或对释义提示更频繁的漂移。在 PBSS 框架中，我们通过计算所有提示对之间的漂移评分 D 来量化这种变异性，并将结果整理成一个诊断漂移矩阵，来总结模型的敏感性概况。

在语义等价性不完美或不确定的情况下，我们引入一种结合语义相似性的混合评分：

$$\text{PBSS}_{\text{hybrid}}(p_i, p_j) = \lambda \cdot \text{Sim}_{\text{sem}}(y_i, y_j) + (1 - \lambda) \cdot \text{PBSS}(p_i, p_j)$$

其中，Sim_{sem} 指的是一种语义相似性度量（例如，S-BERT 余弦相似度），$\lambda \in [0, 1]$ 控制语义保留和风格一致性之间的权衡。我们注意到，Sim_{sem} 是一种可能的实现，并不是 PBSS 内在的。

给定 n 提示 $\{p_1, \ldots, p_n\}$ ，我们计算完整的 PBSS 矩阵 $D \in \mathbb{R}^{n \times n}$ ，其中 $D_{i,j} = \text{PBSS}(p_i, p_j)$ 对于 $i \neq j$ 和 $D_{i,i} = 0$ 。

从这个矩阵中，我们推导出：

- 平均 PBSS: $\mu = \frac{1}{n(n-1)} \sum_{i \neq j} D_{i,j}$
- 最大 PBSS： $\max_{i \neq j} D_{i,j}$
- CDF 曲线： 所有 (i, j) 对中 PBSS 值的经验分布

为了检测显著的行为异常，我们定义了两种 z 评分方法：

（1）所有非对角线项的全局 Z-score：

$$Z_{i,j}^{\text{global}} = \frac{D_{i,j} - \mu_D}{\sigma_D}, \quad i \neq j$$

其中，μ_D 和 σ_D 分别是 D 在 $i \neq j$ 上的全局平均值和标准偏差。

(2) 行归一化的 Z-评分，测量相对于每个提示邻域的偏差：

$$Z_{i,j}^{\text{row}} = \frac{D_{i,j} - \mu_i}{\sigma_i}, \quad i \neq j$$

Table 4: 提示变体记录模式

Field	Description
`origin`	Original task prompt identifier
`variant_id`	Prompt variant ID
`model_name`	Model name (e.g., GPT-2)
`temperature`	Decoding temperature
`prompt`	Prompt string
`output_text`	Model-generated output

Table 5: PBSS 评估总结模式

Field	Description
`origin`	Original task prompt identifier
`model`	Language model used for generation
`temperature`	Decoding temperature during inference
`avg_pbss`	Mean drift score across all prompt variants

其中，μ_i 和 σ_i 是除去对角线的行 i 的均值和标准差。

这些指标允许局部漂移模式的出现，突出显示那些由于措辞而引发过度行为反应的提示。

我们评估了五个大型语言模型：GPT-2 (774M)、GPT-Neo (1.3B)、LLaMA-2 (7B)、Mistral (7B)、和 GPT-3.5-Turbo（通过 OpenAI API）。使用相同的提示批处理在两个解码温度下对模型进行查询：$T = 0.2$（低方差）和 $T = 1.3$（高方差）。该设置旨在测试模型在稳定和随机生成情况下的行为一致性。

开放权重模型通过 HuggingFace Transformers 进行访问，采用贪婪或 top-k 抽样（视情况而定）。GPT-3.5-Turbo 通过 OpenAI Platform API v1 进行访问。

A.1 句子嵌入模型

为了计算 PBSS 漂移分数，我们使用三个 Sentence-BERT (S-BERT) 变体嵌入生成的输出：

- MiniLM-L6-v2：紧凑的 6 层 Transformer，经过优化以实现快速推理；
- MiniLM-L12-v2：具有更高表示粒度的更深版本；
- all-mpnet-base-v2：通过置换语言模型预训练；用于过滤和验证。

这些模型不仅在深度和参数数量上存在差异，而且在训练目标上也不同——例如，被遮掩的语言建模与置换的语言建模。通过在整个评估流程中并行应用这三个编码器，我们评估行为漂移模式在嵌入架构中是否保持一致。这起到了稳健性检查的作用，尽量减少对任何单一语义表示的依赖。

所有嵌入向量均进行余弦归一化，并通过 `sentence-transformers` Python 库进行批量推理处理。

A.2 提示日志架构

所有模型输出都通过一个轻量级、模块化的日志系统捕获，该系统旨在实现可重复性和易于扩展性。为了实现系统化的行为分析，我们进一步形式化了两个结构化的数据模式：提示变体日志模式和 PBSS 评估摘要模式。这些模式作为可扩展的模板，用于记录提示扰动、模型参数和聚合漂移分数，旨在促进下游诊断、比较评估以及与未来工具的集成。

A.3 代码库和可重复性

这种结构能够在不同模型和温度设置中精确跟踪每个提示变体，从而促进对齐的行为比较。

Table 6: 提示扰动维度的定义

Dimension	Description
Stylistic Shift	Varies tone/register (e.g., formal to conversational).
Syntactic Manipulation	Reorders or modifies grammatical structure.
Instructional Perturbation	Alters framing verbs or phrasing.
Contextual Reframing	Wraps tasks in scenario contexts.
Broken Prompt	Injects malformed or compressed input.

Table 7: 提示簇在使用领域中的覆盖

Case ID	Domain	Task Description
C1	Medical Explanation	Explain motion-induced boundary ambiguity in MRI scans
C2	Causal Inference	Identify confounding factors in paradoxical drug effects
C3	Urban Policy	Design fair and effective weekday traffic restriction policy
C4	Public Science	Communicate climate effects in varying tone registers
C5	Primary Education	Explain a basic concept in age-appropriate language
C6	Academic Appeals	Justify a request for grade review
C7	Major Transfer	Motivate an undergraduate program switch convincingly
C8	Financial Forms	Write compliant yet natural reimbursement justifications
C9	Audit Response	Draft formal responses to regulatory inquiries
C10	Request Protocol	Vary formality and tone in formal administrative requests

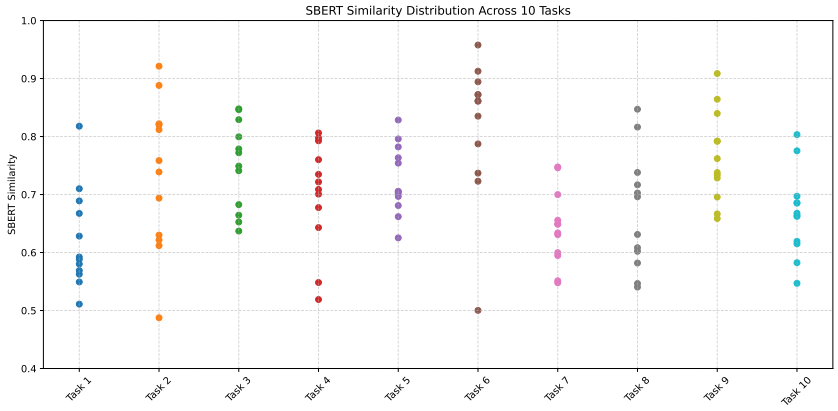
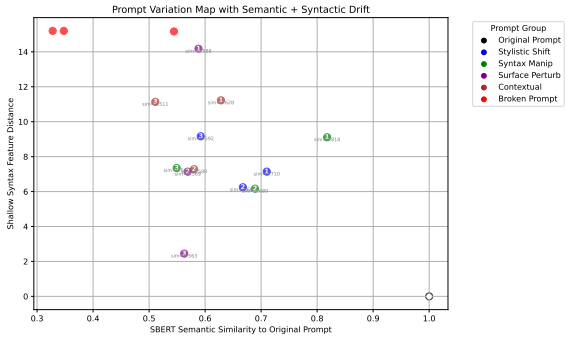


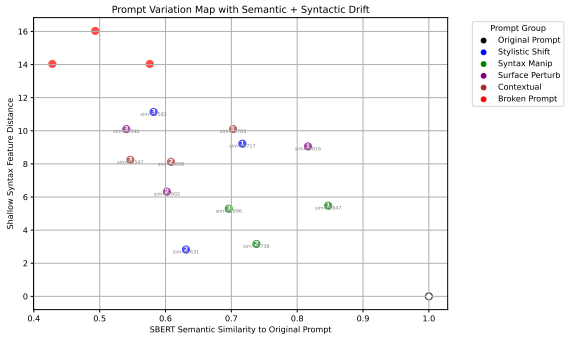
Figure 10: 基于 S-BERT 的提示变体在 10 个任务中的相似度分布。每个任务的提示集包含 15 种与原始指令在语义上对齐的提示变体。

为了支持我们的主要分析，我们在下面提供了扩展的可视化内容。除了主文本中分析的五个核心模型外，我们还引入了五个新的 LLM，以在不同的架构家族中进行交叉验证。我们在 3 任务和 10 任务子集上可视化漂移模式，以评估在不同语义和词汇约束下的一致性。

模型输出通过 S-BERT 嵌入，并采用 t-SNE 进行投影。图表按照模型身份（模型空间）和提示来源（来源空间）进行组织。3 任务设置捕捉更细致的提示级别变化，而 10 任务配置则提供更广泛的领域覆盖。在来源空间中的视觉一致性反映了对释义提示输出的稳定性；分散或碎片化的簇表明较高的行为敏感性。



(a) C1: Medical Explanation



(b) C8: Financial Forms

Figure 11: 每个点代表一个提示变体，由 S-BERT 相似性（x 轴）和浅层语法距离（y 轴）绘制。这个合理性检查在视觉上确认提示集在展示标记级表面变化的同时保持语义一致性。

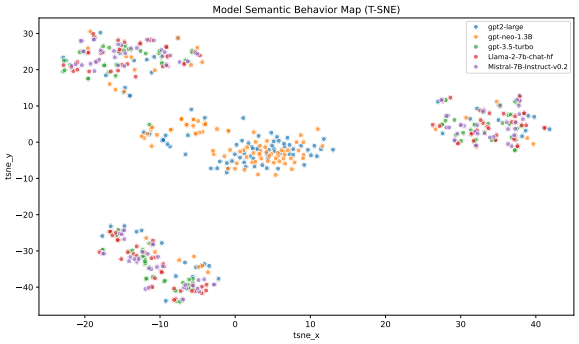


Figure 12: 模型空间 t-SNE（3 个任务，5 个模型）

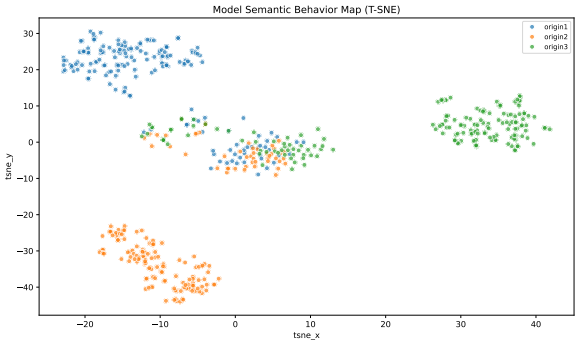


Figure 13: 原始空间 t-SNE（3 个任务，5 个模型）

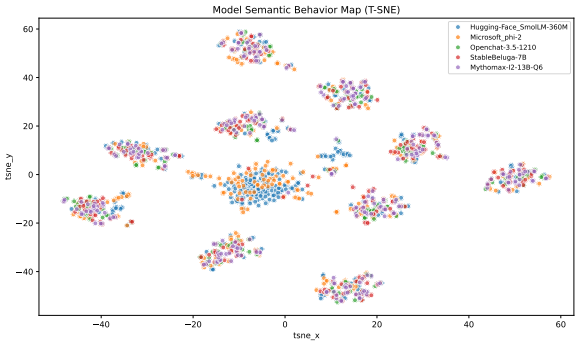


Figure 14: 模型空间 t-SNE（10 个任务，5 个新模型）- 2

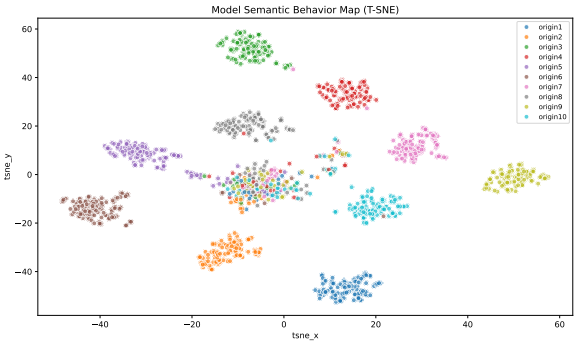


Figure 15: 初始空间 t-SNE（10 个任务，5 个新模型）- 2

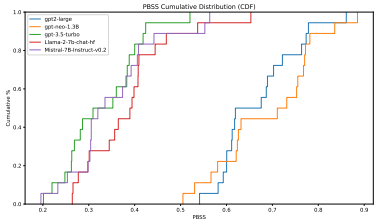


Figure 16: PBSS CDF for MiniLM-L6: 3 tasks.

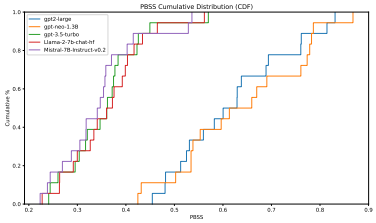


Figure 17: PBSS CDF for MiniLM-L12: 3 tasks.

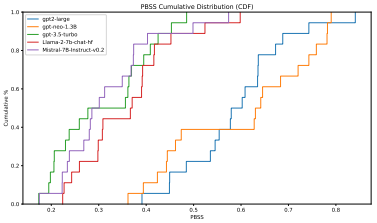


Figure 18: PBSS CDF for MPNet: 3 tasks.

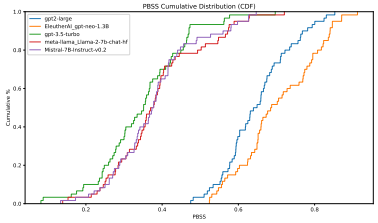


Figure 19: PBSS CDF for MiniLM-L6: 10 tasks.

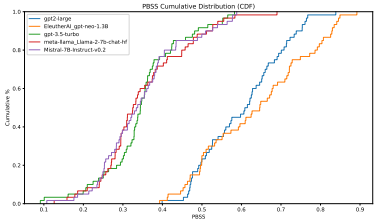


Figure 20: PBSS CDF for MiniLM-L12: 10 tasks.

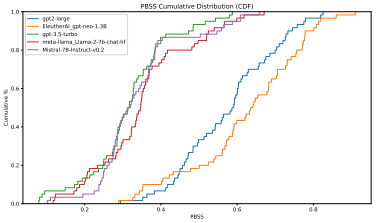


Figure 21: PBSS CDF for MPNet: 10 tasks.

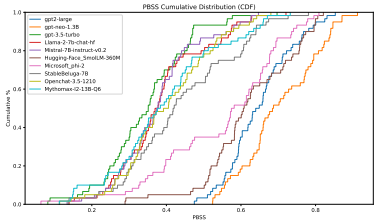


Figure 22: PBSS CDF for MiniLM-L6 (alt case).

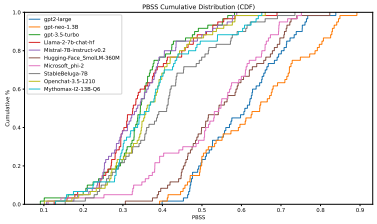


Figure 23: PBSS CDF for MiniLM-L12 (alt case).

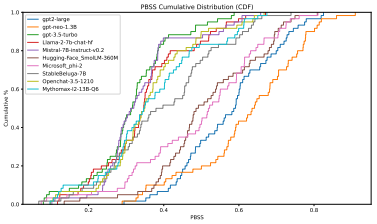


Figure 24: PBSS CDF for MPNet (alt case).

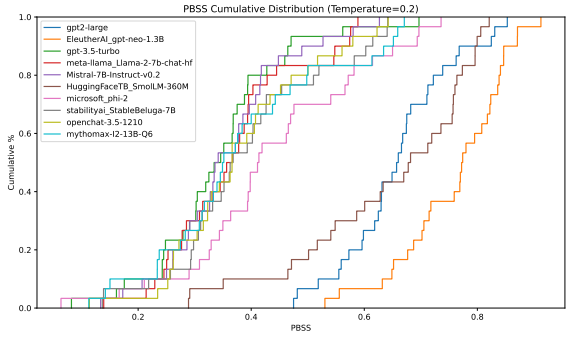


Figure 25: PBSS CDF for MiniLM-L6, $T = 0.2$.

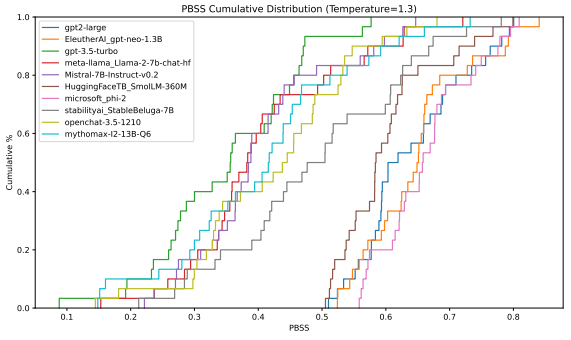


Figure 26: PBSS CDF for MiniLM-L6, $T = 1.3$.

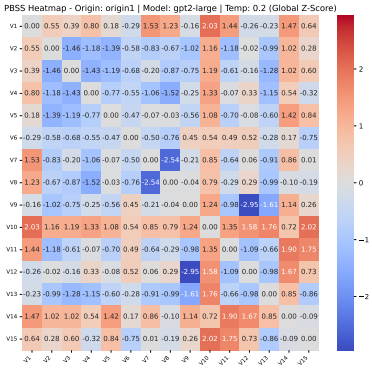


Figure 27: GPT-2：全局 Z 分数

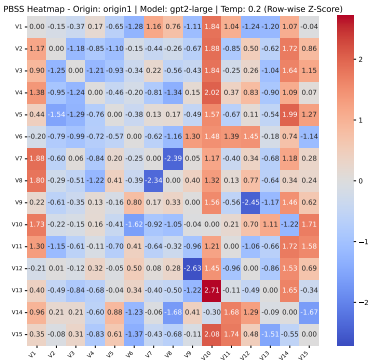


Figure 28: GPT-2: 行内标准分数

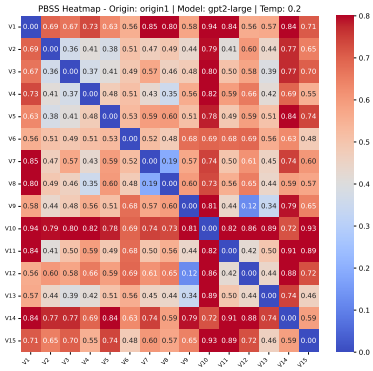


Figure 29: GPT-2：原始相似性

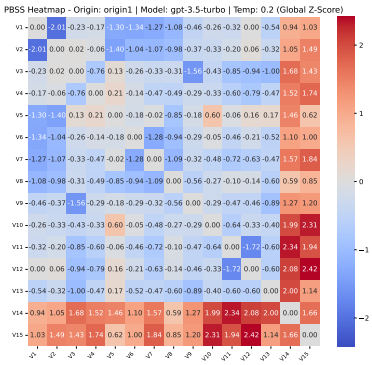


Figure 30: GPT-3.5：全局 Z 分数

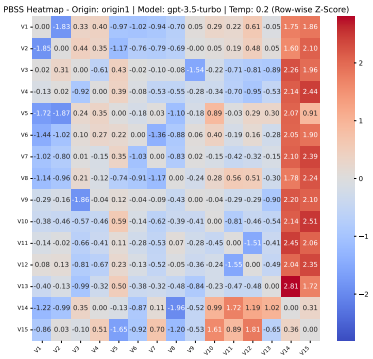


Figure 31: GPT-3.5：行内 Z 分数

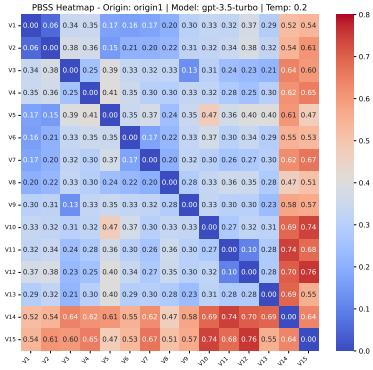


Figure 32: GPT-3.5：原始相似性

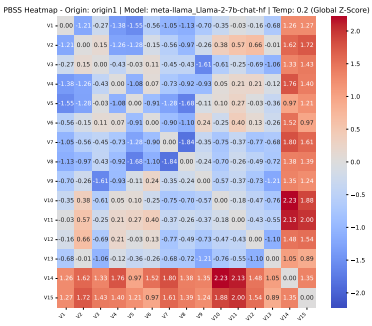


Figure 33: LLaMA-2: 全局 Z-得分

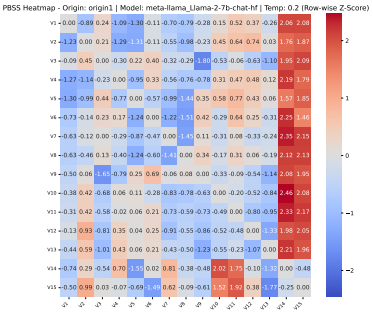


Figure 34: LLaMA-2：行均值标准分数

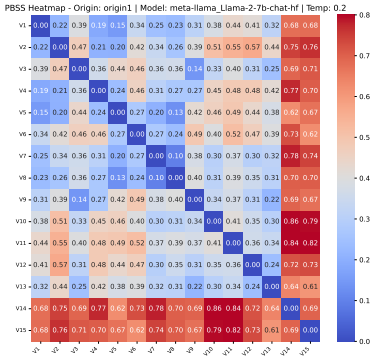


Figure 35:
LLaMA-2: 原始相似性