arXiv:2601.06586v1 [cs.CL] 10 Jan 2026

具备性能保证的大语言模型生成文本检测

周弘毅[1*]、朱进[*2†]、杨颖[1], 施承春[3†]

[1]清华大学, [2]伯明翰大学 &[3]伦敦政治经济学院

摘要

GPT、Claude、Gemini 和 Grok 等大语言模型已深度融入我们的日常生活。如今，它们支持各种各样的任务——从对话和邮件起草，到协助教学和编程、充当搜索引擎等。然而，它们生成高度仿真人类文本的能力引发了诸多严重担忧，包括假新闻传播、误导性政府报告的生成以及学术不端。为解决这一实际问题，我们训练了一个分类器，用于判断一段文本是由大语言模型还是人类撰写的。我们的检测器已部署在一个在线 CPU 平台上 https://huggingface.co/spaces/stats-powered-ai/StatDetectLLM，与现有检测器相比，具有三项创新之处：（i）无需依赖水印或对生成文本所用特定大语言模型的了解等辅助信息；（ii）能更有效地区分人类撰写文本与大语言模型撰写文本；（iii）支持统计推断，而这一点在当前文献中基本缺失。实证结果表明，我们的分类器在保持第一类错误控制、较高统计功效和计算效率的同时，取得了高于现有检测器的分类准确率。

关键词：大语言模型、机器生成文本检测、分类、统计推断。

*前两位作者对本文贡献相同，并按字母顺序排列。
†共同通讯作者。

1 引言

过去几年，通用大语言模型（LLM）迅速发展，例如 GPT（Hurst 等，2024）、DeepSeek（Liu 等，2024）、Claude（Anthropic，2024）、Gemini（Comanici 等，2025）、Grok（xAI，2025）和通义千问（Yang 等，2025）。这些模型在广泛的任务中展现出卓越性能，涵盖传统的问答、摘要、翻译，以及推理和代码生成。如今，它们已深度融入金融、教育、医疗保健、软件工程和新闻业等多个应用领域（Arora& Arora2023、Chan& Hu2023、Hou etal.2024、Liuet al. 2025）。

由于具备生成高度连贯、近似人类文本的能力，这些大语言模型也带来了严峻的社会和伦理挑战，涉及作者归属判定、学术诚信、知识产权以及错误信息的传播。例如，在学术领域，作者可能会将大语言模型生成的观点或文字当作自己的成果发表，却不予适当署名。同样，审稿人可能依赖大语言模型生成的评审意见，而不仔细阅读论文，导致评审质量低下且缺乏建设性。近期一项研究报告称，提交给2024年国际学习表征会议（ICLR）——最负盛名的机器学习会议之一——的评审意见中，至少有15.8%借助了人工智能（Latona等，2024年）。与此同时，2024年发表的PubMed收录论文中，至少有13.5%的摘要经过大语言模型处理（Kobak等，2025年）。在社交媒体平台上，大语言模型可以轻易地大规模助长虚假信息的传播（Weidinger等，2021）。最后，政策制定者和政府日益强调，有必要对生成式人工智能技术加以保护，以确保其负责任且值得信赖（经济合作与发展组织，2024年）。因此，开发能够可靠地区分人类撰写文本与大语言模型撰写文本的算法，已成为当务之急（Crothers等，2023年、Wu等，2025年）。

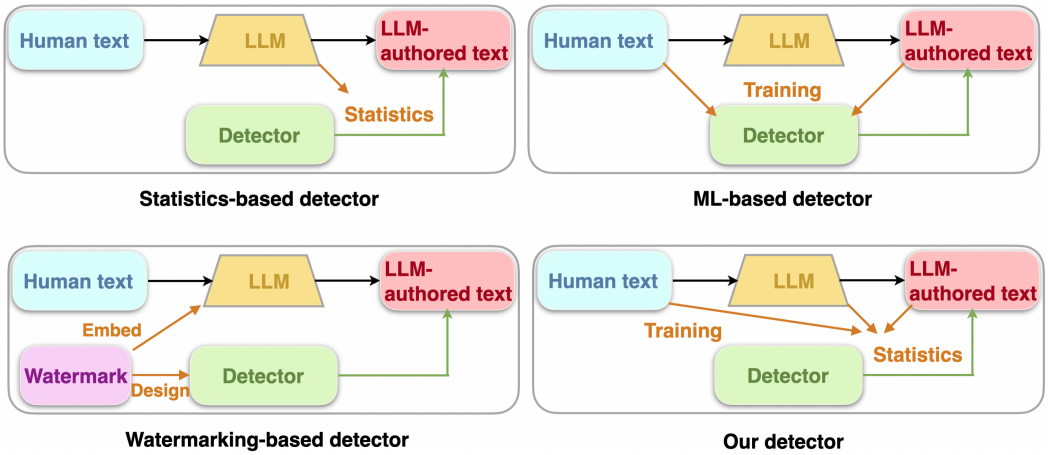

图1：三类现有检测器（基于统计量、基于机器学习、基于水印）的可视化，以及所提出的检测器。

1.1 相关工作

尽管大语言模型近期才出现，但关于检测大语言模型生成文本的研究已经相当丰富。总体而言，现有方法分为三类：基于水印的方法、基于机器学习（ML）的方法以及基于统计的方法。我们将在下文分别回顾这几类方法，并在图1中对这些方法进行可视化。

1. **基于水印的方法**会在目标大语言模型的输出文本中嵌入称为水印的细微信号（见图1左下图）。随后，可以通过检验这些信号是否存在，来识别由该大语言模型生成的文本。这项技术的早期实例之一是 Gumbel 水印（Aaronson & Kirchner2023）。最近，统计学文献中的一系列研究将水印检测视为统计假设检验问题（例如，见Li et al. 2025*a*、*b*、Xie et al. 2025）。在这一框架下，零假设认为输入文本中不存在水印；拒绝零假设则能提供统计证据，表明文本由目标大语言模型生成。然而，这些方法依赖于了解目标大语言模型在生成词元时所使用的特定哈希函数或随机数生成器（RNG），而这些信息通常并未公开。

2. **基于统计的方法**从目标大语言模型中提取信息，构造检验统计量，再据此区分人工撰写的文本和机器生成文本（见图1左上图）。举个简单的例子，假设某些词语在大语言模型生成文本中出现的可能性高于在人工撰写的文本中出现的可能性。那么，这些词语在输入文本中的频率就可以作为检验统计量：如果频率足够高，我们便可得出结论，认为该文本由大语言模型生成。文献中提出了多种统计量，其中最常用的是基于目标大语言模型下一个词元预测分布的 Logits（Mitchell et al. 2023、Su et al.2023、Bao et al. 2024、Hans etal. 2024、Zhouet al. 2025）；有关这些基于 logits 的方法的更多详情，请参见第2 节。其他统计量包括输入文本的 N元语法分布（Solaiman et al. 2019、Yang etal. 2024）、其内在维度（Tulchinskii etal.2023）、输入文本中词元概率绝对排名的分布（Gehrmann et al. 2019）、目标大语言模型所使用的奖励模型（Leeet al. 2024），以及最大均值差异（Zhanget al.2024、Song et al.2025）。然而，这些方法只关注分类，并未研究统计推断。

3. **基于机器学习的方法**利用互联网上现有的大型人类撰写语料库，提示目标大语言模型生成相应的大语言模型撰写文本，然后使用这两类文本训练分类模型进行检测（见图 1 右上方的面板）。根据所用分类模型的不同，这些方法可进一步分为三类：(i) 经典机器学习模型（例如决策树、支持向量机）；(ii) 大语言模型；以及 (iii) (i) 和 (ii) 的混合模型。具体而言，第一类方法从输入文本中提取某些特征，并将其输入分类模型。例如，基于统计的方法中所述的统计量可用作此类特征。其他特征包括经典的词频-逆文档

频率、一元组和二元组（Solaiman et al. 2019）、输入文本与目标大语言模型之间计算得出的交叉熵损失（Guo et al. 2024），以及原始输入文本与其大语言模型改写版本之间的语义差异度量（Maoet al.2024）。第二类方法直接使用大语言模型作为分类模型。用于此目的的语言模型包括 RoBERTa（Solaiman etal. 2019，Guo et al. 2023）、BERT（Ippolito et al. 2020）和 DistilBERT（Mitrović et al.2023）。这种方法有充分的依据，因为大语言模型本就旨在处理文本，而且可以在数据集上微调由此得到的模型参数。最后一类方法将微调后的大语言模型的输出用作经典机器学习分类的输入特征（Abburi etal. 2023）。与基于统计的方法类似，这些方法不考虑统计推断。

总而言之，现有方法都存在一定的局限性。基于水印的方法需要了解目标大语言模型所使用的哈希函数或随机数生成器，但它们能够进行统计推断，从而得出用于不确定性量化的 p-值。相比之下，大多数基于统计和机器学习的方法不需要访问哈希函数或随机数生成器，但通常无法进行统计推断，Zhu 等人(2025) 是一个例外。最后，所有这些方法都是特定模型的：它们旨在检测由特定目标大语言模型生成的文本，无法泛化到其他模型。

1.2 贡献

本文提出了一种检测器，旨在克服现有方法的局限性。我们的主要贡献如下：

1. 我们收集了一个包含一万多篇文本的数据集，涵盖从医学到法律文件等多个领域，文本由人类以及近期流行的大语言模型（如 GPT 系列、Grok 和 Gemini）生成。该数据集可用于训练分类器

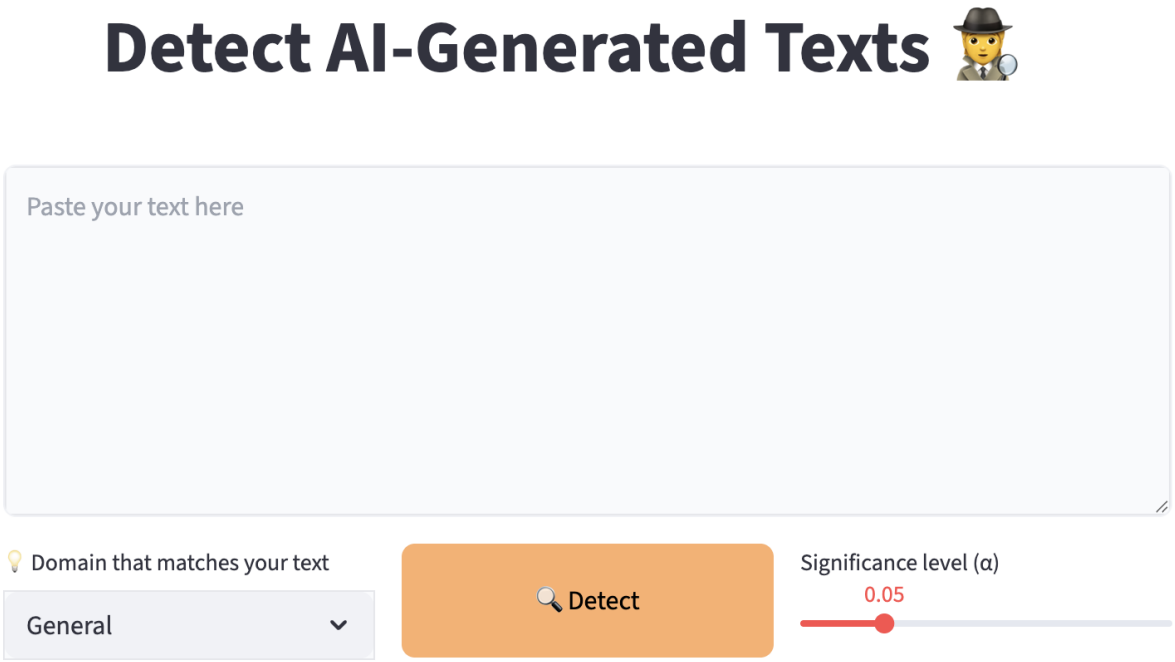

图 2：我们检测器的网站界面：用户可以在标有“在此处粘贴文本”的阴影区域输入文本，选择文本所属的领域（例如金融、法律；默认为通用），并选择显著性水平 α，然后点击“检测”。随后，检测器会生成输出（见图 8）。如未另行指定，显著性水平设为 0.05。

以检测不同领域中的大语言模型生成文本。

2. 我们使用收集的数据集训练了一个检测器，该检测器与目标大语言模型无关，且无需访问模型的哈希函数或随机数生成器。与现有的基于统计量和基于机器学习的检测器相比，它的分类性能更优，同时能够控制第一类错误，并保持与基于水印的方法相近的高检验功效。

3. 我们部署了一个公开可用的网站（见图 2），用于托管我们的检测器，帮助用户检测大语言模型生成文本，而无需下载模型或使用我们的数据重新训练检测器。

1.3 论文结构

本文其余部分安排如下。第 2 节回顾问题设定，并介绍与我们的方法密切相关的基于逻辑值的检测器。第 3节介绍我们为训练检测器而收集的数据集。第 4 节描述我们的训练方法。第 5节报告了我们的方法在分布内和分布外数据上的评估数值结果。第 6 节介绍我们的网站。最后，第 7节总结全文，并讨论未来研究方向。

2 预备知识

2.1 问题设置

本节将介绍相关符号并对问题进行形式化定义。在自然语言处理中，词元是文本的基本单位，例如单词、子词、字符或标点符号。我们将所有此类词元的集合称为词汇表，记作 $\mathcal{V}$。一段文本可以表示为词元序列，记作 $\boldsymbol{X} = (X_1, \ldots, X_T)$，其中 $X_t \in \mathcal{V}$ 表示序列中的第 t 个词元，T 表示该文本段中的词元总数。不失一般性，我们假设所有文本段的长度均为T。具体而言，可以将 T 设置为最大长度，并用零填充较短的文本段，以达到所需长度。

令 $\mathbb{P}$ 和 $\mathbb{Q}$ 分别表示人类撰写和大语言模型生成词元序列的分布函数。由于每段文本 $\boldsymbol{X}$ 都是一个时间序列，因此 $\mathbb{P}$ 和 $\mathbb{Q}$ 均可分解为条件概率分布序列 $\{p_t\}_{t=1}^{T}$ 和 $\{q_t\}_{t=1}^{T}$，其中每个 p_t （以及 q_t）表示在给定所有前置词元的条件下，X_t 的条件分布，前置词元记为 $\boldsymbol{X}_{<t} = (X_1, \cdots \quad X_{t-1})$，即

$$p_t(x_t|\boldsymbol{x}_{<t}) = \mathbb{P}(X_t = x_t|\boldsymbol{X}_{<t} = \boldsymbol{x}_{<t}),\ q_t(x_t|\boldsymbol{x}_{<t}) = \mathbb{Q}(X_t = x_t|\boldsymbol{X}_{<t} = \boldsymbol{x}_{<t}),$$

对于任意 $x_t \in \mathcal{V}$， $\boldsymbol{x}_{<t} \in \mathbb{V}^{t-1}$。

注意， $\mathbb{P}$ 和 $\mathbb{Q}$ 都未知。此外，与大多数现有方法（例如，参见 Mitchell 等（2023）、Bao 等人，2024）不同，我们不将 $\mathbb{Q}$ 限制为某个特定的目标大语言模型。相反， $\mathbb{Q}$ 对应于一个混合分布，代表多个大语言模型分布的混合。定义如下这对假设：

$$\mathcal{H}_0 : \boldsymbol{X} \sim \mathbb{P} \text{ versus } \mathcal{H}_1 : \boldsymbol{X} \sim \mathbb{Q}. \tag{1}$$

我们的目标是引入一条分类规则 $S > c$，使得当 $S > c$ 时，我们将文本分类为大语言模型生成。这里， S 表示用于区分 p 和 q的统计量， c 表示分类阈值。我们的目标有两个方面：

1. 设计一个检验功效更高的统计量 S，使得在改变分类阈值 c时，所得分类器的曲线下面积（AUC）大于文献中提出的方法。

2. 计算一个阈值 c ，使所得分类规则的第一类错误（即，当 $\boldsymbol{X} \sim \mathbb{P}$ 时，错误地得出 $\boldsymbol{X} \sim \mathbb{Q}$结论的概率）控制在预先指定的显著性水平 $0 < \alpha < 1$。

2.2 基于 Logits 的检测器

如第 1.1 节所述，基于 Logits 的检测器是一类广受欢迎的基于统计的检测器，其统计量根据目标大语言模型的 Logits 构造。Logit 是大语言模型 $\mathcal{M}$ 在应用 Softmax 函数将分数转换为概率之前，为每个可能的下一个词元分配的原始、未归一化分数。用数学形式表示，令 $\ell_t^{\mathcal{M}}(x_t|\boldsymbol{x}_{<t})$表示 $\mathcal{M}$ 在给定前文 $\boldsymbol{x}_{<t}$时，为词元 x_t 分配的 Logit。那么，模型对词元 x_t 的预测概率为

$$q_t^{\mathcal{M}}(x_t|\boldsymbol{x}_{<t}) = \frac{\exp\{\tau^{-1}\ell_t^{\mathcal{M}}(x_t|\boldsymbol{x}_{<t})\}}{\sum_{x\in\mathcal{V}}\exp\{\tau^{-1}\ell_t^{\mathcal{M}}(x|\boldsymbol{x}_{<t})\}}, \tag{2}$$

其中， $\tau > 0$ 表示温度参数。

一种值得关注的基于 logits 的方法是 Fast-DetectGPT（Bao 等人，2024），它是早期 DetectGPT（Mitchell 等（2023））的一种计算效率更高的变体。Fast-DetectGPT 基于以下统计量来检测由 $\mathcal{M}$ 生成的文本：

$$\frac{\sum_t \log q_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \sum_t \mathbb{E}_{\widetilde{X}_t \sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}[\log q_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t \sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}(\log q_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t}))}}, \tag{3}$$

其中， $q_t^{\mathcal{S}}$ 表示采样模型$\mathcal{S}$的下一个词元预测分布，用于对给定的 $\boldsymbol{X}_{<t}$进行采样 $\widetilde{X}_t$ ，且它可能与评分模型$\mathcal{M}$相同，也可能不同。

(3) 分子中的第一项是目标模型 $\boldsymbol{X}$ 在目标模型$\mathcal{M}$下的对数似然。分子中的第二项和分母分别充当中心化项和归一化项，确保当 $\boldsymbol{X}$ 由采样模型$\mathcal{S}$生成时，该统计量近似服从零均值、单位方差分布。

根据（2），显然，中心化项的存在会抵消归一化常数 $\sum_{x \in \mathcal{V}} \exp\{\tau^{-1} \ell_t^{\mathcal{M}}(x|\boldsymbol{X}_{<t})\}$ ，该常数出现在（2）的分母中。因此，（3）的分子变为

$$\frac{1}{\tau} \sum_t [\ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \mathbb{E}_{\widetilde{X}_t \sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})} \ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})].$$

此外，分母中的归一化项会抵消温度参数τ，因此（3）恰好变为 Logits 的标准化版本：

$$S_{\mathrm{Fast}}(\boldsymbol{X}) = \frac{\sum_t \ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \sum_t \mathbb{E}_{\widetilde{X}_t \sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}[\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t \sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}(\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t}))}}. \tag{4}$$

使用（4）作为统计量的理由是，与人工撰写的文本相比，大语言模型生成文本平均而言往往会产生更大的统计量值。为证明这一观察结果，我们给出以下定理：

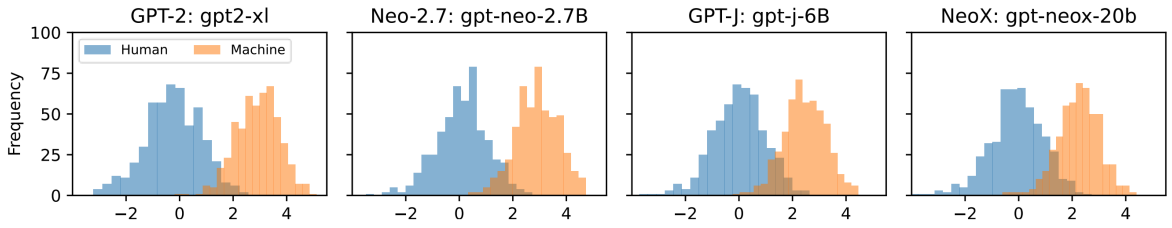

图 3：Fast-DetectGPT 统计量在人工撰写文本和由各种大语言模型生成的文本上的直方图，摘自 Bao 等人 （2024）。

定理 1. 假设采样模型 S 与评分模型相同（即目标模型 M）。那么，当温度参数为 $\tau \to 0^+$ 时，我们有

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} S_{Fast}(\boldsymbol{X}) \leq \mathbb{E}_{\boldsymbol{X}\sim\mathbb{Q}^{\mathcal{M}}} S_{Fast}(\boldsymbol{X}),$$

其中， $\mathbb{Q}^{\mathcal{M}}$ 表示目标模型 M 生成文本的概率分布。

定理 1 的证明很简单。当 $\mathcal{S} = \mathcal{M}$ 时，只需证明

$$\sum_t \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \sum_t \mathbb{E}_{\substack{\boldsymbol{X}_{<t}\sim\mathbb{P} \\ \widetilde{X}_t\sim q_t^{\mathcal{M}}(\bullet|\boldsymbol{X}_{<t})}} [\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})] \leq 0 \tag{5}$$

当温度参数趋近于0时，从 $\widetilde{X}_t$ 中 $q_t^{\mathcal{M}}$采样会使所有概率质量集中于使 Logit $\ell_t^{\mathcal{M}}(\bullet|\boldsymbol{X}_{<t})$ 达到最大值的词元上。因此，无论分布 $\mathbb{P}$如何，式（5）的左侧始终非正。

因此，定理 1 证明了温度足够低时该结果成立。然而在实践中，现代大语言模型使用的温度通常不为零。尽管如此，实验证据始终表明，与人工撰写的文本相比，该统计量在大语言模型生成文本上往往取值更大；图 3 展示了相关示例。因此，当该统计量超过某个阈值时，可以将输入文本判定为大语言模型生成文本。

然而，当温度较高时，仅依赖 Logits 不足以稳定地区分不同的分布；参见Zhou 等人 (2025) 讨论的“比特王国”示例。为了解决这一局限，Zhou 等人(2025) 开发了 AdaDetectGPT，

基于以下统计量

$$S_{\mathrm{Ada}}(\boldsymbol{X}) = \frac{\sum_t w(\log q_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t})) - \sum_t \mathbb{E}_{\widetilde{X}_t\sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}[w(\log q_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t}))]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t\sim q_t^{\mathcal{S}}(\bullet|\boldsymbol{X}_{<t})}(w(\log q_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})))}}. \tag{6}$$

比较 (6) 和 (3) 可见，这两个统计量的结构非常相似。唯一的区别是，(6) 会先对对数概率应用一维见证函数 w，然后执行相同的中心化和归一化步骤。Zhou 等人(2025) 提议使用 B 样条基函数对该见证函数进行参数化 w，并通过最大化所得检测器的曲线下面积，根据数据对其进行自适应学习。

他们的关键观察在于，(6) 的分子相对于滤子 $\{\sigma(\boldsymbol{X}_{<t})\}_{t\geq 1}$ 构成了一个鞅差序列，其中 $\sigma(\boldsymbol{X}_{<t})$ 表示由 σ-代数生成的 $\boldsymbol{X}_{<t}$。这使得我们能够应用经典的鞅中心极限定理（MCLT，Hall 与 Heyde（2014）），推导出所得分类器的假阴性率（FNR）和真阴性率（TNR）的闭式表达式，进而便于推导用于优化 w的目标函数。从理论上说，他们为该分类器建立了统计保证，包括其假阴性率和假阳性率（FPR）的下界，以及其真阴性率和真正率（TPR）的上界。

然而，使用一维见证函数 w 不足以显著提升分类准确率。此外，Fast-DetectGPT 和 AdaDetectGPT 均无法控制第一类错误。这两种方法也都依赖模型：其检验统计量针对特定目标模型，因此需要指定目标模型。然而在实践中，用户通常并不掌握此类先验信息——他们希望判断一段文本是否由大语言模型生成，而无需预先知道可能生成该文本的是哪个特定大语言模型。我们将在接下来的两节中解决这些局限。

3 数据

为了解决 DetectGPT 和 AdaDetectGPT 的上述局限，我们构建了一个大型数据集，其中包含人工撰写的文本和大语言模型生成文本，用于训练我们的检测器。本节将详细介绍这两类数据。

3.1 人工撰写的文本

数据来源。 我们的初步研究表明，基于机器学习的检测器的性能在很大程度上取决于训练数据。例如，如果训练数据集包含大量来自某一领域的文本，而其他领域的文本较少，那么由此得到的检测器往往在该领域的分类准确率较高，但在其他代表性不足的领域表现较差。这一发现促使我们收集尽可能多样化的训练数据，涵盖广泛的领域。

基于这一原则，我们收集了涵盖八个领域的大型人工撰写文本语料库，并按字母顺序列出如下：(i) 学术，(ii) 金融，(iii) 政府，(iv) 知识，(v) 立法，(vi) 医学，(vii) 新闻，以及 (viii) 用户评论。表 1 概述了各领域所用的数据集。

这些文本收集自 Kaggle、GitHub 和 Hugging Face 等公开平台。在每个领域中，我们至少从两个不同来源收集数据，以确保多样性，最终共得到 18 个数据集。为避免大语言模型生成内容造成污染，我们仅保留 2022年11月之前编写的数据集（ChatGPT 的公开发布日期），并舍弃此后发布的所有数据。

数据处理。 收集完成后，我们对所有文本进行处理，确保其质量达到较高水平，再将其输入算法以训练检测器。构建训练数据集时，我们从每个领域随机抽取数量相等的文本，以避免某些领域占比过高

学术 (50000)	来自 arXiv 的论文摘要（arXiv.org 投稿者 2024） 发表于 PubMed 的研究论文（Cohan 等人，2018年）
金融 (49254)	Reddit 上公开的金融讨论 金融观点挖掘与问答数据集（Maia 等人，2018年） 2006年至2013年的彭博金融文章（Flowers，2025年）
政府 (38887)	美国政府报告和专家撰写的摘要（Huang 等人，2021年）
知识 (49251)	2022年4月之前的维基百科文章（Foundation，日期不详。）
立法 (49987)	美国和欧盟立法文件（Chalkidis 等人，2021，2022年）
医学 (44917)	来自开源数据集的公开患者记录（Khandekar 等，2024年） 医学入学考试题目的专家答案（Pal 等，2022年）
News (49954)	CNN 和 BBC 的文章（Greene 和 Cunningham，2006年，Narayan 等，2018年）
用户评论 （47902）	来自 Yelp、IMDb 和亚马逊的评论（Maas 等人，2011 年、Zhang 等人，2015 年）

表1：按领域汇总的人工撰写文本数据集。括号中的数字表示每个类别收集的文本总数。

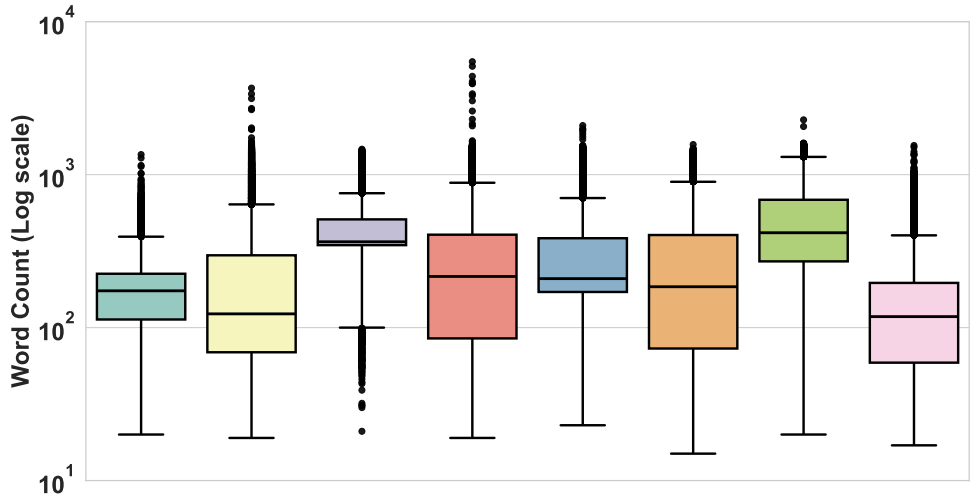

图4：收集的数据集中各领域人工撰写文本词数的箱线图。

任何领域。接下来，我们对每篇文本应用以下处理步骤：移除控制字符、去除多余的空格和换行符、删除特殊标记，以及消除重复项（有关每个步骤的详细信息，请参见补充材料中的第 B.1 节）。

经过这一处理后，每个领域都包含超过38,000篇文本，合计得到超过370,000篇人工撰写文本。我们在表1中汇总了各领域的文本总数。图4展示了各领域人工撰写文本词数的箱线图。可以看出，文本平均长度从几十词到几百词不等。这种差异反映了不同领域在写作风格和格式方面的多样性。

3.2 大语言模型生成文本

接下来，我们提示大语言模型根据上述人工撰写的文本生成大语言模型撰写文本。生成这些文本有两个目的：（i）用于训练我们的检测器；（ii）用于评估我们的

检测器相对于现有方法的表现。为确保所得检测器不依赖于特定模型，我们使用多种广泛应用的大语言模型生成文本，包括 GPT-4o（Hurst 等，2024）、Claude（Anthropic，2024）、Gemini（Comanici 等，2025）和 Grok（xAI，2025）。此外，我们采用多种提示词，使大语言模型能够根据相同的人工撰写的输入生成多样化的输出，包括改写、润色和扩写，以及摘要。有关数据生成流程的更多细节，请参阅补充材料第 B.2 节。

4 方法

我们首先介绍用于训练统计量的过程 S，该统计量由我们的检测器使用（第4.1节）。接下来，我们详细说明如何确定阈值 c，以确保假设检验有效（第4.2节）。

4.1 训练

我们的检测器基于 Fast-DetectGPT 构建。如第2.2节所述，当温度较高时，仅依赖对数概率来构建统计量可能并非最优。AdaDetectGPT 通过对数概率应用见证函数，在一定程度上缓解了这一问题；然而，一维见证函数的表达能力仍然有限。为了解决这些局限，我们考虑以下统计量：

$$S(\boldsymbol{X}) := \frac{\sum_{t=1}^{T}[w_t(X_t|\boldsymbol{X}_{<t}) - \mathbb{E}_{\widetilde{X}_t\sim q_t(\bullet|\boldsymbol{X}_{<t})}\{w_t(\widetilde{X}_t|\boldsymbol{X}_{<t})\}]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t\sim q_t(\bullet|\boldsymbol{X}_{<t})}(w_t(\widetilde{X}_t|\boldsymbol{X}_{<t}))}}, \tag{7}$$

对于一组函数$\{w_t\}_t$，这些函数通过第3节所述的构建的数据集进行自适应学习。

我们的统计量与 AdaDetectGPT 的统计量之间的关键区别，在于所学习函数的灵活性。在（7）中，每个 w_t 都可以是关于以下两者的任意函数

前置词元 $\boldsymbol{X}_{<t}$ 和当前词元 X_t，并且可以随 t变化。相比之下，AdaDetectGPT 使用时不变的见证函数 w，其输入仅为一维对数概率。实际上，令 $w_t = w \circ \log q_t^{\mathcal{M}}$，我们的统计量便可化简为 $S_{\text{Ada}}(\boldsymbol{X})$。更一般地，使用一系列函数$\{w_t\}_{t=1}^T$ 会显著提升所得统计量的表达能力。我们将在第5节通过实证展示这一优势。

接下来，我们定义一个用于学习 $\{w_t\}_t$ 的目标函数，以最大化所得统计量的分类准确率。由于这本质上是一个二分类问题，一个自然的目标是曲线下面积。注意，最大化曲线下面积等价于在任意固定的假阴性率下最大化检测器的真阴性率。我们遵循 Zhou 等人 (2025) 推导我们的学习目标。

具体而言，与 (6) 类似，当 $\boldsymbol{X} \sim \mathbb{Q}$ 时，对于任意序列 $\{w_t\}_t$，(7) 的分子构成相对于 $\{\sigma(\boldsymbol{X}_{<t})\}_{t\geq 1}$ 的鞅差序列。将分类阈值 c 设为标准正态分布的上侧 α 分位数（记为 z_α），根据鞅中心极限定理，所得检测器的假阴性率 $S(\boldsymbol{X})$ 渐近地受控于水平 α。再次应用鞅中心极限定理，可以推导出其对应真阴性率的下界，我们将在下面的定理中对其进行形式化表述。

定理 2。 在补充材料第 A 节所规定的条件下，我们的检测器在假阴性率为 α 时，其真阴性率的渐近下界为 $\min\{\alpha + \phi(z_\alpha)L_w,\ 1-\alpha\}$，其中 ϕ 表示标准正态分布的概率密度函数，并且

$$L_w = \frac{\sum_t [\mathbb{E}_{\substack{\widetilde{X}_t \sim q_t(\bullet|\boldsymbol{X}_{<t}) \\ \boldsymbol{X}_{<t} \sim \mathbb{P}}} \{w_t(\widetilde{X}_t|\boldsymbol{X}_{<t})\} - \mathbb{E}_{\substack{\widetilde{X}_t \sim p_t(\bullet|\boldsymbol{X}_{<t}) \\ \boldsymbol{X}_{<t} \sim \mathbb{P}}} \{w_t(X_t|\boldsymbol{X}_{<t})\}]}{\sqrt{\sum_t Var_{\substack{\widetilde{X}_t \sim q_t(\bullet|\boldsymbol{X}_{<t}) \\ \boldsymbol{X}_{<t} \sim \mathbb{P}}} \{w_t(\widetilde{X}_t|\boldsymbol{X}_{<t})\}}}.$$

我们有几点说明。首先，L_w 是一个标量，其取值取决于 $\{w_t\}_t$。其次，最大化定理 2 中的下界等价于最大化 L_w。第三，因为

L_w 不依赖于假阴性率水平 α，因此优化 L_w 会同时最大化所有 α 对应的真阴性率下界。因此，L_w 的最大化解也会最大化所得检测器的曲线下面积下界。最后，虽然可以推导出真阴性率本身（而非其下界）的闭式表达式，但这样的表达式通常取决于 α。直接对其进行最大化会得到一个针对特定假阴性率水平的优化器，却无法保证在其他水平下也能达到最优性能。相比之下，最大化基于下界的目标函数 L_w 会得到一个“α-统一”的优化器。

基于上述讨论，自然可以将 L_w 设为优化的目标函数。我们进行了一些修改，以进一步简化优化。首先，我们将 L_w 分子第一项中的期望 $\boldsymbol{X}_{<t} \sim \mathbb{P}$ 替换为 $\boldsymbol{X}_{<t} \sim \mathbb{Q}$。这样，分子便可简化为

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\Big[\sum_t w_t(X_t|\boldsymbol{X}_{<t})\Big] - \mathbb{E}_{\boldsymbol{X}\sim\mathbb{Q}}\Big[\sum_t w_t(X_t|\boldsymbol{X}_{<t})\Big]. \tag{8}$$

请注意，最大化（8）与机器学习领域广泛研究的最大均值差异度量密切相关（Gretton 等人 2012）。此外，（8）中的两个期望都可以通过对我们构建的数据集中由人类撰写和大语言模型生成的文本计算经验平均值来近似。

其次，按照对分子的这一修改，我们相应地将 L_w的分母调整为经验平均值的标准差。具体而言，我们将分母设为 $\sqrt{\mathrm{Var}_{\boldsymbol{X}\sim\mathbb{P}}(\sum_t w_t(X_t|\boldsymbol{X}_{<t}) + \mathrm{Var}_{\boldsymbol{X}\sim\mathbb{Q}}(\sum_t w_t(X_t|\boldsymbol{X}_{<t}))}$，由此得到以下双样本 t-检验型目标函数，

$$\frac{\widehat{\mathbb{E}}_{\boldsymbol{X}\sim\mathbb{Q}}\{\sum_t w_t(X_t|\boldsymbol{X}_{<t})\} - \widehat{\mathbb{E}}_{\boldsymbol{X}\sim\mathbb{P}}\{\sum_t w_t(X_t|\boldsymbol{X}_{<t})\}}{\sqrt{\widehat{\mathrm{Var}}_{\boldsymbol{X}\sim\mathbb{P}}(\sum_t w_t(X_t|\boldsymbol{X}_{<t}) + \widehat{\mathrm{Var}}_{\boldsymbol{X}\sim\mathbb{Q}}(\sum_t w_t(X_t|\boldsymbol{X}_{<t}))}}, \tag{9}$$

其中，$\widehat{\mathbb{E}}$ 和 $\widehat{\mathrm{Var}}$ 分别表示基于构建的数据集计算得出的经验均值和抽样方差估计量。

为了优化（9），我们需要为 $\{w_t\}_t$ 指定一个函数类。在这里，我们将 $\{w_t\}_t$ 参数化

使用基础语言模型 $\{q_t^{\mathcal{B}}\}_t$，因为 $q_t^{\mathcal{B}}$和 w_t 都以 X_t 和 $\boldsymbol{X}_{<t}$ 为输入，并输出一个标量值。然后，我们在构建的数据集上微调 $\{q_t^{\mathcal{B}}\}_t$ 的参数，以最大化目标函数（9）。最后，我们将微调后的模型代入（$\{w_t\}_t$ （7）中，从而得到我们的统计量。

4.2 检验

在本节中，我们讨论如何选择阈值 c，以控制检验假设中的第一类错误（例如，FPR）（1）。与在 $\boldsymbol{X} \sim \mathbb{Q}$ 下评估的假阴性率不同，其（7）的分子构成鞅差序列，从而能够使用鞅中心极限定理刻画其渐近分布；而假阳性率则在$\boldsymbol{X} \sim \mathbb{P}$下评估。在这种情况下，检验统计量的渐近分布分析要困难得多。

为应对这一挑战，我们利用收集的人工撰写文本对零分布进行经验估计，记为 $\{\boldsymbol{X}^{(j)}\}_{j=1}^{m}$。对于每个 $\boldsymbol{X}^{(j)}$，我们计算统计量$S(\boldsymbol{X}^{(j)})$。给定一篇待检验的新文本 $\boldsymbol{X}$，我们计算其 p-值如下：

$$p\text{-value} = \frac{1 + \sum_{j=1}^{m} \mathbb{I}\Big(S(\boldsymbol{X}) < S(\boldsymbol{X}^{(j)})\Big)}{1+m},$$

并将阈值 c 设为 $S(\boldsymbol{X})$的最大值，使所得的 p-值不大于预先指定的显著性水平 $0 < \alpha < 1$。如果 p-值不大于 α，或等价地，如果 $S(\boldsymbol{X}) > c$，我们就拒绝零假设，并得出 $\boldsymbol{X}$ 是大语言模型生成的结论。理论上，以下定理确立了该程序的有效性。**定理 3**。所提出的检验渐近地将第一类错误控制在 $m \to \infty$。

从经验上看，我们发现，$S(\boldsymbol{X})$ 的分布在不同类别的人工撰写文本之间存在显著差异。为考虑这种异质性，在检验新文本$\boldsymbol{X}$时，我们会请用户指定该文本属于第3 节所述的八个类别中的哪一类

，然后使用该类别的人工撰写文本计算相应的 p值和分类阈值。如果没有领域信息，我们会采用保守策略，报告八个类别中最大的 p值。图10 展示了不同类别人工撰写文本的 $S(\boldsymbol{X})$ 零分布。

5 真实数据分析

本节评估所提出的检测器在有限样本下的性能。我们使用两类数据：一类是从第 3 节所述的八个类别中抽取的分布内数据，并通过样本拆分进行评估（第 5.1 节）；另一类是来自外部数据集的分布外数据（第 5.2 节）。最后，我们在第 5.3 节报告计算成本。

5.1 分布内评估

我们首先评估检测器在第 3 节收集的数据集上的性能，并将其与文献中的 9 种代表性基线检测器进行比较，涵盖基于统计量和基于机器学习的方法：(i) 似然（Gehrmann 等人，2019）；(ii) 对数秩（Gehrmann 等人，2019）；(iii) 对数秩比（LRR，Su 等人，2023）；(iv) Fast-DetectGPT（FDGPT，Bao 等人，2024）；(v) Binoculars（Hans 等人，2024）；(vi) RoBERTa（Solaiman et al. 2019）；(vii) RADAR（Hu 等人，2023）；(viii) 先模仿后检测（ImBD，Jiaqi 等人，2025）；(ix) AdaDetectGPT（Zhou 等人，2025）。由于这些检测器主要是为分类而非统计推断设计的，因此通常会构造统计量 S，却不指定用于控制第一类错误的分类阈值 c。因此，我们使用曲线下面积（AUC）评估其性能，该指标用于衡量S 独立于 c的质量。

为确保公平　　为了进行比较，我们将第3 节中介绍的八种数据类别划分为

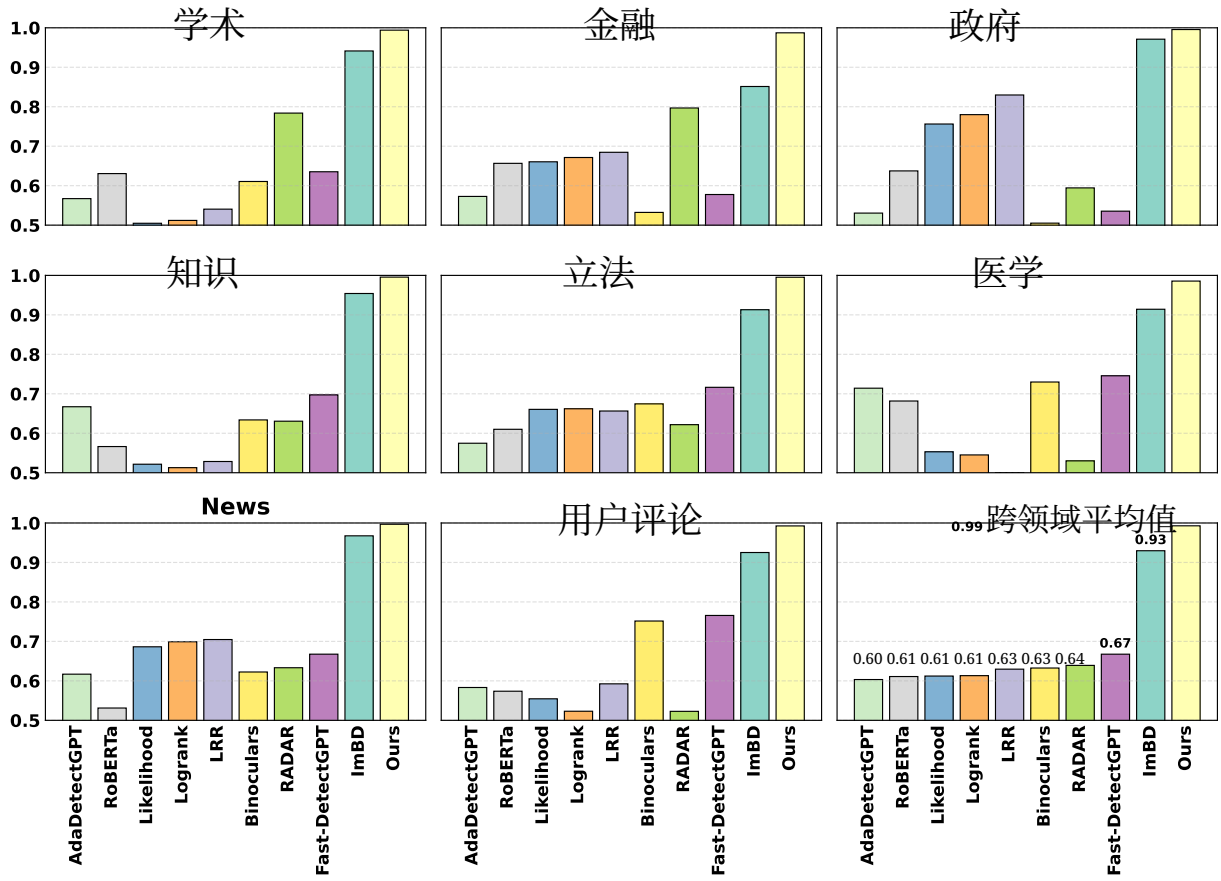

图5：各种检测器在第3节所述的八个数据领域上训练和评估时的曲线下面积。每个面板报告一个领域的曲线下面积。右下方面板报告八个领域的平均曲线下面积。

训练子集和测试子集。所有方法均使用相同的训练数据进行训练，并使用相同的测试数据计算曲线下面积（AUC）分数。此外，对于所有需要采样的方法（例如 FastDetectGPT 和 AdaDetectGPT），我们在构建统计量时使用相同的采样模型 $q^{\mathcal{S}}$ 。

我们在图 5中报告了这些检测器在不同数据领域上训练和评估时的 AUC值。可以看出：(i) 我们所提出的检测器在所有领域中的曲线下面积始终接近 1.0，优于所有基线方法，且跨领域的平均曲线下面积超过 0.99；(ii) *ImBD* 表现次佳，但我们的方法

表 2：我们的方法在八类人工撰写文本上的第一类错误，分别对应三个名义显著性水平 α。

α	学术	金融	政府	知识	立法	医学	News	用户评论
0.01	0.000	0.000	0.013	0.007	0.013	0.013	0.027	0.000
0.05	0.007	0.020	0.053	0.047	0.033	0.040	0.067	0.060
0.10	0.053	0.073	0.093	0.073	0.080	0.080	0.113	0.100

表 3：我们的方法在三个名义显著性水平 α 下对八类大语言模型生成文本的检验功效。

α	学术	金融	政府	知识	立法	医学	News	用户评论
0.01	0.933	0.827	0.987	0.947	0.900	0.873	0.973	0.953
0.05	0.967	0.933	0.993	0.980	0.973	0.960	0.993	0.973
0.10	0.980	0.953	0.993	0.987	0.987	0.960	0.993	0.973

大幅优于它，相对提升幅度为我们的方法曲线下面积–ImBD的曲线下面积1–ImBD的曲线下面积，在大多数情况下超过 90%；(iii) 所有其他基线方法的表现都显著逊于$ImBD$。

最后，我们通过评估检测器在三个显著性水平下的第一类错误（即FPR）和检验功效（即TPR），来考察所提出的分类阈值选择程序 c，$\alpha \in \{0.01, 0.05, 0.1\}$。八个数据类别的经验第一类错误率和检验功效分别报告于表2 和3中。如这些表所示，第一类错误率在名义水平上得到了良好控制，而经验检验功效在大多数情况下接近1 。这些结果与我们在图5中的发现一致，并证明了所提出检测器用于统计推断的有效性。

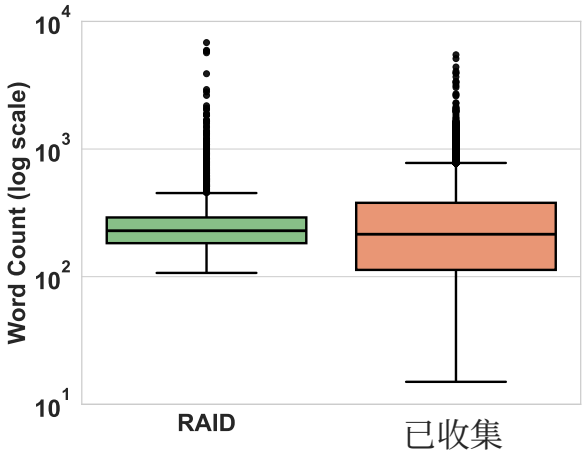

图6：RAID 和我们收集的数据集中人工撰写文本的词数分布。

5.2 分布外评估

本节中，我们考察所提出检测器在外部基准数据集RAID上的性能[1]（Dugan等人（2024））。该数据集广泛用于文献中对大语言模型生成文本检测器进行基准测试。它在三个方面不同于第3 节所述的我们收集的数据集。首先，两个数据集中的人工撰写文本在内容上有所不同。图6展示了RAID和我们收集的数据集中人工撰写文本的词数分布。尽管RAID中的平均文本长度与我们收集的数据集相近（且略长），但其变异性显著更小。其次，RAID中的大语言模型生成文本并非由我们数据集中使用的同一组语言模型生成。最后，RAID还包括在11种对抗性攻击设置和4种大语言模型解码策略下生成的文本（Dugan等人（2024））。

为了评估检测器在RAID数据集上的表现，训练时我们没有采用第 5.1 节所述的样本拆分程序，而是使用全部数据进行检测。表 4报告了所提出的检测器的经验I类错误率和检验功效。如表中所示，I类错误率总体上得到了良好控制。对于 $\alpha = 0.01$ 和 0。05，经验

[1]https://huggingface.co/datasets/liamdugan/raid

表4：所提出的检测器在RAID数据集中的2000篇人工撰写文本和2000篇GPT-4生成文本上的经验I类错误率和检验功效。

α	0.01	0.05	0.10
第一类错误率	0.007	0.050	0.127
经验检验功效	0.450	0.780	0.906

第一类错误率不高于名义显著性水平，仅在$\alpha = 0.1$时略有膨胀。此外，经验检验功效随α增加而提高，在 $\alpha = 0.05$时接近 80%，并在 $\alpha = 0.1$时超过 90%。鉴于 RAID 中的大语言模型生成文本采用了不同于训练阶段的解码温度，并且是在对抗性设置下生成的，这些结果尤为出色。这表明，在分布偏移情况下，即测试数据与训练数据不同时，我们的方法仍能为统计推断提供可靠保证。

最后，我们根据分类准确率，将检测器与第5.1节中考察的同一组基线方法进行比较。同样，我们使用曲线下面积作为评估指标。结果汇总于表5。可以看出，我们的检测器仍取得了最高的曲线下面积，超过 0.95。相比之下，最佳基线检测器的曲线下面积不超过 0.88。最后，值得注意的是，尽管$ImBD$ 在训练数据和测试数据服从相同分布时（图5 ）于第 5.1 节中表现良好，但在该外部数据集上的曲线下面积大幅下降至 0.8 以下。这些结果表明，与$ImBD$等基线检测器相比，我们的方法对训练数据与测试数据之间的分布偏移具有显著更强的鲁棒性。

表5：在 RAID 上评估的各种检测器的 AUC值。

	AdaDetectGPT	Binoculars	Fast-DetectGPT	**Ours**	ImBD
AUC	0.688	0.875	0.862	0.954	0.771

	似然	对数秩	LRR	RoBERTa	RADAR
AUC	0.831	0.817	0.755	0.555	0.867

5.3 计算成本

我们在图7中报告了所提出检测程序的运行时间（秒）和内存使用量（吉字节），并展示其随输入文本中词元数量的变化。我们有以下观察结果。首先，该程序通常能在几秒内完成。同时，运行时间随词元数量近似呈二次方增长，这与基于 Transformer 的架构的计算复杂度一致；在这类架构中，自注意力机制会计算输入序列中所有词元之间的成对交互（Vaswani 等，2017）。其次，在大多数情况下，内存使用量低于 8 GB，并随词元数量近似线性增长。（iii）我们的评估所用输入长度约从 20 个词元到超过 $2^{12} = 4096$ 个词元不等。综合来看，这些结果表明，我们的方法能够在标准图形处理器硬件上轻松处理短文档和中等长度的长文档，例如短篇文章和报告、新闻标题及文章。

6 案例研究

我们提供一个可公开访问的网站[2]，用于托管我们训练好的检测器。这有两个目的。首先，用户无需下载

[2]https://huggingface.co/spaces/stats-powered-ai/StatDetectLLM

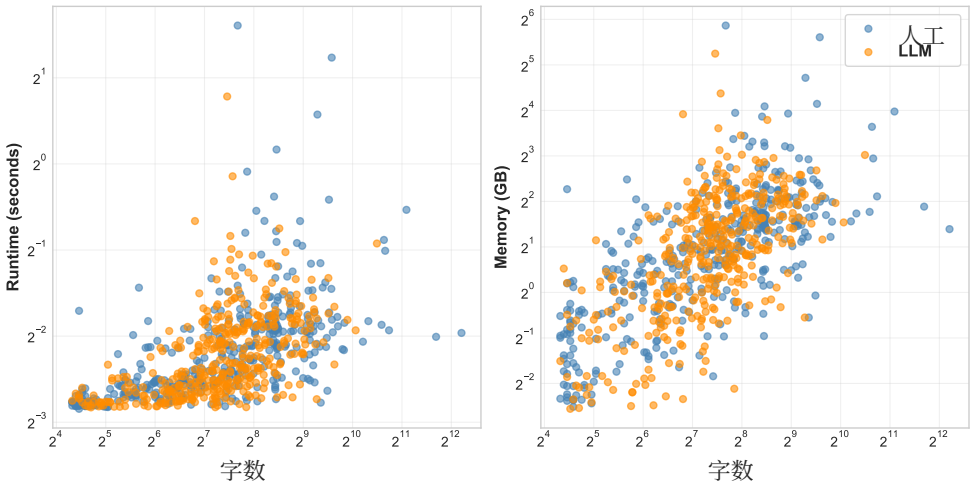

图 7：左图：我们方法的运行时间（纵轴为 y）与待检测文本中的词元数量（横轴为 x）之间的关系。右图：内存使用量（纵轴为 y）与词元数量之间的关系。运行时间以秒为单位，内存使用量以吉字节（GB）为单位。

模型到本地，也无需使用我们的数据重新训练模型，即可直接使用我们的检测器。其次，大语言模型不断发展，更新速度很快。对于几年前发布的大语言模型有效的检测器，可能会随着这些模型被替换或更新而过时。通过在线托管我们的检测器，我们计划定期更新用于生成训练数据的热门大语言模型集合，并据此重新训练检测器，确保它能跟上大语言模型的快速发展。

接下来，我们将介绍如何使用我们的检测器；图 2 展示了我们网站的快照。上方面板说明用户提供的输入，并由以下组件组成：

- **输入框**（灰色）：用户可以在此处粘贴待检测的文本。

- **领域列表**（灰色）：用户可以指定输入文本所属的领域。第 3 节中介绍的八个领域均可选择。如果未指定此类信息，则默认使用“通用”选项，并且我们的检测器会报告所有八个领域中的最大 p值，以控制第一类错误（详情见第 4.2 节）。

- **显著性　　级别** α **滑块**（红色）：用户可以调整滑块以选择合适的

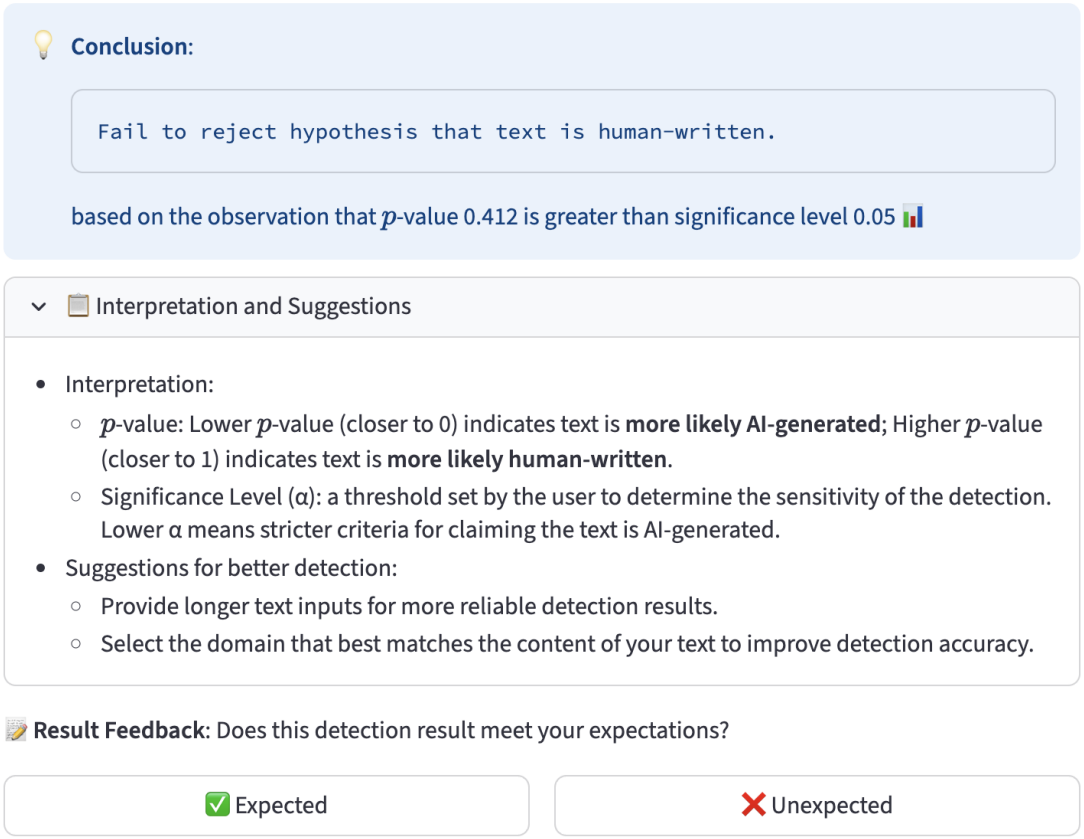

图 8：检测器的输出是一个结论面板，用于指示文本是由人类撰写还是由大语言模型生成，并附带相应的 p-值。

显著性水平介于 0.01 和 0.20 之间。若未指定数值，则默认显著性水平设为 0.05。

- **检测按钮** （橙色）：点击此按钮可处理输入文本，此时我们的检测器 (i) 计算式 (7) 中的统计量，(ii) 使用选定领域计算相应的 p值，并 (iii) 根据用户指定的显著性水平得出结论。

点击检测按钮后，结果将显示出来（示例见图 8 ）。结果面板会呈现检测器得出的结论：

- **结论** box （蓝色）：它显示了我们的统计假设检验结论，

并说明得出该结论的依据：报告根据经验零分布计算出的相应 p 值，以及用户指定的显著性水平 α。

- **解释与建议**（蓝色）：向统计学界以外的用户解释如何解读 p-值以及显著性水平 α 的含义，同时提供有关如何使用该模型的实用指导。

- **用户反馈面板**（白色）：允许用户指出检测结果是否符合其预期。这有助于我们收集反馈数据，以便今后改进。作为一个具体示例，我们将 Li & Yu（2021）论文的摘要粘贴到图 2上方面板的输入框中。该论文发表于 2021 年，早于 ChatGPT 发布，因此很可能由人类撰写。我们的检测器报告的 p-值为 0.412，因此无法拒绝文本由人类撰写这一零假设。

接下来，我们提示 GPT-5 改写同一篇摘要（提示词和生成文本见补充材料第 C.2 节），并将改写后的版本粘贴到输入框中。如图 2下方的面板所示，所得 p-值非常小（0.0000），因此拒绝零假设。这个案例研究展示了我们的方法在区分人工撰写的文本和大语言模型生成文本方面的有效性。

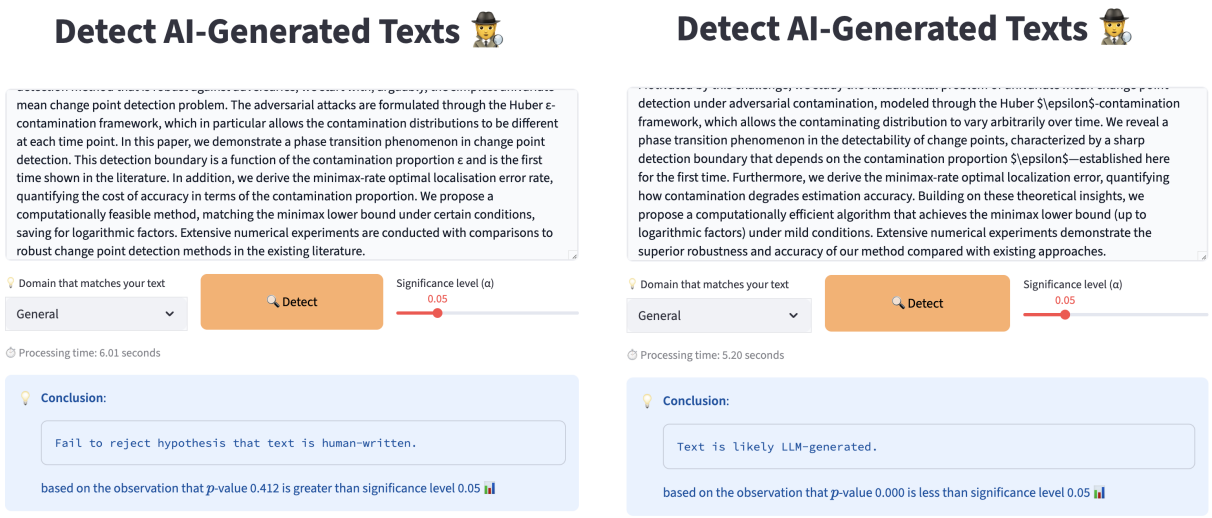

图 9：案例研究：检测论文 Li &Yu（2021）的摘要（左图）

及其改写版本（右图）是否由大语言模型生成。

7 讨论

本文开发了一款公众可用的大语言模型检测工具。与文献中现有的检测器相比，我们的方法不依赖水印，也不需要了解用于生成文本的具体大语言模型。此外，在控制第一类错误的同时，该工具在区分人工撰写文本和大语言模型撰写文本方面展现出更优的实证性能。鉴于生成式人工智能能力迅速发展，涵盖文本、图像和视频，将我们提出的方法扩展到检测此类人工智能生成内容仍是未来研究的重要方向。

参考文献

Aaronson, S.& Kirchner, H.(2023年), 大语言模型水印, 载于《西蒙斯计算理论研究所大语言模型与Transformer研讨会》。

Abburi, H., Roy, K., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E.& Bhattacharya, S. (2023年), 一种简单而高效的人工智能生成文本检测集成方法, 载于S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus & H. Sedghamiz, 编, 《第三届自然语言生成、评估与度量研讨会论文集（GEM）》, 第413–421页。

Anthropic (2024), 'Claude 3：新一代人工智能模型', https://www.anthropic.com/claude。

Arora, A. & Arora, A. (2023年), '大语言模型在医疗保健领域的前景', 《柳叶刀》401(10377), 641。

arXiv.org 投稿者（2024）, 'arxiv数据集', https://www.kaggle.com/dsv/7548853。

Bao, G., Zhao, Y., Teng, Z., Yang, L. & Zhang, Y.（2024），Fast-DetectGPT：通过条件概率曲率高效进行机器生成文本的零样本检测，发表于‘第十二届国际学习表征会议’。

Bolthausen, E.（1982年），《某些鞅中心极限定理中的精确收敛率》，《概率年刊》10(3)，672 – 688。

Chalkidis, I., Fergadiotis, M. & Androutsopoulos,I.（2021），Multieurlex——用于零样本跨语言迁移的多语言多标签法律文档分类数据集，载于《2021年自然语言处理经验方法会议论文集》，计算语言学协会。

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M. & Aletras, N.（2022年），Lexglue：用于英语法律语言理解的基准数据集，载于《计算语言学协会第六十届年会论文集》，爱尔兰都柏林。

Chan, C. K. Y. & Hu, W.（2023），《学生对生成式人工智能的看法：高等教育中的认知、益处与挑战》，《高等教育教育技术国际期刊》20(1)，43。

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W. &Goharian, N. (2018), ‘长文档抽象式摘要的话语感知注意力模型’，*arXiv*预印本 *arXiv:1804.05685*。

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E. 等（2025），《Gemini 2.5：以高级推理、多模态、长上下文和下一代智能体能力推动前沿》，*arXiv* 预印本 *arXiv:2507.06261*。

Crothers, E. N., Japkowicz, N. & Viktor, H. L.（2023），《机器生成文本：威胁模型与检测方法的综合调查》，*IEEE Access* **11**，70977–71002。

Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J. M., Xu, H., Ippolito, D. & Callison- Burch, C. (2024), RAID：用于稳健评估机器生成文本检测器的共享基准，载于《第62届计算语言学协会年会论文集（第1卷：长篇论文）》。

Flowers, J. G. (2025),《Finance-instruct-500k》，https://huggingface.co/datasets/Josephgflowers/Finance-Instruct-500k。

Foundation, W.（无日期），《维基媒体下载》，https://dumps.wikimedia.org。

Gehrmann, S., Strobelt, H. & Rush, A. M. (2019),《GLTR：生成文本的统计检测与可视化》，*arXiv* 预印本 *arXiv:1906.04043*。

Greene, D. & Cunningham, P. (2006年), 核文档聚类中对角占优问题的实用解决方案, 载于 《第23届国际机器学习会议论文集》，第377–384页。

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012),《核双样本检验》，机器学习研究期刊 **13**(1), 723–773。

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J. & Wu, Y. (2023年),《ChatGPT 与人类专家有多接近？比较语料库、评估与检测》，*arXiv* 预印本 *arXiv:2301.07597*。

Guo, H., Cheng, S., Jin, X., Zhang, Z., Zhang, K., Tao, G., Shen, G. & Zhang, X. (2024),《Biscope：通过检查对前置词元的记忆进行人工智能生成文本检测》，神经信息处理系统进展**37**, 104065–104090。

Hall, P. & Heyde, C. C. (2014), 鞅极限定理及其应用，学术出版社。

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J. & Goldstein, T. (2024),《用双筒望远镜识别大语言模型：机器生成文本的零样本检测》， *arXiv* 预印本 *arXiv:2401.12070* 。

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J. & Wang, H. (2024), '软件工程中的大语言模型：系统性文献综述',*ACM*软件工程与方法学汇刊 **33**(8), 1–79.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al. (2022), 'LoRA：大语言模型的低秩适配。', 国际学习表征会议**1**(2), 3.

Hu, X., Chen, P.-Y. & Ho, T.-Y. (2023), 'RADAR：通过对抗学习实现稳健的人工智能文本检测',神经信息处理系统进展 **36**, 15077–15095.

Huang, L., Cao, S., Parulian, N., Ji, H. & Wang, L. (2021), 面向长文档摘要的高效注意力机制, 载于 '《2021年北美计算语言学协会分会会议论文集：人类语言技术》', 计算语言学协会，线上，第 1419–1436 页。

Huang, W., Murakami, A. & Grieve, J. (2025),'通过作者语言模型的困惑度归因作者身份',*PLOS ONE* **2**0(7),e0327081.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford,A.等(2024),'GPT-4o系统卡',*arXiv*预印本 *arXiv:2410.21276* .

Ippolito, D., Duckworth, D., Callison-Burch, C. & Eck, D. (2020), 当人类受骗时，生成文本的自动检测最容易，载于D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, 编，'第58届计算语言学协会年会论文集'，计算语言学协会，页码1808–1822。

Jiaqi, C., Zhu, X., Liu, T., Chen, Y., Chen, X., Yuan, Y., Tou, L. C., Li, Z., Tang, L. & Zhang, L. (2025), 先模仿再检测：为机器改写文本检测对齐机器风格偏好，载于 ‘AAAI 人工智能会议论文集’，第39卷，页码23559–23567。

Khandekar, N., Jin, Q., Xiong, G., Dunn, S., Applebaum, S., Anwar, Z., Sarfo-Gyamfi, M., Safranek, C., Anwar, A., Zhang, A. 等（2024），《Medcalc-bench：评估用于医学计算的大语言模型》，神经信息处理系统进展**37**, 84730–84745。

Kobak, D., González-Márquez, R., Horvát, E.-Á. & Lause, J.（2025），《通过过量词汇探究生物医学出版物中的大语言模型辅助写作》，《科学进展》11(27)，eadt3813。

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V. & West, R.（2024），《人工智能审稿彩票：广泛使用人工智能辅助同行评审会提高论文评分和录用率》，*arXiv* 预印本 *arXiv:2405.02150* 。

Lee, H., Tack, J. & Shin, J. (2024), Remodetect：奖励模型能够识别对齐的大语言模型生成文本，载于 《第三十八届神经信息处理系统年会》。

Li, M. & Yu, Y. (2021),《对抗鲁棒变点检测》，神经信息处理系统进展**34**，22955–22967。

Li, X., Ruan, F., Wang, H., Long, Q. &Su, W.J. (2025a)，《人工编辑下大语言模型水印的鲁棒检测》，英国皇家统计学会杂志*B*辑：统计方法论，第 qkaf056 页。

Li, X., Ruan, F., Wang, H., Long, Q. & Su, W. J. (2025b)，《大语言模型水印的统计框架：

Pivot、检测效率与最优规则》，《统计年鉴》53(1)，322–351。

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng,C., Zhang,C., Ruan, C. 等（2024），《Deepseek-v3 技术报告》，*arXiv* 预印本 *arXiv:2412.19437*。

Liu, Z., Guo, X., Yang, Z., Lou, F., Zeng, L., Li, M., Qi, Q., Liu, Z., Han, Y., Cheng, D., Feng, X., Wang, H. J., Shi, C. & Zhang, L.（2025），《Fin-r1：通过强化学习实现的金融推理大语言模型》，*arXiv* 预印本 *arXiv:2503.16252*。

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C.（2011），学习用于情感分析的词向量，载于《计算语言学协会第四十九届年会论文集：人类语言技术》，计算语言学协会，美国俄勒冈州波特兰，第142–150页。

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M. & Balahur, A.（2018），WWW'18 开放挑战：金融观点挖掘与问答，载于《2018年世界互联网大会伴随论文集》，WWW '18，国际万维网会议指导委员会，第 1941–1942 页。

Mao, C., Vondrick, C., Wang, H. & Yang, J.（2024），Raidar：通过改写检测生成式人工智能，载于《第十二届国际学习表征会议》。

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. & Finn, C.（2023），DetectGPT：利用概率曲率进行零样本机器生成文本检测，载于《国际机器学习会议》，机器学习研究会议录，第 24950–24962 页。

Mitrović, S., Andreoletti, D. & Ayoub, O.（2023年），《ChatGPT还是人类？检测并解释：解释用于检测短篇ChatGPT生成文本的机器学习模型所作出的决策》，*arXiv*预印本 *arXiv:2301.13852* 。

Narayan, S., Cohen, S. B. & Lapata, M.（2018），《别给我细节，只给我摘要！用于极限摘要的主题感知卷积神经网络》，*arXiv:1808.08745* 。经济合作与发展组织（2024），《人工智能治理：政府准备好了吗？》。

Pal, A., Umapathi, L. K. & Sankarasubbu, M.（2022年），Medmcqa：用于医学领域问答的大规模多学科选择题数据集，载于《健康、推理与学习会议论文集》，第 174 卷《机器学习研究会议录》，PMLR，第 248–260 页。

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S. 等（2019），《语言模型的发布策略及其社会影响》，*arXiv* 预印本 *arXiv:1908.09203* 。

Song, Y., Yuan, Z., Zhang, S., Fang, Z., Yu, J. & Liu, F.（2025），用于机器生成文本检测的深度核相对检验，载于 《第十三届国际学习表征会议》。

Su, J., Zhuo, T. Y., Wang, D. & Nakov, P.（2023年），《DetectLLM：利用对数秩信息进行机器生成文本的零样本检测》，*arXiv:2306.05540* 。

Tulchinskii, E., Kuznetsov, K., Kushnareva, L.,Cherniavskii, D., Nikolenko,S., Burnaev, E., Barannikov, S. & Piontkovskaya, I.（2023年），《用于稳健检测人工智能生成文本的内在维度估计》，神经信息处理系统进展**36**，39257–39276。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I.（2017），《注意力就是你所需要的一切》，神经信息处理系统进展**30**。

Weidinger，L.，Me llor, J.、Rauh, M.、Griffin, C.、Uesato, J.、Huang, P.-S.、Cheng, V.、Balle ,

B., Kasirzadeh, A., Biles, C. 等（2021），《大语言模型的伦理与社会风险》，*arXiv* 预印本 *arXiv:2112.04359*。

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S. & Wong, D. F.（2025），《大语言模型生成文本检测综述：必要性、方法与未来方向》，计算语言学，第1–66页。

xAI（2025），《Grok（4版）》，https://grok.x.ai。

Xie, Y., Li, X., Mallick, T., Su, W. & Zhang, R.（2025），《通过最大耦合为大语言模型去偏水印》，美国统计学会会刊0（0），1–11。

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C. 等（2025），《通义千问3技术报告》，*arXiv*预印本 *arXiv:2505.09388* 。

Yang, X., Cheng, W., Wu, Y., Petzold, L. R., Wang, W. Y. & Chen, H. (2024), DNA-GPT：用于免训练检测 GPT生成文本的发散式 n 元语法分析，载于 ‘第十二届国际学习表征会议’。

Zhang, S., Song, Y., Yang, J., Li, Y., Han, B. & Tan, M. (2024), 通过面向多群体的优化最大化均值差异，以检测机器生成文本，载于‘第十二届国际学习表征会议’。

Zhang, X., Zhao, J. & LeCun, Y. (2015), ‘用于文本分类的字符级卷积网络’，神经信息处理系统进展 **28**。

Zhou, H., Zhu, J., Su, P., Ye, K., Yang, Y., Gavioli-Akilagun, S. A. O. B. & Shi, C. (2025), AdaDetectGPT：具有统计保证的大语言模型生成文本自适应检测，载于 ‘第三十九届神经信息处理系统年会’。

Zhu, X., Ren, Y., Cao, Y., Lin, X., Fang, F. & Li, Y. (2025), ‘可靠地界定假阳性

：一种基于多尺度保形预测的零样本机器生成文本检测框架’，*arXiv* 预印本 *arXiv:2505.05084* 。

A 证明

A.1 定理1

对于任意固定的 t，回顾一下，$q_t^{\mathcal{M}}$定义为

$$q_t^{\mathcal{M}}(x|\boldsymbol{x}_{<t}) = \frac{\exp\left(\frac{1}{\tau}\ell_t^{\mathcal{M}}(x|\boldsymbol{x}_{<t})\right)}{\sum_{x\in\mathcal{V}}\exp\left(\frac{1}{\tau}\ell_t^{\mathcal{M}}(x|\boldsymbol{x}_{<t})\right)}.$$

设$\mathcal{A}_t(\boldsymbol{x}_{<t}) = \arg\max_{v\in\mathcal{V}} \ell_t^{\mathcal{M}}(v|\boldsymbol{x}_{<t})$为使 $\ell_t^{\mathcal{M}}(\bullet|\boldsymbol{x}_{<t})$达到最大值的词元集合。由此可知，当 $\tau \to 0$时，对于任意 $x \in \mathcal{V}$，$q_t^{\mathcal{M}}(x|\boldsymbol{x}_{<t}) \to |\mathcal{A}_t|^{-1}\mathbb{I}\{x \in \mathcal{A}_t(\boldsymbol{x}_{<t})\}$。当$\mathcal{S} = \mathcal{M}$时，对于所有 $t \geq 1$ 以及任意分布 $\mathbb{P}$,

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) \leq \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\{\max_{v\in\mathcal{V}}\ell_t^{\mathcal{M}}(v|\boldsymbol{X}_{<t})\} = \mathbb{E}_{\substack{\boldsymbol{X}_{<t}\sim\mathbb{P}\\ \widetilde{X}_t\sim q_t^{\mathcal{M}}(\bullet|\boldsymbol{X}_{<t})}}[\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})]. \tag{10}$$

因此,

$$\sum_t \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \sum_t \mathbb{E}_{\substack{\boldsymbol{X}_{<t}\sim\mathbb{P}\\ \widetilde{X}_t\sim q_t^{\mathcal{M}}(\bullet|\boldsymbol{X}_{<t})}}[\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})] \leq 0. \tag{11}$$

这表明 $\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}S_{\text{Fast}}(\boldsymbol{X}) \leq 0$。另一方面，当温度 $\tau \to 0$时，若$\mathbf{X} \sim \mathbb{Q}^{\mathcal{M}}$,

$$\ell_t^{\mathcal{M}}(X_t|\boldsymbol{X}_{<t}) - \mathbb{E}_{\substack{\boldsymbol{X}_{<t}\sim\mathbb{Q}^{\mathcal{M}}\\ \widetilde{X}_t\sim q_t^{\mathcal{M}}(\bullet|\boldsymbol{X}_{<t})}}[\ell_t^{\mathcal{M}}(\widetilde{X}_t|\boldsymbol{X}_{<t})] = 0$$

几乎必然地，这表明 $\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}S_{\text{Fast}}(\boldsymbol{X}) = 0$。证明至此完成。

A.2 定理 2 的证明

我们首先介绍证明定理 2 所需的技术条件 .

假设 **定理 1** (方差相等)**.** 对于任意非恒定见证函数 w，定义

$$\sigma_{q,L}^2 := \frac{1}{L}\sum_{t=1}^{L}\mathrm{Var}_{\widetilde{X}_t\sim q_t}\left(w_t(\widetilde{X}_t|\widetilde{X}_{<t})\right), \quad \sigma_{p,L}^2 := \frac{1}{L}\sum_{t=1}^{L}\mathrm{Var}_{\widetilde{X}_t\sim p_t}\left(w_t(\widetilde{X}_t|\widetilde{X}_{<t})\right).$$

$\sigma_{q,L}^2$,$\sigma_{p,L}^2$ 几乎必然有某个常数 $\sigma_w^2 > 0$ 作为下界。此外，当 $L \to \infty$ 时，$\sigma_{q,L}/\sigma_{p,L} \to 1$依概率收敛。

假设 2. 对于任意见证函数 w，定义

$$\bar{\sigma}_{q,L}^2 = \frac{1}{L}\sum_{t=1}^{L} \mathrm{Var}_{\boldsymbol{X}\sim q}\left(w_t(\widetilde{X}_t|\widetilde{X}_{<t})\right), \quad \bar{\sigma}_{p,L}^2 = \frac{1}{L}\sum_{t=1}^{L} \mathrm{Var}_{\boldsymbol{X}\sim p}\left(w_t(\widetilde{X}_t|\widetilde{X}_{<t})\right).$$

若 $\boldsymbol{X} \sim q$，则 $\bar{\sigma}_{q,\ L}^2/\sigma_{q,\ L}^2 \to 1$ 依概率成立。若 $\boldsymbol{X} \sim p$，则 $\bar{\sigma}_{p,\ L}^2/\sigma_{p,\ L}^2 \to 1$ 依概率成立。

引理 1. 设 $\boldsymbol{X} = (X_1, \ldots X_n)$ 是实值随机变量序列，且对所有 $1 \le t \le n$,

$$\mathbb{E}(X_t|X_{<t}) = 0 \quad \text{almost surely.}$$

设 $\sigma_t^2 = \mathbb{E}(X_t^2\big|X_{<t})$、$\bar{\sigma}_t^2 = \mathbb{E}(X_t^2)$、$s_n^2 = \sum_{t=1}^n \bar{\sigma}_t^2$和 $V_n^2 = \sum_{t=1}^n \sigma_t^2/s_n^2$。假设对所有 n，$|X_n|$ 几乎必然被某个常数所界，且 $s_n/\sqrt{n}$ 远离零，并且$V_n^2 \to 1$ 属于 L^1。那么

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sum_{t=1}^n X_t}{\sqrt{\sum_{t=1}^n \sigma_t^2}} \le z\right) - \Phi(z)\right| \to 0,$$

其中，$\Phi(\bullet)$是标准正态分布的累积分布函数。

证明。结论可直接由鞅中心极限定理推出，参见 Bolthausen (1982年)。

□

引理 2. 设 X 是一个随机变量。令 Φ 和 ϕ 分别为标准正态分布的累积分布函数和概率密度函数。则对于任意随机变量 X,

$$\mathbb{E}\Phi(z_\alpha + X) \ge \min\{1-\alpha, \alpha + \Phi'(z_\alpha)\mathbb{E}X\},$$

其中，$0 < \alpha < 1/2$, z_α 是标准正态分布的第 α 分位数。

证明。 该证明直接源自 Zhou 等人（2025）的引理S2。 □

现在，我们开始证明定理2。注意到，我们的检验统计量可以分解为$S(\boldsymbol{X}) = S^{(1)}(\boldsymbol{X}) - S^{(2)}(\boldsymbol{X})$ 其中 $S^{(1)}(\boldsymbol{X})$ $S^{(2)}(\boldsymbol{X})$ 定义为:

$$\begin{aligned} S^{(1)}(\boldsymbol{X}) &= \frac{\sum_t [w_t(X_t|X_{<t}) - \mathbb{E}_{\widetilde{X}_t \sim p_t} w_t(\widetilde{X}_t|X_{<t})]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t \sim q_t}(w_t(\widetilde{X}_t|X_{<t}))}} \\ S^{(2)}(\boldsymbol{X}) &= \frac{\sum_t [\mathbb{E}_{\widetilde{X}_t \sim q_t} w_t(\widetilde{X}_t|X_{<t}) - \mathbb{E}_{\widetilde{X}_t \sim p_t} w_t(\widetilde{X}_t|X_{<t})]}{\sqrt{\sum_t \mathrm{Var}_{\widetilde{X}_t \sim q_t}(w_t(\widetilde{X}_t|X_{<t}))}}, \end{aligned} \tag{12}$$

真阴性率可以表示为

$$\mathbb{P}_{\boldsymbol{X}\sim p}(S(\boldsymbol{X}) \le z_\alpha) = \mathbb{P}_{\boldsymbol{X}\sim p}\left(S^{(1)}(\boldsymbol{X}) \le z_\alpha + S^{(2)}(\boldsymbol{X})\right) \tag{13}$$

容易验证，当 $\boldsymbol{X} \sim p$, $S^{(1)}(\boldsymbol{X})\sigma_{q,L}/\sigma_{p,L}$ 收敛于标准正态分布。具体而言，根据引理 1，我们得到

$$\begin{aligned} \mathbb{P}_{\boldsymbol{X}\sim p}(S(\boldsymbol{X}) \le z_\alpha) &= \mathbb{P}_{\boldsymbol{X}\sim p}\left(S^{(1)}(\boldsymbol{X})\frac{\sigma_{q,L}}{\sigma_{p,L}} \le (z_\alpha + S^{(2)}(\boldsymbol{X}))\frac{\sigma_{q,L}}{\sigma_{p,L}}\right) \\ &\ge \Phi(z_\alpha + S^{(2)}(\boldsymbol{X})) + \left(\Phi\left((z_\alpha + S^{(2)}(\boldsymbol{X}))\frac{\sigma_{q,L}}{\sigma_{p,L}}\right) - \Phi(z_\alpha + S^{(2)}(\boldsymbol{X}))\right) \\ &\quad + o_p(1) \\ &\ge \Phi(z_\alpha + S^{(2)}(\boldsymbol{X})) - \sup_{z\in\mathbb{R}}|\phi(z)| \times \left|z_\alpha + S^{(2)}(\boldsymbol{X})\right| \times \left|\frac{\sigma_{q,L}}{\sigma_{p,L}} - 1\right| + o_p(1) \end{aligned}$$

在假设 1 下， $\sigma_{q,L}/\sigma_{p,L} \to 1$ 依概率收敛，我们得到

$$\mathbb{P}_{\boldsymbol{X}\sim p}(S(\boldsymbol{X}) \le z_\alpha) \ge \Phi(z_\alpha + S^{(2)}(\boldsymbol{X})) + o_p(1).$$

此外，余项 $o_p(1)$ 是一致可积的，因为 $\mathbb{P}_{\boldsymbol{X}\sim p}(S(\boldsymbol{X}))$ 和 $\Phi(z_\alpha + S^{(2)}(\boldsymbol{X}))$ 均有上界。对等式两边取期望，根据假设1，可得

$$\mathbb{P}_{\boldsymbol{X}\sim p}(S(\boldsymbol{X}) \le z_\alpha) \ge \mathbb{E}\Phi(z_\alpha + S^{(2)}(\boldsymbol{X})) + o(1).$$

接下来，定义 $\widetilde{\sigma}^2_{q,\ L} = \mathbb{E}_{\boldsymbol{X}\sim p}\sigma^2_{q,\ L}$。由此可得 $L_w = \mathbb{E}\left\{S^{(2)}(\boldsymbol{X})\frac{\sigma_{q\ L,\ ,\ L}}{\widetilde{\sigma}_q}\right\}$。根据假设 1 中的等方差假设，我们还有 $\sigma_{q,\ L} - \widetilde{\sigma}_{q,\ L} \to 0$ 依概率成立。由此可得

对于任意 $\epsilon > 0$,

$$
\begin{aligned}
& \mathbb{E}\Phi(z_\alpha + S^{(2)}(\boldsymbol{X})) \qquad (14)\\
= \;& \mathbb{E}\Phi(z_\alpha + S^{(2)}(\boldsymbol{X}))\mathbb{I}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| \le \epsilon\} \\
& \quad + \mathbb{E}\Phi(z_\alpha + S^{(2)}(\boldsymbol{X}))\mathbb{I}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| > \epsilon\} \\
\ge \;& \mathbb{E}\Phi(z_\alpha + S^{(2)}(\boldsymbol{X}))\mathbb{I}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| \le \epsilon\} \\
\ge \;& \mathbb{E}\Phi\left(z_\alpha + S^{(2)}(\boldsymbol{X})\frac{\sigma_{q,L}}{\widetilde{\sigma}_{q,L} + \mathrm{sgn}(S^{(2)}(\boldsymbol{X}))\epsilon}\right)\mathbb{I}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| \le \epsilon\} \\
\ge \;& \mathbb{E}\Phi\left(z_\alpha + S^{(2)}(\boldsymbol{X})\frac{\sigma_{q,L}}{\widetilde{\sigma}_{q,L} + \mathrm{sgn}(S^{(2)}(\boldsymbol{X}))\epsilon}\right) \\
& \quad - \mathbb{E}\Phi\left((z_\alpha + S^{(2)}(\boldsymbol{X}))\frac{\sigma_{q,L}}{\widetilde{\sigma}_{q,L} + \mathrm{sgn}(S^{(2)}(\boldsymbol{X}))\epsilon}\right)\mathbb{I}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| > \epsilon\} \\
\ge \;& \mathbb{E}\Phi\left((z_\alpha + S^{(2)}(\boldsymbol{X}))\frac{\sigma_{q,L}}{\widetilde{\sigma}_{q,L} + \mathrm{sgn}(S^{(2)}(\boldsymbol{X}))\epsilon}\right) - \mathbb{P}(|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| > \epsilon),
\end{aligned}
$$

其中，第一个不等式是由于 Φ非负而得到的，第二个不等式则是由于 Φ的单调性和有界性而成立。结合引理2 和假设1，我们得到

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{X}\sim p}\left(S(\boldsymbol{X}) \le z_\alpha\right) \ge & \min\left\{1-\alpha, \alpha + \phi(z_\alpha)\mathbb{E}\left\{S^{(2)}(\boldsymbol{X})\frac{\sigma_{q,L}}{\widetilde{\sigma}_{q,L}}\right\}\right\}\frac{\widetilde{\sigma}_{q,L}}{\widetilde{\sigma}_{q,L} + \mathrm{sgn}(S^{(2)}(\boldsymbol{X}))\epsilon} \qquad (15)\\
& - \mathbb{P}\{|\sigma_{q,L} - \widetilde{\sigma}_{q,L}| \ge \epsilon\} + o(1).
\end{aligned}
$$

令 $L \to \infty$，并利用 $\mathbb{E}\left\{S^{(2)}(\boldsymbol{X})\frac{\sigma_{q,L,,L}}{\widetilde{\sigma}_q}\right\} = L_w$，可得真阴性率在渐近意义下的下界为 $\min\{1-\alpha, \alpha + \phi(z_\alpha)L_w)\}\frac{\sigma_{q,L,,\mathrm{sgn}(S^{(2)})\epsilon}}{\widetilde{\sigma}_{q,L}+}$ 取 $\epsilon \to 0$，则定理2 的结论成立。

A.3 定理3的证明

根据格利文科-坎特利定理,

$$
\sup_{s\in\mathbb{R}}\left|\frac{1 + \sum_{j=1}^m \mathbb{I}\{s < S(\boldsymbol{X}_j)\}}{m+1} - (1 - F(s))\right| \to 0
$$

在概率意义下，其中 F 是 $S(\boldsymbol{X}_j)$ 的累积分布函数。注意，在零假设下， $S(\boldsymbol{X})$ 服从分布 F。因此， $F(S(\boldsymbol{X}))$ 严格服从均匀

分布 $U(0,\ 1)$。因此，对于任意 $\alpha \in (0,\ 1)$

$$\mathbb{P}_{\boldsymbol{X}\sim\mathbb{P}}\left(\text{p-value} \leq \alpha\right) = \mathbb{P}_{X\sim\mathbb{P}}\left(1 - F(S(\boldsymbol{X})) \leq \alpha\right) + o_p(1) \to \alpha \tag{16}$$

作为 $m \to \infty$。至此，定理3的证明结束。

B 数据

B.1 人类文本：来源与处理详情

Huma人类文本来源：详情。我们介绍本文使用的数据资源 r:

- 学术：PubMed 中的论文[3] 以及 arXiv 中的摘要[4]。
- 金融：金融对话[5]、关于金融问题的基于观点的问答[6]以及金融新闻[7]
- 知识：2022年3月之前清理过的维基百科文章[8]
- 政府：美国政府报告数据集[9] 及其由人类专家撰写的对应摘要[10]
- 立法：美国法院案件中法律判决的裁判要旨[11] 以及以英语撰写的欧盟法律[12]
- 医学：患者描述[13] 以及关于答案的专家解释[3]

[3] https://github.com/armancohan/long-summarization?tab=readme-ov-file
[4] https://www.kaggle.com/datasets/Cornell-University/arxiv
[5] https://huggingface.co/datasets/ceadar-ie/FinTalk-19k
[6] https://huggingface.co/datasets/LLukas22/fiqa
[7] https://huggingface.co/datasets/danidanou/Bloomberg_Financial_新闻
[8] https://huggingface.co/datasets/legacy-datasets/wikipedia
[9] https://huggingface.co/datasets/launch/gov_report
[10] https://huggingface.co/datasets/ccdv/govreport-summarization
[11] https://huggingface.co/datasets/coastalcph/lex_glue
[12] https://huggingface.co/datasets/coastalcph/multi_eurlex
[13] https://huggingface.co/datasets/ncbi/Open-Patients

经过基准测试的医学问题[14]。

- 新闻：来自 BBC[15][16] 和 CNN[17]的新闻报道。

- 用户评论：亚马逊[18]上的美食评论、Yelp[19]上的产品评论，以及 IMDb[20]上的电影评论。

文本预处理详情。为确保所收集文本的高质量，我们剔除了长度不超过 20 个词的文本，因为理论和实证研究（Bao 等人，2024，Zhou 等人，2025）均表明，过短的文本所含信息有限，难以判断其是否由大语言模型生成。我们还避免使用过长的文本，因为经典Transformer架构的运行时间呈二次方增长，会在微调过程中带来沉重的计算负担。对于极长的文本（例如维基百科中的某些条目），我们改为从原始内容中随机选取连续的 7–8 个句子。此外，我们剔除了包含大量重复内容的文本。具体而言，我们计算了每篇文本的三元语法统计，其取值范围为 0 到 1，数值越大表示重复越多。根据Huang 等人（2025）中的建议，我们将三元语法统计的阈值设为 0.4。

B.2 机器数据生成

各类别大语言模型生成文本的生成流程如下。首先，我们从每种文本类别中随机抽取 225 篇文本。然后，对于每篇选中的文本，我们从 `grok-3-mini`、`gemini-2.5-flash` 和 `gpt-4o` 中随机选择一个大语言模型，以

[14]https://huggingface.co/datasets/openlifescienceai/medmcqa[15]
https://huggingface.co/datasets/SetFit/bbc-news[16]
https://huggingface.co/datasets/EdinburghNLP/xsum[17]
https://huggingface.co/datasets/AyoubChLin/CNN_新闻_文章_2011-2022年[18]
https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews[19]
https://huggingface.co/datasets/Yelp/yelp_评论_full/[20]
https://huggingface.co/datasets/stanfordnlp/imdb

生成相应的文本。此外，我们从 20 多个候选提示词组成的集合中随机选取一个提示词，这些提示词旨在捕捉不同模型的多样化语言风格。典型示例如下：

用于生成大语言模型文本的提示词

以有说服力的风格扩写段落，但不要延长其篇幅。

我们还添加了必要的系统提示词，以确保大语言模型不会生成不必要的文本。改写、润色和扩写任务的具体系统提示词如下。

改写任务的系统提示词

你是一名专业的改写专家，可以在不遗漏原始细节的情况下，帮助将该段落用英语进行改写。请保持改写后文本的长度与原文相近。仅返回改写后的版本。不要解释修改内容，不要提供多个选项，也不要添加评论。

润色任务的系统提示词

你是一位专业的润色专家，可以帮助润色这段文字。只返回润色后的版本。不要解释修改内容，不要提供多个选项，也不要添加评论。

润色任务的系统提示词

你是一位专业的写作专家，可以帮助扩写这段文字。只返回扩写后的版本。不要解释，不要提供多个选项，也不要添加评论。

调用这些大语言模型的应用程序接口时，为简便起见，我们没有设置温度，也没有进行 top-k 或核采样。

C 实验：详情

C.1 实现细节

我们方法的微调设置。在我们的实现中，我们使用 w 函数，并通过 `google/gemma-3-1b-pt`[21]对模型进行初始化。随后，使用低秩适应（Hu 等，2022）对模型进行微调，该方法通过 `peft` 库实现，其中秩参数设为 4，`lora_alpha` 设为 16，`lora_dropout` 设为 0.05。其他所有参数均保持默认设置。

估计 $\mathcal{H}_0$ 下的统计量分布。在外部数据集上进行评估时，我们使用未参与微调的人工撰写文本来估计 $\mathcal{H}_0$ 下的检验统计量分布。相应的直方图见图 10。从图中可以看出，在大多数领域中，经验零分布都偏离正态性。这表明，推导 $\mathcal{H}_0$ 下的闭式渐近分布似乎很困难，因此为使用经验分布估计 p值提供了实证依据。

基线方法的实现。为确保公平，所有无需训练的基线方法均使用与我们的方法相同的采样模型 `google/gemma-3-1b-pt`。对于需要评分模型的方法，我们使用 `google/gemma-3-1b-it`[22]，即`google/gemma-3-1b-pt`的指令微调版本。对于基于训练的方法 ImBD，我们使用与我们的方法相同的采样模型，并采用其论文中提供的默认低秩适应设置。对于 RoBERTa 检测器和 RADAR，我们直接使用 Hugging Face 提供的检查点，即`openai-community/roberta-large-openai-detector`[23] 和 `TrustSafeAI/RADAR-Vicuna-7B`[24]。

[21]https://huggingface.co/google/gemma-3-1b-pt[22]
https://huggingface.co/google/gemma-3-1b-it[23]
https://huggingface.co/openai-community/roberta-large-openai-detector[24]
https://huggingface.co/TrustSafeAI/RADAR-Vicuna-7B

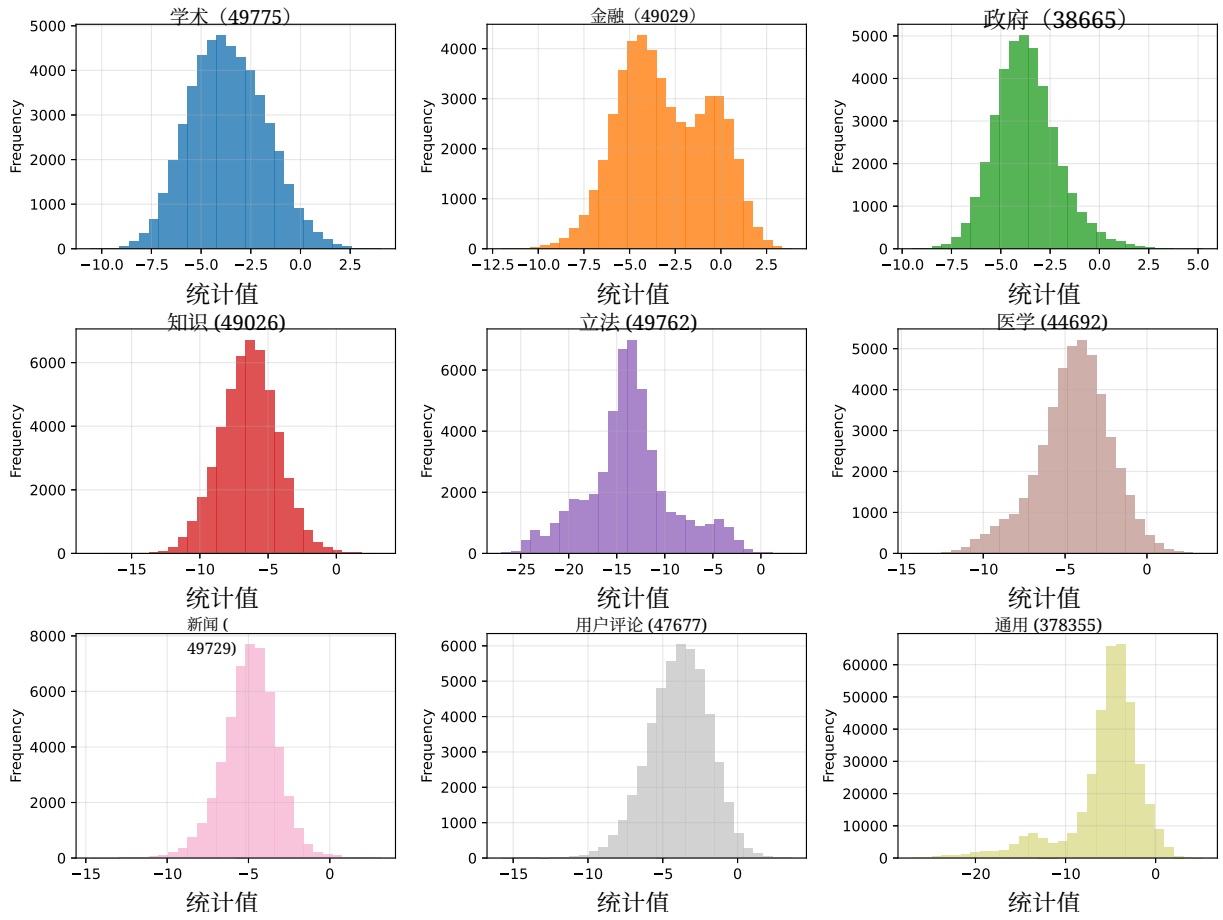

图 10：人工撰写文本的检验统计量的经验分布。

计算设备。微调流程和实验在 Ubuntu 22.04 平台上进行，该平台配备 16 个虚拟 CPU、AMD EPYC 9K84 96核处理器和 150 GB 内存。系统连接了一块配备 96 GB 显存的 NVIDIA H20-NVLink GPU，所有实验均使用 Python 3.10 运行。

C.2 第 6 节的现实应用6

向 GPT5 提示

请改写一篇论文的摘要：“变点检测在许多应用领域日益受到关注。一方面，大多数具有理论依据的方法都是在不存在模型违背的理想环境中进行研究，或者仅对随时间保持相同的重尾噪声分布和／或孤立异常值具有鲁棒性；另一方面，我们注意到，对手发起的攻击呈指数级增长，他们可能会对数据进行系统性污染，蓄意制造虚假的变点或掩盖真实变点。鉴于当前迫切需要一种能够抵御对手的变点检测方法，我们从最简单的单变量均值变点检测问题着手。我们通过 Huber ϵ 污染框架对对抗性攻击进行建模，该框架尤其允许各时间点的污染分布各不相同。本文揭示了变点检测中的一种相变现象。该检测边界是污染比例 ϵ 的函数，也是文献中首次展示这一现象。此外，我们推导出极小极大速率最优定位误差率，量化了污染比例所带来的准确率代价。我们提出一种计算上可行的方法，在特定条件下，该方法除对数因子外与极小极大下界相匹配。我们还进行了大量数值实验，并将结果与现有文献中的鲁棒变点检测方法进行比较。”

GPT5 生成的文本

以下是润色后、更自然的摘要改写版本——在保留所有技术内容的同时，改善了行文流畅度、清晰度和学术语气：摘要（改写）：变点检测在众多应用领域日益受到关注。尽管许多现有方法具有坚实的理论保证，但它们通常基于理想化假设开发，例如模型正确性或齐次噪声；其鲁棒性至多体现在能够应对孤立异常值，或重尾但分布相同的噪声。相比之下，现代数据环境日益面临对抗性污染，即有人蓄意引入系统性扰动，以掩盖真实变点或伪造虚假变点。

受这一挑战的启发，我们研究了对抗性污染下单变量均值变点检测这一基本问题，并通过 Huberϵ-污染框架对其进行建模；该框架允许污染分布随时间任意变化。我们揭示了变点可检测性中的相变现象，其特征在于一个取决于污染比例 ϵ 的精确检测边界——这一结果在此首次建立。此外，我们推导出了极小极大速率最优定位误差，量化了污染如何降低估计精度。在这些理论洞见的基础上，我们提出了一种计算高效的算法，在温和条件下达到了极小极大下界（差一个对数因子）。大量数值实验表明，与现有方法相比，我们的方法具有更强的鲁棒性和更高的准确率。

您希望我将其稍微缩短并增强表达力度（适合 NeurIPS/ICML 风格的摘要），还是保持正式完整（适合 Annals of Statistics 或 JASA 等统计学期刊）？