

Statistical Hypothesis Testing for Auditing Robustness in Language Models

Paulius Rauba¹ Qiyao Wei¹ Mihaela van der Schaar¹

英寸

Abstract

考虑一个问题，即测试大型语言模型（LLM）系统的输出在任意干预下是否发生变化，例如输入扰动或模型变体更改。我们不能简单地比较两个 LLM 输出，因为它们可能由于系统的随机性而不同，也不能比较整个输出分布，因为计算上不可行。尽管现有的方法可以分析基于文本的输出，但它们着眼于根本不同的问题，如测量偏见或公平性。为此，我们引入了基于分布的扰动分析，这是一种将 LLM 扰动分析重新表述为频率假设检验问题的框架。我们通过蒙特卡罗采样在低维语义相似性空间中构建经验上的无效和备选输出分布，从而在无需限制性分布假设的情况下实现可处理的推断。该框架具有以下特点：(i) 与模型无关，(ii) 支持对任何黑盒 LLM 的任意输入扰动评估，(iii) 产生可解释的 p 值；(iv) 通过受控错误率支持多种扰动；(v) 提供标量效应大小。我们通过多个案例研究展示了该框架的有用性，展示了如何量化响应变化、测量真/假阳性率，并评估与参考模型的对齐。最重要的是，我们将其视为一个可靠的 LLM 审计频率假设检验框架。

1. 介绍

大型语言模型（LLMs）通过对文本输入进行条件迭代采样生成输出，这些输出来源于一个 token 分布。因此，由于随机采样过程的存在，LLMs 的输出表现出固有的变化性，这一过程通过温度或 top-k (Renze & Guven, 2024) 等参数进行控制。这意味着评估干预措施——例如输入扰动、模型变体变化、提示系统修改——对输出的影响并不是一件简单的事情。理解和量化这种扰动的影响在高风险应用中至关重要——例如法律文件起草或医学诊断——因为错误或意外行为可能带来严重后果 (Meskó & Topol, 2023; Helberger et al., 2023)。

系统地评估输入扰动对输出响应的影响，是理解 LLM 行为的基础。它提供了关于模型在不同输入条件下的鲁棒性和输出一致性的定量洞察。可以使用模型响应

的统计假设检验来量化这些效果，并隔离随机变异的影响。在多个方面，这都是直接相关的。首先，它通过量化潜在的针对性攻击脆弱性来帮助识别脆弱性 (Goodfellow et al., 2014)。其次，它有助于偏差发现，其中潜在偏见或意外行为可能通过单一维度的审核方法不会显现出来 (Ribeiro et al., 2016)。第三，它可以在合规框架内发挥作用。用于评估模型行为的可量化框架对于遵循新兴的伦理和法律责任法规是必不可少的 (Doshi-Velez et al., 2017)。在语言模型的背景下，病史的细微变化可能导致截然不同的诊断；而拥有几乎相同健康记录的患者可能由于小的语法变化而接受截然不同的治疗计划。鉴于这些关键功能，显然需要一个围绕可靠统计工具的全面 LLM 审核框架。

当前用于分析大型语言模型行为变化的方法通常侧重于简化的指标，例如词语重叠或直接的对数概率比较。虽然在某些情况下这些方法有效，但它们未能考虑到大型语言模型处理的语义信息的细微且高维的性质 (Bhandari et al., 2020; Lin, 2004)。此外，现有的方法通常缺乏严格的统计基础，难以将模型行为的变化与输出生成过程中固有的随机性区分开来。为解决这一问题的努力包括专业的归因方法、特征重要性技术和反事实推理框架 (Ribeiro et al., 2016; Garg et al., 2019)。然而，这些方法往往是特定于某一模型的，依赖于对数据或模型的限制性假设，或者未能提供可解释和可推广的指标。

我们的解决方案。在这项工作中，我们引入了基于分布的扰动分析 (DBPA)，这是一个将大语言模型 (LLM) 扰动分析问题重新表述为频率假设检验任务的框架。DBPA 通过蒙特卡罗采样构建经验输出分布，以捕捉 LLM 固有的随机性，并在低维语义相似性空间内实现扰动效应的统计假设检验。通过使用统计假设检验，该框架能够进行稳健且可解释的推断，以判断输入扰动是否以及如何实质性地影响 LLM 输出。DBPA 是模型无关的，计算效率高，并且灵活到可以适应任何黑箱 LLM 的任意扰动。它还提供可解释的 p 值、标量效应大小，并支持具有控制误差率的多重检验，使其成为用于后验解释性和 LLM 可靠性评估的多功能工具。¹



Contributions . ① We identify limitations in existing methods for evaluating language model outputs (Sec. 2). ② We introduce distribution-based perturbation analysis which is a model-agnostic sensi-

¹ 剑桥大学. Correspondence to: Paulius Rauba <pr501@cam.ac.uk>.

¹代码可以在 <https://github.com/vanderschaarlab/dbpa> 找到

itivity technique that can test the effect of any perturbation with statistical significance measures (Sec. 3). ③ We perform multiple case studies to show the usefulness of DBPA (Sec. 4).

2. 通过频率论假设检验分析查看 LLM 输出

2.1. 问题表述

令 $\mathcal{X}$ 表示机器学习系统的输入空间, $\mathcal{Y}$ 表示输出空间。我们将该系统定义为一个随机映射 $\mathcal{S} : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$, 其中 $\mathcal{P}(\mathcal{Y})$ 是 $\mathcal{Y}$ 上的概率分布空间。这一表述捕捉到了现代机器学习系统, 包括大型语言模型 (LLMs) 所固有的随机性。

我们的目标是解决以下研究问题: 给定机器学习系统 $\mathcal{S}$, 一个输入 $x \in \mathcal{X}$, 以及一个输入扰动 Δ_x , 我们如何在一般的敏感性概念下系统地测量和解释 Δ_x 对 $\mathcal{S}(x)$ 输出分布的影响?

Definition 2.1 (Sensitivity). 机器学习系统 $\mathcal{S}$ 对输入扰动 $\Delta_x : \mathcal{X} \rightarrow \mathcal{X}$ 或系统扰动 $\Delta_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{S}$ 在输入 $x \in \mathcal{X}$ 上的敏感性由输出分布 $\mathcal{S}(x)$ 和 $\mathcal{S}(\Delta_x(x))$, 或 $\mathcal{S}(x)$ 和 $\Delta_{\mathcal{S}}(\mathcal{S})(x)$ 之间的非负差异度量 d 来描述。

我们旨在对黑箱 LLMs $\mathcal{S}$ 就系统和输入进行统计假设检验。因此, (i) 我们不假设可以访问底层系统的参数, 例如其权重或偏差; 并且 (ii) 我们不假设可以访问真实标签, 因为评估是基于系统生成的输出进行的。

2.2. 分布测试作为一个频率主义假设检验问题

评估 $\mathcal{S}$ 灵敏度的挑战可以重新构建为频率论的假设检验问题。一种常见的方法是查询语言模型一次以获得答案, 修改输入, 再次查询, 然后比较输出:

$$y \sim \mathcal{S}(x), \quad y' \sim \mathcal{S}(x'). \quad (1)$$

由于 LLM 系统的随机性, 这种方法存在高度方差, 因此统计效率低。因为 $\mathcal{S}(x)$ 为任何固定输入 x 生成一个随机变量, 所以 y 和 y' 代表输出分布中的单一实现。 y 和 y' 间观察到的任何差异因此可能来自固有的随机性, 而不是 Δ_x 的真实效果。

为了解决这个局限性, 我们建议从分布测试的角度重新审视这个问题。理想情况下, 我们希望比较整个输出分布, 而不是比较单个输出。为了方便表示, 令 $\mathcal{D}_x := \mathcal{S}(x)$ 表示输入 x 的输出分布。令 Y_x 和 $Y_{x'}$ 为随机变量, 其中 $Y_x \sim \mathcal{D}_x$ 和 $Y_{x'} \sim \mathcal{D}_{x'}$ 。

Definition 2.2 (Output Distribution). 对于给定的输入 $x \in \mathcal{X}$ 和一个 LLM $\mathcal{S} : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$, 输出分布 $\mathcal{D}_x$ 是从中得到的概率分布, 当由 $\mathcal{S}(x)$ 生成时随机变量 Y_x 被抽取。即, $Y_x \sim \mathcal{D}_x$ 。概率质量函数由以下给出:

$$p_x(y) = \mathbb{P}(Y_x = y), \quad \forall y \in \mathcal{Y}. \quad (2)$$

我们的目标是确定这些分布是否存在显著差异。这可以表述为一个假设检验:

$$H_0 : \mathcal{D}_x = \mathcal{D}_{x'} \quad (\text{The perturbation has no effect}) \quad (3)$$

$$H_1 : \mathcal{D}_x \neq \mathcal{D}_{x'} \quad (\text{The perturbation has an effect}) \quad (4)$$

这里, $x' := \Delta_x(x)$ 是扰动输入的简写。这样一种分布形式的主要好处在于它捕捉到 $\mathcal{S}$ 的完整随机行为, 而不仅仅是单一的实现。这意味着我们可以通过直接比较这些分布来进行统计推断, 并了解整个输出空间中的输出已经发生了多大的变化, 从而检测到可能不会从个别样本中显现的细微变化。

Takeaway . To understand the sensitivity of a language model to changes in input, we need to look at the entire range of possible outputs it can produce, not just single examples.

显然, 在语言模型的背景下直接使用这个框架存在挑战: 我们如何实际估计 $\mathcal{D}_x$ 或如何实际解释它? 我们在接下来的部分讨论这些问题。

在评估输入扰动对输出的影响时, 比较输出分布有两个主要挑战: 计算上的难以处理和解释性差。

► 挑战 1: 计算难解性。即使我们将语言模型限制为固定长度 L 的序列, 输出空间也是指数级的: $\mathcal{Y} = V^L$ 以及 $|\mathcal{Y}| = |V|^L$ 。因此, 对于给定输入 x , 其分布 $\mathcal{D}_x$ 完全由其概率质量函数描述。枚举每个 $y \in \mathcal{Y}$ 的 $p_x(y)$ 需要 $O(|V|^L)$ 次评估, 因此在实际操作中是不可行的。

► 挑战 2: 可解释性。第二个主要问题是分布 $\mathcal{D}_x$ 并没有提供对 LLM 输出的可解释理解。随着 LLM 被越来越多地用作推理引擎 (Yao et al., 2022; Hao et al., 2023), 我们关心的是它们的输出在语义上是否不同, 而不是在概率上。例如, 假设 $\mathcal{S}$ 对治疗建议的问题输出了两个答案

y_1 : “建议进行靶向放射治疗”, y_2 : “我们建议进行靶向放射治疗”。

在 p_x 下, 两个输出可能具有不同的概率, 即使在语义上, 它们传达了相同的建议。

理想情况下, 我们希望能够同时解决这两个问题——(i) 能够在计算上逼近分布, 以及 (ii) 评估这些差异是否在语义上有意义, 而不只是概率上不同。我们展示了如何通过有限样本逼近来实现这两点。

Takeaway . Analyzing output distributions of language models faces two unique challenges: the computational intractability due to the enormous output space, and the need for semantic rather than just probabilistic interpretation of differences.

2.3. 有限样本对输出分布的逼近

我们已经确定，使用分布来分析 LLM 结果与使用单个结果相比具有显著优势。这些优势带来了两个实际挑战：(i) 计算难以解决和 (ii) 解释性差。在此，我们建议使用有限样本近似可以解决这两个挑战。

► 解决挑战 1：计算复杂性。我们通过蒙特卡罗抽样来解决计算复杂性。我们定义输入 $x \in \mathcal{X}$ 及其扰动 x' 的输出分布的有限样本近似为：

$$\hat{D}_x = \{y_i\}_{i=1}^k, \quad y_i \stackrel{i.i.d.}{\sim} \mathcal{S}(x), \quad (5)$$

$$\hat{D}_{x'} = \{y'_i\}_{i=1}^k, \quad y'_i \stackrel{i.i.d.}{\sim} \mathcal{S}(x') \quad (6)$$

其中 k 是样本量。 k 的选择既影响估计量的方差（其比例为 $O(1/k)$ ），又影响后续假设检验的效能。蒙特卡罗样本的数量 k 可以视为一个超参数，并应根据问题的具体情况进行调整（对于相似性方差和统计效能的不同需求）。

► 应对挑战 2：可解释性。给定一组有限样本，我们希望衡量在输入扰动的情況下输出的变化程度。为了量化输入扰动引起的分布变化，我们引入了相似性度量 $s : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ 。该度量允许我们构建成对相似性的经验分布：

$$P_0 = \{s(y_i, y_j) : y_i, y_j \in \hat{D}_x, i \neq j\}, \quad (7)$$

$$P_1 = \{s(y_i, y'_j) : y_i \in \hat{D}_x, y'_j \in \hat{D}_{x'}\} \quad (8)$$

在这里， P_0 反映了原始输出分布中的内在变异性，而 P_1 反映了原始输出与扰动输出之间的跨分布相似性。因此，我们构建了两个相似性分布，作为敏感度（定义 ??）的代理，表示答案相似性的变异性。

□

Takeaway . Finite sample approximations using Monte Carlo sampling and pairwise similarity measures offer a computationally feasible and interpretable approach to analyzing output distributions of language models.

3. 基于分布的扰动分析

我们提出了一种模型无关的方法，用于评估大型语言模型（LLM）对抗扰动的敏感性。我们的方法避免了限制性的分布假设，并利用了 $\mathcal{S}$ 的整个输出分布，捕捉 LLM 响应中的内在变异性。通过构建原假设和备择假设的分布，我们可以使用 p 值进行频率统计假设检验。重要的是，我们的方法框架适用于任何扰动和任何语言模型，最低要求是能够从语言模型的输出分布中采样并构建嵌入。此后，我们假设嵌入函数是确定的和稳定的，同时嵌入保持语义不变：如果两个输出被人

类判断为同样相似，那么它们的嵌入之间的距离应该相等，至多单调变换。

3.1. 该过程

目标。通过这种方法，我们能够评估两个值。首先，我们计算效应大小（“反应分布发生了多大的变化？”）。其次，我们计算 p 值（“新旧分布之间的差异是否显著？”）。我们遵循一个简单的程序，其一般形式如下所述。

Distribution-based perturbation analysis: a quick overview of the procedure

Distribution-based perturbation analysis proceeds in four steps: response sampling, distribution construction, distributional comparison, and statistical inference.

I. Response Sampling . Draw k independent outputs from the original prompt and k from the perturbed prompt

$$\hat{D}_x = \{y_i\}_{i=1}^k, \quad \hat{D}_{x'} = \{y'_i\}_{i=1}^k,$$

where $y_i \stackrel{i.i.d.}{\sim} \mathcal{S}(x)$ and $y'_i \stackrel{i.i.d.}{\sim} \mathcal{S}(x')$ with $x' := \Delta_x(x)$. Define the pooled vector $Z = (z_1, \dots, z_{2k})$ with

$$z_i = y_i \quad (1 \leq i \leq k), \quad z_{k+i} = y'_i \quad (1 \leq i \leq k).$$

II. Distribution construction . Using a similarity function $s : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, build

$$P_0 = \{s(y_i, y_j) : 1 \leq i < j \leq k\}, \\ [2pt] P_1 = \{s(y_i, y'_j) : 1 \leq i, j \leq k\}.$$

III. Distributional comparison . Measure the discrepancy between P_0 and P_1 with any non-negative functional

$$\omega : \mathcal{P} \times \mathcal{P} \longrightarrow \mathbb{R}_{\geq 0}, \quad T_{\text{obs}} = \omega(P_0, P_1).$$

IV. Statistical inference . Formulate the hypotheses

$$H_0 : \mathcal{S}(x) = \mathcal{S}(x'), \quad H_1 : \mathcal{S}(x) \neq \mathcal{S}(x').$$

We can evaluate this hypothesis via a simple permutation procedure described in Algorithm 1.

Objective . If $\hat{p}$ is small, this suggests that T_{obs} is unusually large relative to its null distribution. The value T_{obs} itself serves as the effect-size estimate, whereas the permutation test provides frequentist p -values (Knijnenburg et al., 2009; Phipson & Smyth, 2010).

然后，我们可以通过置换检验来进行统计推断。该过程简单，并且对原始输出仅需最小的可交换性假设。由于相似性分数是相关的，通过置换它们无法保证可交换性。

Assumption 3.1 (Exchangeability Under the Null Hypothesis). 在原假设 $H_0 : \mathcal{D}_x = \mathcal{D}_{x'}$ 下，合并的输出 $Z = (y_1, \dots, y_k, y'_1, \dots, y'_k)$ 是可交换的。也就是说，对于指数集合 $\{1, \dots, 2k\}$ 的每一个排列 π ，我们有

$$(y_1, \dots, y_k, y'_1, \dots, y'_k) \stackrel{d}{=} (y_{\pi(1)}, \dots, y_{\pi(2k)}).$$

在这个假设下，我们可以将置换检验的算法形式化如下。

在严格的零假设条件下，任何池化结果的函数是可交换的，因此我们的基于排列的 p 值是有效的。我们选择仅通过语义相似度函数 s （例如，对句子嵌入的余弦）来汇总输出。因此，该检验仅对相似度得分中出现的差异敏感。拒绝意味着可以检测到的意义转变；不拒绝意味着“在 s 下没有检测到的语义转变”，但并不意味着完整的文本生成分布是相同的。关于为何这在章节 ?? 的挑战 2 中是可取的，我们进行了讨论。

Algorithm 1 基于分布的扰动分析的排列检验

Require: Pooled vector $Z = (z_1, \dots, z_{2k})$, similarity function s , discrepancy measure ω , number of permutations B

Ensure: p -value $\hat{p}$

```

1: Compute observed test statistic:
2:  $P_0 = \{s(z_i, z_j) : 1 \leq i < j \leq k\}$ 
3:  $P_1 = \{s(z_i, z_j) : 1 \leq i \leq k < j \leq 2k\}$ 
4:  $T_{\text{obs}} = \omega(P_0, P_1)$ 
5: Initialize counter: count = 0
6: for  $\pi = 1$  to  $B$  do
7:   Randomly select subset  $I \subset \{1, \dots, 2k\}$  of size  $k$ 
8:    $O = \{z_i : i \in I\}$ ,  $P = \{z_j : j \notin I\}$ 
9:    $P_0^{(\pi)} = \{s(u, v) : u, v \in O, u \neq v\}$ 
10:   $P_1^{(\pi)} = \{s(u, v) : u \in O, v \in P\}$ 
11:   $T^{(\pi)} = \omega(P_0^{(\pi)}, P_1^{(\pi)})$ 
12:  if  $T^{(\pi)} \geq T_{\text{obs}}$  then
13:    count = count + 1
14:  end if
15: end for
16:  $\hat{p} = \frac{1 + \text{count}}{1 + B}$ 
    return  $\hat{p}$ 

```

3.2. 从理论到实践：分布式扰动分析的设计选择

在实施所述程序时，需要考虑一些实用的实施要点。

► 为什么使用标量成对余弦相似度而不是直接使用高维嵌入？使用未简化的嵌入来构建零和备择分布面临两个主要障碍。(a) 高维性。嵌入空间通常有数百或数

千个维度，因此在该空间中估计完整分布是不可行的，因为维数诅咒的影响。(b) 多余的语义信息。嵌入捕捉了丰富的语义——其中大部分是无关的，当唯一的目的是量化响应在输入扰动时移动的程度，而不是每个响应在嵌入流形中的位置时。将每对嵌入简化为一个相似度得分可以规避这两个问题。

► 什么是合适的相似性函数？我们使用一个相似性函数而不是明确的降维映射，该函数将两个输出映射到 $\mathbb{R}$ 。在几个可行的选择中（例如，负的 ℓ_1 或 ℓ_2 距离），我们采用余弦相似性

$$s(y_i, y_j) = \frac{\langle e(y_i), e(y_j) \rangle}{\|e(y_i)\| \|e(y_j)\|},$$

，其中 $e(\cdot)$ 是一个假定的确定性、稳定的嵌入函数（在大多数实验中为 ada-002）。由于三角不等式失效，余弦相似性不是一个度量。然而，我们只需要一个保持意义的、有界的相似性，而不是一个度量结构。

► 什么是合适的距离度量 ω 呢？之所以存在对 ω 的选择，是因为我们正处理两个分布之间的比较。这不同于传统的基于重抽样的方法，这些方法构建一个零分布，并针对单个实例进行评估 (Yu, 2003)。虽然 ω 的选择可能会根据应用的不同而有所变化，但我们采用詹森-香农散度 (JSD) 作为 ω 的度量： $\text{JSD}(P_0 \| P_1) = \frac{1}{2} (D_{\text{KL}}(P_0 \| M) + D_{\text{KL}}(P_1 \| M))$ ，其中 $M = \frac{1}{2}(P_0 + P_1)$ 和 D_{KL} 是 Kullback-Leibler 散度。我们将每个多重集合转换为一个经验概率度量，并在这些经验分布之间计算 JSD。这是因为 JSD 具有三个用于评估分布的有用性质：(i) 对称性，确保度量对比较的分布顺序不变；(ii) 有界性，为不同输入和扰动的解释提供一致的尺度；以及 (iii) 对分布的位置和形状差异的敏感性。

为什么执行基于排列的测试而不是直接从 $\mathcal{S}$ 进行蒙特卡罗抽样？原则上，可以通过重复抽样 $\mathcal{S}(x)$ 和 $\mathcal{S}(\Delta_x(x))$ 来近似虚无分布。然而，新的模型调用通常是主要的成本。排列测试重新使用相同的 $2k$ 生成的输出，但是——因为组合向量在 H_0 （假设 ??）下是可交换的——可以用更少的查询提供一个有效的 p 值。我们将在附录 A 中对此进行扩展。

► 我们在测试许多扰动时应该如何控制多重性？当 DBPA 应用于一组扰动 $\{\Delta_i\}_{i=1}^m$ 时，每个假设

$$H_0^{(i)} : \mathcal{S}(x) = \mathcal{S}(\Delta_i(x))$$

都通过其自身的置换 p 值 $\hat{p}_i$ 进行评估。并行运行 m 个独立测试会增加至少一个 I 类错误的概率。为了将全家错误率 (FWER) 保持在期望的水平 α ，可以默认采用 Bonferroni 校正：

$$\text{FWER} = 1 - (1 - \alpha)^m \approx m\alpha, \quad \alpha_{\text{adj}} = \frac{\alpha}{m}, \quad p_{\text{crit}} = \frac{\alpha}{m}.$$

一种常见的对策是增加用于构建经验分布 P_0 和 P_1 的蒙特卡罗样本量 k 。在通常的正态近似下，

$$k_{\text{adj}} \propto \left(\frac{z_{1-\alpha/(2m)} + z_{1-\beta}}{\omega} \right)^2,$$

Experiment	Purpose	Finding	Name	Case study	Insight
Persona Perturbation	Test role instruction effect	Large LMs stable on medical roles; small LMs variable	Table 1	✓	✗
TPR/FPR Tradeoff	Measure TPR/FPR in healthcare	Model choice depends on FPR threshold	Fig. 2	✓	✗
Model Alignment	Compare outputs to GPT-4	Most models align; some small LMs differ	Fig. 3	✓	✗
Input Length Impact	Assess effect of prompt length	Longer contexts reduce detectability	Table A.1	✗	✓
Alt. Distance Measures	Compare JSD, Euclid., etc.	Significance decisions consistent across measures	Table A.2	✗	✓
Embedding Costs	Estimate embedding expenses	Perturbations remain inexpensive even at scale	Table A.3	✗	✓
Text Similarity Alt.	Use BLEU/ROUGE in DBPA	Highly sensitive to small changes; less reliable	Table A.4	✗	✓
Impact of Embeddings	Test various embedding models	Creative personas' effects consistent across embeds	Table A.5	✗	✓

Table 1: 实验和案例研究的简要总结。我们进行了三个案例研究（主要部分）和五个附加的见解研究（附录）以了解我们的框架可能在何时、为何以及如何有用。

，其中 ω 表示目标效应大小， z_q 是第 q 个标准正态分位数， β 是可容忍的 II 类错误。因此，实践者可以通过增加额外的模型调用来恢复统计检验的能力，同时仍然享有严格的 FWER 控制；当适当时，可以替换为不那么保守的多重性调整方法。

4. 案例研究

我们在多种用例中展示了我们方法的有效性。在接下来的小节中，我们将展示我们的方法可以：(1) 捕捉那些显著的和并不显著的回答分歧 (2) 分析语言模型对提示中无关变化的鲁棒性 (3) 评估与参考语言模型的一致性。默认情况下，我们在 5 个种子上运行实验，并报告测量值的平均值和标准差。我们计算距离度量 ω ，该度量是作为零分布和备选分布之间的 JSD 距离计算的，同时计算 p 值。

4.1. 案例研究 1：在不同的输入扰动下，响应变化有多大？

语言模型被认为具有很强的角色扮演能力，会影响它们的回应 (Wang et al., 2023; Kong et al., 2023; Chen et al., 2024)。我们利用这一特性来评估在不同分配角色下回应的变化。

• **设置** . We evaluate whether we can capture LLM response variability to prompt perturbations. We compare response distributions before and after a prompt perturbation. First, we query an LLM with a healthcare question to establish a null distribution from the baseline responses. Then we pre-pend a role-playing instruction of the form ‘Act as { role }’, where { role } describes a specific given role. We divide the setups into medical professions (medical supervisor, therapist, doctor, medical student) and other roles (comedian, robot from the future, neurips reviewer, child). We then quantify the effect size ω and the p-value. We run the case study with larger models (GPT-4 and GPT-3.5) as well as smaller open-source models which are generally considered less capable.

目标 . We aim to showcase that we can quantify the shift in the output distribution by pre-pending

a different role to a language model and that this shift is both model and prompt dependent.

案例研究 1 的结果展示在表 2 中。鉴于提示中的场景是医疗场景，我们期望前四个人物角色（代表医疗职业）的输出分布相似性没有差异，而其他四个人物角色则应该存在这样的差异。我们看到，这在较大模型 (GPT-4 和 GPT-3.5) 中是成立的，其中医疗人物角色的效应大小均不显著，而大多数非医疗人物角色则显著。

另一方面，较小的模型表明人格与输出分布变化之间没有这样的关系。我们将其解读为较小的模型表现出更高水平的答案多样性，并且通常比更大、更强大的模型更不一致/对特定指令不敏感。事实上，我们认为这一分析展示了为何使用较小的模型会不合适，因为通过添加人格而获得的响应与我们的预期相差太大。改变蒙特卡罗样本数量是使估计更精确的一种方法。

这些结果直接与我们关于审计稳健性的主要问题相关。我们观察到，较小的模型在角色方面表现出更大的变异性，而直观上，这些角色不应影响输出响应。因此，我们建议较小的模型比较大的模型更不稳健。

4.2. 案例研究 2：我们能否测量真正阳性和假阳性反应的变化？

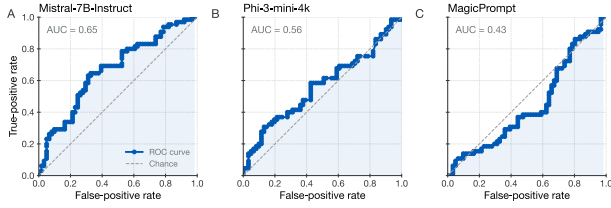
通过量化 TPR 和 FPR 来测量语言模型的鲁棒性。我们调整框架，通过平衡 LLM 响应的真实率和错误率，设计一种新的语言模型敏感性度量。设 $\{\Delta_i\}_{i=1}^m$ 为一系列扰动数据，分为控制指标 C （在扰动下输出不应该变化）和目标指标 T （在扰动下输出应该变化）。对每个 i ，定义 $x'_i = \Delta_i(x)$ ，并设 p_i 为比较 $\mathcal{D}_x$ 和 $\mathcal{D}_{x'_i}$ 的置换检验 p 值。确定一个显著性水平 α 。如果 $p_i < \alpha$ 则对于扰动 i 我们声明“检测到变化”。然后 $\text{FPR} := \frac{1}{|C|} \sum_{i \in C} \mathbb{I}[p_i < \alpha]$ 和 $\text{TPR} := \frac{1}{|T|} \sum_{i \in T} \mathbb{I}[p_i < \alpha]$ 。

显然，这可以通过改变 $\alpha \in [0, 1]$ 来扩展。首先，这允许用户在给定的期望错误水平下分析 FPR 和 TPR 之间的权衡。其次，它使得通过用一个标量总结性能的方式，获得模型对 FPR 和 TPR 敏感性的整体视图。在获得数据集 $(\text{FPR}(\alpha), \text{TPR}(\alpha))$ 对于 $\alpha \in [0, 1]$ 之后，LLM 总体权衡由简单的积分给出

$$\text{AUC} = \int_0^1 \text{TPR}(\alpha) d[\text{FPR}(\alpha)].$$

Table 2: 基于第 4.1 节所描述的情境，针对所有模型按角色估算的效应量 ω 。星号表示双侧显著性水平：* $p < 0.05$ ，** $p < 0.01$ ，*** $p < 0.001$ 使用算法 1 计算得出。效应量越高，表示在选定的蒙特卡罗样本数下，输出分布已经发生了更大的偏移。作为一个有用的参考，我们期望在四个医疗角色（表格顶部）下输出保持稳定，但在其他四个角色（底部）下发生偏移。

Persona	Large models		Smaller models						
	GPT-4	GPT-3.5	LLama-3.1-8B	GPT-2	Gemma-2-9B	SmolLM-135M	Phi-3-mini	MagicPrompt	Mistral-7B
Doctor	0.22	0.21	0.20	0.20	0.41 **	0.14	0.16	0.17	0.31 *
Nurse	0.20	0.20	0.25	0.30 *	0.24 *	0.13	0.22	0.13	0.24
Medical Practitioner	0.25	0.19	0.25	0.33 **	0.30 *	0.29 **	0.18	0.20	0.20
Medical Supervisor	0.22	0.17	0.20	0.19	0.17	0.17	0.25	0.26	0.20
Comedian	0.32 ***	0.24	0.36 ***	0.26	0.23	0.12	0.23	0.24	0.20
Robot From The Future	0.15	0.26 *	0.31 **	0.23	0.18	0.16	0.23	0.25	0.24
NeurIPS Reviewer	0.28 **	0.28 *	0.30 **	0.17	0.19	0.13	0.24	0.17	0.28
Child	0.27	0.33 **	0.19	0.27 *	0.14	0.28 **	0.20	0.19	0.37 **



*Higher curves (upper left) indicate better discriminatory power

Figure 1: 三个选定语言模型的 TPR/FPR 权衡。图中显示了三个面板：A、B 和 C，描述了三个模型的 TPR 和 FPR 之间的关系，以及基于第 4.2 节的方法评估的 AUC 分数。ROC-AUC 描述了随我们改变 $\alpha \in [0, 1]$ 时模型性能的变化。具有更高 AUC 的模型更好，因为它们能够检测真实变化并在与无关变化的情况下不回答不同。

就我们的目的而言，较高的 AUC 表明在用户定义的目标和控制设定中，LLM 能更好地区分目标扰动与控制扰动。在实践中，我们通过对不同阈值水平进行标准的梯形求和来计算它。现在，我们通过实验证明如何利用这一点来评估 LLMs。

- **设置** . We create healthcare prompts with patient varying patient features sampled from a distribution and query the LLMs for assessing the cardiovascular disease (CVD) recommendations for the given patient based on the NICE guidelines. We create two types of perturbations: (1) control perturbations that should not change medical recommendations (e.g. changing the patient's name or modifying a covariate that should not affect the guidelines) and (2) target perturbations that should change recommendations (e.g. changing a patient's cholesterol or diabetes information to move them into a different risk level which requires a different

recommendation). For each perturbation type, we calculate the p-values and obtain TPR and FPR rates.

- **目标** . We aim to show that our method can be used to assess LLM reliability for perturbations by quantifying their true positive and false positive rates for custom-designed scenarios.

我们进行了 TPR/FPR 分析，并在图 1 中展示了三个模型。我们发现，并非所有的模型在选定的情况下都同样强大。在我们的示例中，我们展示了表现最佳、中等和最差的模型。希望在 TPR/FPR 上获得最高 ROC-AUC 的人应选择图 A 中的模型。然而，我们发现虽然这是选择单一模型的直观方法，但最佳模型可能取决于具体的错误率。例如，在图 2 中，我们展示了在同一示例中，对于允许的 1% 的 FPR，我们会选择 Phi-3，而允许的 5% 的 FPR 会建议选择 Mistral-7B。

显然，所选择的模型或设置没有什么特别之处。这种分析可以在任何选定的语言模型和任何设置中进行。对最终用户的重要要求是指定在何时语言模型应该（或不应该）更改其响应。因为结果依赖于模型和问题设置，我们不认为仅仅根据 AUC 分数得出有意义的解释，并建议仅在评估多个模型或系统时，将其作为一个比较系统。

4.3. 案例研究 3：我们能用参考语言模型评估对齐吗？

到目前为止，我们在保持模型不变的情况下更改了输入提示。我们也可以反过来——在改变模型的同时保持提示不变。这与第 2.1 节中的两种表述相符。我们可以将此用作评估不同语言模型之间对齐度的方法。结果展示在图 3 中。

- **设置** . We evaluate alignment between two language models by sampling responses from a given language model and fixing the input prompt but

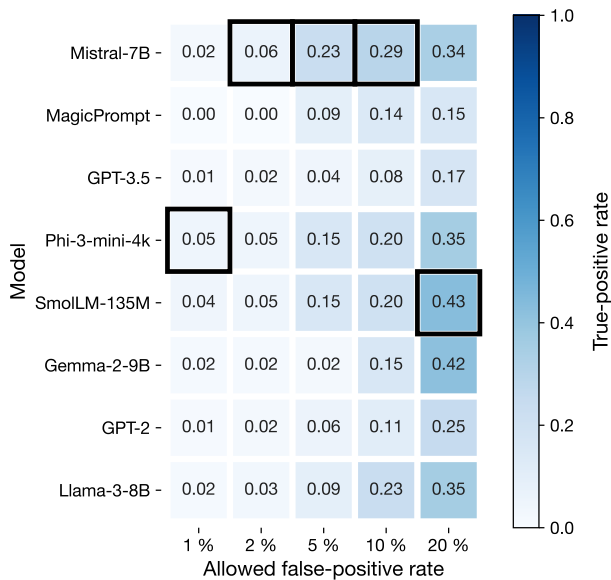


Figure 2: 不同模型在选定 FPR 下的 TPR。我们展示了如何在选定的 FPR 下选出 TPR 最高的最佳模型。允许的假阳性率下的最佳模型已在黑框中突出显示。我们发现，选择最佳模型可能依赖于允许的假阳性率。在给定 FPR 下，较高的 TPR 会更好。

swapping out the model. Then, the differences between the output distributions correspond to the distance between model responses. We evaluate this across multiple open- and closed-source models.

- **目标** . We aim to show that we can understand how well two language models answer across their response distributions.

图 3 的结果表明，我们可以量化模型间响应分布的差异，从而提供一种具体的跨模型对齐测量方法，可能对语言模型输出的比较分析有用。鉴于使用的蒙特卡罗样本数，只有两个模型——SmolLM 和 Llama-3.1——与其他模型的响应显著不同，其他模型展示的差异较小。显然，何为“更好”取决于使用场景：如果我们旨在确保语言模型之间的输出不发生偏差，我们希望获得较低的 ω ；反之，如果我们期望输出发生变化，例如在对语言模型应用奖励建模的迭代之后，则情况正好相反。

4.4. 其他评价

以上个案研究展示了我们认为假设检验用于审计语言模型鲁棒性最重要的使用案例。在这里，我们展示了一些其他使用案例并介绍我们进行的一些分析。语言模型在处理长上下文时表现不佳，这一现象被称为“中途丢失” (Liu et al., 2023)。我们研究了更长输入长度对我们的 Persona 实验 (第 C.1 节) 的影响，发现较长

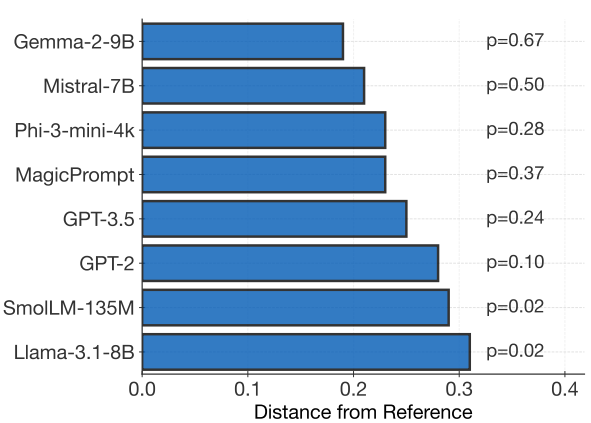


Figure 3: 与参考语言模型的响应距离。我们从一个基础模型 (GPT-4) 收集 k 个响应，并评估左侧报告的多个其他语言模型的响应，使用算法 1 计算它们的 ω 和 p-值。更高的距离意味着相似性分布的差异更大。

上下文窗口中由于输入扰动而检测变化更为困难。我们评估了替代距离度量 (第 C.2 节) 并发现评估对于所选择的距离度量较为敏感，但在效果量的显著性方面保持相对一致。由于该框架依赖于蒙特卡罗抽样和嵌入模型，我们评估了使用各种嵌入模型的成本，发现即使在大规模下，进行此类扰动分析的成本依旧高效 (第 C.3 节)。我们将我们的模型与不在语义空间中操作的其他传统替代方法进行比较，如 BLEU 或 ROUGE 指标 (第 C.4 节)，并评估使用不同嵌入模型的效果 (第 C.5 节)，普遍发现结果对于计算 p 值和效果量的嵌入较为敏感。

5. 相关工作

有三种主要方法来量化与我们方法相关的基于文本的输出。这些方法包括测量模型输出中无意的偏见，开发反事实公平方法，以及文本摘要度量。

► **测量非预期偏差**。总体而言，与之最相关的工作是在测量非预期偏差 (Borkan et al., 2019; Dixon et al., 2018; Park et al., 2018)。这些指标量化了模型在子群体之间的现有偏差。这个领域更广泛的工作可以在 (Friedler et al., 2021; Kleinberg et al., 2016; Menon & Williamson, 2018) 中找到。然而，这需要人工标注，仅涉及公平性，并假定各子群体之间存在可靠的标签。

► **反事实公平性**。这种方法 (Garg et al., 2019) 检查如果敏感属性不同，预测结果将如何改变。它可以计算效应大小，但不能应用于黑箱模型，不允许任意的扰动，也不支持统计推断。需要人工输入来标记答案的特定属性 (例如，毒性)，并作出一些假设，例如，非毒性的例子相对于有毒的例子更不容易包含不对称的反事实。

► **文本摘要指标**。这些指标 (Bhandari et al., 2020; Zhang

et al., 2019; Zhao et al., 2019; Lin, 2004) 评估文本摘要的质量。它们可以计算效应量,但不适用于黑箱模型,不允许任意扰动,也不支持统计推断。它们需要人工输入并做出某些假设。各种指标如 BERTScore、MoverScore 和 ROUGE 变体衡量系统摘要与参考摘要之间相似性的不同方面。

为了更好地解释我们的不同之处,我们根据五个重要标准对每个领域进行比较:(i) 该方法是否可以应用于任何黑箱模型;(ii) 是否可以应用和测量任何扰动;(iii) 该方法是否支持统计推断;(iv) 该方法是否允许计算变化的效应量;(v) 是否有任何假设;(vi) 是否需要人类作为输入的一部分。我们在表 8 中展示了这一点。

6. 讨论

随着对评估 LLM 系统需求的增长,我们需要基于统计的方法来理解 LLM 的输出。我们知道,我们不能简单地询问语言模型在给定情况下它们的答案是否会有所不同,因为它们可能并不真实(Si et al., 2023)。那么,有没有方法可以系统性地评估模型的响应如何影响输出?

在我们的论文中,我们提供了一种实现这一目标的方法。具体来说,我们认识到,为了完全分析语言模型的响应,必须考虑 LLM 响应的整个输出分布。因此,比较单个的两个点估计是不足够的。为了解决这个问题,我们引入了基于分布的扰动分析,作为一种使用蒙特卡罗采样和基于频率推断的方法来量化两个输出之间差异的方法。重要的是,我们的模型具有理想的统计特性,使其在许多应用领域中具有直接的实用性,例如与模型无关、支持用户可以指定的任何输入扰动、提供统计估计、为多重扰动测试提供控制误差率,以及通过效应量比较不同扰动之间的关系。

我们看到这在高风险领域具有直接用途,例如模型审计、模型透明性、合规性,或者在语言模型需要可解释性并作为更广泛统计系统的一部分使用的领域。此外,假设检验框架的核心组件是能够为给定的用户问题设置选择和量化真阳性率和假阳性率的能力。这提供了极大的灵活性,可以为期望的假阳性率选择最佳模型,这种设置在高风险环境中极为有用。

限制。本文首次提出使用频率假设检验来审计语言模型的方法。然而,本文的本质在其范围内是概念性的。未来的工作应增加更有力的实证证据,并在其他领域测试该方法。关键的设计选择——如距离度量、嵌入函数及其对泛化的影响——将从更深入的分析 and 应在未来工作中建立的更明确的选择指南中获益。此外,尽管 DBPA 为模型审计提供了宝贵的见解,将这些发现转化为增强模型稳健性并使输出与人类偏好对齐的实际策略仍然是一个重要的挑战。

对于解释黑箱行为并提供保证的可解释和基于统计的系统的需求正在增长。我们希望我们的工作可以成为

朝这个方向迈出的第一步,并能在理论和实际应用中得到进一步的拓展。

7.

影响声明 本论文的目标是推动机器学习领域中的可解释性。我们将广泛的评估问题重新构建为正式的假设检验，从而为专业人员、政策制定者和独立审计员提供一种统计上有原则的工具，用于检测大型语言模型中真实行为变化。我们认识到这可能对语言模型系统的安全性和审核具有许多积极影响。同时，我们也意识到双重使用风险，例如对手使用程序中的逻辑来逃避检测或优化对抗攻击。然而，我们认为在现实中这样的风险极小。

References

- Bhandari, M., Gour, P., Ashfaq, A., Liu, P., and Neubig, G. Re-evaluating evaluation in text summarization. arXiv preprint arXiv:2010.07100, 2020.
- Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In Companion proceedings of the 2019 world wide web conference, pp. 491–500, 2019.
- Chen, N., Wang, Y., Deng, Y., and Li, J. The oscars of ai theater: A survey on role-playing with language models. arXiv preprint arXiv:2407.11484, 2024.
- Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73, 2018.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O’Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al. Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134, 2017.
- Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Communications of the ACM, 64(4):136–143, 2021.
- Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual fairness in text classification through robustness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 219–226, 2019.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
- Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. arXiv preprint arXiv:2305.14992, 2023.
- Helberger, N., Diakopoulos, N., et al. Chatgpt and the ai act. Internet Policy Review, 12(1):1–6, 2023.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.
- Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. Fewer permutations, more accurate p-values. Bioinformatics, 25(12):i161–i168, 2009.
- Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. Better zero-shot reasoning with role-play prompting. arXiv preprint arXiv:2308.07702, 2023.
- Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74–81, 2004.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172, 2023.
- Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In Conference on Fairness, accountability and transparency, pp. 107–118. PMLR, 2018.
- Meskó, B. and Topol, E. J. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. NPJ digital medicine, 6(1):120, 2023.
- Park, J. H., Shin, J., and Fung, P. Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231, 2018.
- Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Statistical applications in genetics and molecular biology, 9(1), 2010.
- Renze, M. and Guven, E. The effect of sampling temperature on problem solving in large language models. arXiv preprint arXiv:2402.05201, 2024.
- Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, 2016.
- Romero-Alvarado, D., Hernández-Orallo, J., and Martínez-Plumed, F. How resilient are language models to text perturbations? In International Conference on Intelligent

- Data Engineering and Automated Learning , pp. 85–96. Springer, 2024.
- Schubert, E. A triangle inequality for cosine similarity. In International Conference on Similarity Search and Applications , pp. 32–44. Springer, 2021.
- Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daumé III, H., and Boyd-Graber, J. Large language models help humans verify truthfulness—except when they are convincingly wrong. arXiv preprint arXiv:2310.12558 , 2023.
- Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746 , 2023.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 , 2022.
- Yu, C. H. Resampling methods: concepts, applications, and justification. Practical Assessment, Research & Evaluation , 8(19):1–23, 2003.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 , 2019.
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622 , 2019.

A. 蒙特卡罗采样

为了理解为什么我们使用蒙特卡罗采样，让我们考虑在这两种方法下 ω 的期望值。对于置换检验，这产生：

$$\mathbb{E}[\omega(P_0^*, P_1^*)] = \mathbb{E}[\omega(\{s(z_i, z_j)\}, \{s(z_i, z_k)\})] \quad z_i, z_j, z_k \stackrel{i.i.d.}{\sim} \mathcal{Z} \quad (9)$$

其中 $\mathcal{Z} = \{y_i, \dots, y_k, y'_1, \dots, y'_k\}$ 是 $S(x)$ 和 $S(\Delta(x))$ 的 $2k$ 个样本的结合集。对于重采样，期望是：

$$\mathbb{E}[\omega(P'_0, P'_1)] = \mathbb{E}[\omega(\{s(y_i, y_j)\}, \{(y_i, y'_k)\})] \quad y_i, y_j \stackrel{i.i.d.}{\sim} \mathcal{S}(x), y'_k \stackrel{i.i.d.}{\sim} \mathcal{S}(\Delta(x)) \quad (10)$$

在 H_0 条件下， $\mathcal{S}(x) = \mathcal{S}(\Delta(x))$ ，因此 $\mathcal{Z}$ 中的所有样本遵循同一分布。因此，从 $\mathcal{Z}$ 中随机抽样等同于从 $\mathcal{S}(x)$ 或 $\mathcal{S}(\Delta(x))$ 中抽样。因此，期望值是相同的，并且置换检验在极限情况下将产生相同的 ω 值分布。

B. 提示输入示例

B.1. 治疗建议的示例输入

在治疗建议的练习中，我们使用以下输入提示（其中特征可以有所不同）：

约翰有以下特征：年龄：70，BMI：30，血压：114/61 mmHg，总胆固醇：6.7 mmol/L，HDL 胆固醇：1.0 mmol/L，吸烟状况：目前吸烟者，糖尿病：无糖尿病，心血管疾病家族史：有，种族：华裔。仅基于这些特征，提供基于 NICE 的心血管疾病指南建议。

John 具有这些特征：年龄：70 岁，BMI：29，血压：114/61 mmHg，总胆固醇：6.7 mmol/L，HDL 胆固醇：1.0 mmol/L，吸烟状态：当前吸烟者，糖尿病：无糖尿病，CVD 家族史：有，种族：华人。仅根据这些特征，基于 NICE 为此人提供 CVD 指南的建议。

B.2. 治疗建议的示例输出

以下是示例输出：

```
1 ["Based on these characteristics, NICE guidelines for CVD recommend the following:\n\n
  n1. Encouraging John to quit smoking and providing support and resources to help him
  do so.\n2. Promoting a healthy diet and regular physical activity to help lower his
  BMI and improve his cholesterol levels.\n3. Monitoring and managing his blood pressure
  to keep it within a healthy range.\n4. Considering his family history of CVD, it may
  be advisable to assess his overall cardiovascular risk and consider interventions to
  reduce that risk.\n5. Providing culturally sensitive and tailored support for his
  Chinese ethnicity, taking into account any specific risk factors or considerations.\n\n
  nIt's important for John to work with his healthcare provider to develop a
  personalized plan for cardiovascular health and to address any individualized risk
  factors or concerns.",
2 "Based on the provided characteristics, John's CVD risk would be assessed as moderate to
  high. The NICE guidelines recommend the following for individuals with similar
  characteristics:\n\n
  n1. Lifestyle changes: John should be advised to make lifestyle
  changes such as quitting smoking, following a healthy diet to reduce cholesterol
  levels and maintain a healthy weight, and engage in regular physical activity.\n\n
  n2. Assess CVD risk: John should undergo a formal assessment of CVD risk using a
  cardiovascular risk assessment tool, taking into account his age, BMI, blood pressure,
  cholesterol levels, smoking status, and family history of CVD to determine the need
  for medication or further interventions.\n\n
  n3. Regular monitoring: John should have
  regular monitoring of his blood pressure, cholesterol levels, and overall
  cardiovascular health to manage and reduce his risk of CVD.\n\n
  n4. Consider medication: Depending on the results of his CVD risk assessment, John may need to be considered
  for medication such as statins to lower his cholesterol levels and reduce his overall
  CVD risk.\n\n
  nIt's important for John to work closely with a healthcare professional to
  develop a personalized plan for managing his CVD risk based on the NICE guidelines.",
3 "Based on the provided characteristics, NICE guidelines for CVD prevention would
  recommend the following:\n\n
  n1. Encouraging John to quit smoking and providing support
  and resources to help him quit.\n2. Recommending lifestyle modifications such as
  regular physical activity and a healthy diet to help manage his BMI and cholesterol
```


levels.\n3. Monitoring and managing his blood pressure to keep it within a healthy range.\n4. Consideration of cholesterol-lowering medication if lifestyle modifications are not sufficient to achieve target cholesterol levels.\n5. Regular monitoring and management of other CVD risk factors, especially considering his family history of CVD.\n\nIt's important for John to work closely with his healthcare provider to create a personalized plan for CVD prevention and management based on his individual risk factors and health needs.",

4 "Based on the provided characteristics, the following recommendations can be made for CVD guidelines based on NICE:\n\n1. Lifestyle changes: John should be advised to quit smoking and adopt a healthier lifestyle, including regular physical activity and a balanced diet to help lower his BMI and cholesterol levels.\n\n2. Cardiovascular risk assessment: John's overall cardiovascular risk should be assessed using a validated tool, taking into account his age, blood pressure, and cholesterol levels.\n\n3. Medical assessment: John should be assessed for any potential risk factors for cardiovascular disease, such as high blood pressure, and be offered appropriate management and treatment as necessary.\n\n4. Family history: Given John's family history of CVD, he should be counseled on the potential genetic predisposition to cardiovascular disease and the importance of managing other risk factors effectively.\n\n5. Ethnicity: Being of Chinese ethnicity, John should be aware that certain populations may have a higher risk for specific cardiovascular conditions, and this should be taken into consideration when assessing his overall cardiovascular risk.\n\nIt is important to note that these recommendations are general and may need to be tailored specifically to John's individual health needs and circumstances. It is advisable for John to seek personalized medical advice and undergo a comprehensive cardiovascular risk assessment and management plan under the care of a healthcare professional.",

C. 扩展实验

C.1. 提示长度的影响

实验设计。为了解决关于 DBPA 在更长和更复杂的输入序列上有效性的问题，我们进行了额外的实验，将提示长度从 100 个令牌变化到 900 个令牌，并使用不同的人设指令。我们重复使用相同的八种不同的人设，并改变输入提示的长度。

主要结果。我们发现，对于超过 800 个标记的情况，个人标记数量并未显示出输出分布的任何显著变化。观察到的模式是在较短提示中效果更强，而在较长序列中逐渐减弱，这与大型语言模型中记载良好的经验现象一致，其中扩展的上下文可能导致更一致但潜在上不那么独特的输出。无论如何，这种对提示长度效应的定量表征可能会开启一种新方法研究上下文依赖模型行为，并可能成为为实践者提供有关不同应用的最佳提示设计策略指导的另一种方式。

Table 3: 不同文本量下的角色效应值估计。星号表示双侧显著性水平：* $p < 0.05$ ，** $p < 0.01$ ，*** $p < 0.001$ 。较高的效应值表明模型输出受角色指令的影响有较大分布变化。医学角色（上方）预期表现稳定，效应值较低，而具有创造性/角色扮演的角色（下方）应表现出更高的变异性 and 更强的效果。

Persona	Token Count								
	100	200	300	400	500	600	700	800	900
Act as a doctor	0.14	0.25	0.20	0.22	0.20	0.16	0.16	0.22	0.19
Act as a nurse	0.18	0.27 **	0.23	0.20	0.25	0.37 ***	0.19	0.17	0.18
Act as a medical practitioner	0.16	0.23	0.22	0.23	0.21	0.45 ***	0.17	0.22	0.18
Act as a medical supervisor	0.24	0.25	0.21	0.26	0.21	0.33 **	0.20	0.17	0.17
Act as a comedian	0.24	0.41 ***	0.34 ***	0.26	0.41 ***	0.18	0.28 *	0.24	0.17
Act as a robot from the future	0.26	0.28 **	0.32 ***	0.42 **	0.29	0.19	0.22	0.19	0.20
Act as a NeurIPS reviewer	0.40 ***	0.36 ***	0.39 **	0.35 *	0.23	0.37 *	0.26	0.25	0.17
Act as a child	0.34 ***	0.31 *	0.32 *	0.27	0.37 ***	0.26 *	0.31 ***	0.18	0.23

C.2. 距离度量的影响

实验设计。为了评估 DBPA 在不同分布距离度量下的鲁棒性，我们使用六种度量方法复制了核心实验：余弦相似度的能源距离、欧几里得距离、Wasserstein 距离、Jensen-Shannon 散度 (JSD)、以及 L1 和 L2 范数的能源距离。

主要结果。结果表明，尽管由于固有的尺度和灵敏度性质不同，导致不同的距离度量产生不同的幅度，但各个角色之间的相对模式依然一致。医学角色在所有度量标准中表现出较低的效应量，而创意和角色扮演角色则表现出显著更高的数值。角色排名在各度量标准上的一致性验证了我们方法的稳健性，而不同的灵敏度则强调了不同的距离度量如何在不同的提示条件下为模型行为提供互补的见解。

Table 4: 跨角色分布距离度量，用于测量相对于基线输出的偏离。数值代表平均距离及其显著性水平：* $p < 0.05$ ，** $p < 0.01$ ，*** $p < 0.001$ 。更高的数值表示与模型默认响应模式的偏离程度更大。

Persona	Distance Measures					
	Energy (Cosine)	Euclidean	Wasserstein	JSD	Energy (L1)	Energy (L2)
Doctor	0.01 ***	31.95	2.72	0.20	1.97 ***	0.06 ***
Nurse	0.01 ***	34.63	2.56	0.20	1.57 ***	0.05 ***
Practitioner	0.01 **	40.88	1.57	0.22	1.27 *	0.04 **
Supervisor	0.01 ***	37.34	1.60	0.21	2.15 ***	0.07 ***
Comedian	0.11 ***	54.99 ***	4.36 ***	0.31	9.11 ***	0.30 ***
Robot	0.01 ***	40.95	2.47	0.22	1.71 **	0.05 ***
Reviewer	0.05 ***	52.63 ***	4.55 ***	0.30 ***	4.84 ***	0.15 ***
Child	0.14 ***	57.03 ***	4.57 ***	0.31 **	11.29 ***	0.36 ***

C.3. 运行扰动的财务成本

设计。为了解决关于 DBPA 嵌入需求计算成本的实际问题，我们进行了不同规模的扰动研究的成本分析。我们评估了三种最先进的 OpenAI 嵌入模型：text-embedding-3-small（每百万个标记 2 美分），text-embedding-3-large（每百万个标记 13 美分），和 text-embedding-ada-002（每百万个标记 10 美分）。我们的分析假设每次扰动平均 200 个标记，包括输入提示和生成的输出，并且研究规模从小型研究（100 次扰动）到大型分析（50,000 次扰动）。

主要结果。成本分析表明，即使在大规模条件下，基于嵌入的扰动研究仍然具有很高的性价比。对于涉及 1,000 到 5,000 次扰动的典型研究应用，成本根据所选模型不同，范围在 4¢ 到 \$ 1.30 之间。即使是包括 50,000 次扰动的大规模研究，嵌入成本也仅需 \$ 2.00 到 \$ 13.00。text-embedding-3-small 模型提供了最经济的选择，同时维持了具有竞争力的性能，使其特别适合大规模扰动分析。这些发现支持了 DBPA 的实用性，因为与生成模型输出本身的计算开销相比，嵌入成本几乎不构成采纳的障碍。

Table 5: 基于嵌入的扰动研究在不同 OpenAI 嵌入模型中的成本分析。成本是根据截至 2025 年 1 月的官方 API 定价计算的。数值显示了处理所有标记所需的总成本，假设每次扰动平均 200 个标记，括号中显示的美元金额表示超过 100 美分（1 美元）的成本。

# of Perturbations	Total Tokens	OpenAI Embedding Models		
		text-embedding-3-small (2¢/1M)	text-embedding-3-large (13¢/1M)	text-embedding-ada-002 (10¢/1M)
100	200,000	0.4¢	2.6¢	2.0¢
200	400,000	0.8¢	5.2¢	4.0¢
500	1,000,000	2.0¢	13.0¢	10.0¢
1,000	2,000,000	4.0¢	26.0¢	20.0¢
2,000	4,000,000	8.0¢	52.0¢	40.0¢
5,000	10,000,000	20.0¢	130.0¢ (\$ 1.30)	100.0¢ (\$ 1.00)
10,000	20,000,000	40.0¢	260.0¢ (\$ 2.60)	200.0¢ (\$ 2.00)
20,000	40,000,000	80.0¢	520.0¢ (\$ 5.20)	400.0¢ (\$ 4.00)
50,000	100,000,000	200.0¢ (\$ 2.00)	1,300.0¢ (\$ 13.00)	1,000.0¢ (\$ 10.00)

C.4. 文本相似性替代方案

实验设计。我们研究了传统的 NLP 指标，如 BLEU 和 ROUGE，是否可以作为替代嵌入相似性的指标，纳入 DBPA 框架中。我们使用 BLEU 分数和三种 ROUGE 变体（ROUGE-1、ROUGE-2 和 ROUGE-L）作为距离度量，复制了我们的角色扮演实验，并计算了与我们主要实验中使用的同样的八个人格之间的效应大小。这些指标直接在文本空间中运作，而不是语义嵌入空间，从而提供了对分布变化非常不同的视角。

主要结果。分析表明，传统的 NLP 指标在技术上可以在 DBPA 框架内运行，但有重要的局限性。BLEU 分数对角色引发的变化表现出特别高的敏感性，像“喜剧演员”和“儿童”这样的创造性角色实现的效果量为 0.65-0.66，明显高于相应的基于嵌入的测量。ROUGE 指标显示出更适中的但仍显著的效果，特别是对于创造性角色。然而，我们警告不要使用这些指标，因为它们主要依赖于基于文本的标记。

C.5. 嵌入的影响

实验设计。我们评估了四种不同的嵌入模型：OpenAI 的 text-embedding-ada-002 (Ada)，以及三种替代模型 (Jasper、Stella 和 Kalm)。每个模型都代表了不同的语义表示方法。我们在所有四种嵌入函数上重复了核心角色实验。

主要结果。分析表明，尽管绝对效应大小在不同的嵌入模型之间变化显著，但相对模式保持相当一致。像“儿童”和“喜剧演员”这样的创造性角色在所有嵌入架构中始终产生最大的效应大小，而医学角色则无论嵌入选择如何，始终表现出较小的效应。这种稳定性表明 DBPA 捕捉到了真实的语义变化，而不是特定嵌入实现的产物。绝对值的变化可能反映了每个模型对情境变化和语义细粒度的敏感性差异。值得注意的是，Ada 嵌入对医学角色显示出更保守的效应大小，但在角色类型之间保持强大的区分能力，这与其作为优化为宽泛语义理解的通用嵌入设计是一致的。

Table 6: 在 DBPA 框架中使用传统 NLP 指标（BLEU 和 ROUGE）作为角色扮演实验的距离衡量标准来估计效应大小。星号表示显著性水平：* $p < 0.05$ 、** $p < 0.01$ 、*** $p < 0.001$ 。虽然这些指标在技术上可以替换我们框架中的嵌入相似度，但它们对小幅扰动非常敏感，可能无法有效捕获零分布。医疗角色（顶部）通常显示较低的效应大小，而创造性角色（底部）表现出显著的效果，特别是对于 BLEU 分数。这些结果应谨慎解释，因为 BLEU/ROUGE 在文本空间而非语义空间中操作。

Persona	Traditional NLP Metrics			
	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Act as a doctor	0.19	0.16	0.18	0.26
Act as a nurse	0.40 ***	0.36 ***	0.31 *	0.28
Act as a medical practitioner	0.20	0.27 *	0.20	0.19
Act as a medical supervisor	0.24	0.34	0.34 **	0.26
Act as a comedian	0.65 ***	0.37 ***	0.38 ***	0.26
Act as a robot from the future	0.30 **	0.26 *	0.25	0.20
Act as a NeurIPS reviewer	0.26	0.22	0.21	0.24
Act as a child	0.66 ***	0.44 ***	0.49 ***	0.33 ***

Table 7: 不同嵌入模型中，各种人格对效应值的估计。星号表示双侧显著性水平：* $p < 0.05$ ，** $p < 0.01$ ，*** $p < 0.001$ 。较高的效应值表明在给定人格指令情况下，模型输出的更大分布变化。预计医疗类人格（顶部）将显示稳定且较低的效应值，而创意/角色扮演类人格（底部）应展示更高的可变性和在不同嵌入模型中更强的效应。我们发现，创意人格在所有嵌入架构中一致地产生较大的效应值，其中儿童人格表现出最强的效应。

Persona	Embedding Model			
	Ada	Jasper	Stella	Kalm
Act as a doctor	0.21	0.07 ***	0.07 ***	0.03 ***
Act as a nurse	0.20	0.03 ***	0.03 ***	0.02 ***
Act as a medical practitioner	0.19	0.05 ***	0.04 ***	0.02 **
Act as a medical supervisor	0.17	0.09 ***	0.08 ***	0.05 ***
Act as a comedian	0.24	0.36 ***	0.31 ***	0.23 ***
Act as a robot from the future	0.26 *	0.04 ***	0.04 ***	0.05 ***
Act as a NeurIPS reviewer	0.28 *	0.14 ***	0.13 ***	0.13 ***
Act as a child	0.33 **	0.48 ***	0.41 ***	0.35 ***

D. 扩展的相关工作

表格 8 根据在分布基础上的扰动分析背景下重要的五个标准评估了主要的相关工作。

Method	Example Works	(I)	(II)	(III)	(IV)	(V)	Representative Question
Measuring Unintended Bias	(Borkan et al., 2019; Dixon et al., 2018; Park et al., 2018)	✓	✗	✓	✓	✗	Does this model have unintended biases in certain subgroups?
Counterfactual Fairness	(Garg et al., 2019)	✗	✗	✗	✓	✗	How would the prediction change if the sensitive attribute were different?
Text Summarization	(Bhandari et al., 2020; Zhang et al., 2019; Zhao et al., 2019; Lin, 2004)	✗	✗	✗	✓	✓	How well is this text summarized?
Distribution-based perturbation analysis	This work	✓	✓	✓	✓	✓	Do the responses change if we change any input in the prompt? If so, how?

Table 8: 相关的量化文本输出的指标。缩写: (I): 可以用于任何黑盒模型; (II): 可以应用任何扰动; (III): 支持统计推断; (IV): 计算效应大小; (V): 无假设。基于分布的方法是唯一可以通过评估所有可能输出的分布的期望来量化任何扰动对结果的影响的方法。这很重要, 因为单点评估无法捕捉大型语言模型的随机特性。