arXiv:2602.01240v1 [cs.CL] 1 Feb 2026

最小化不匹配风险：用于零样本大语言模型生成文本检测的基于原型的路由框架

Ke Sun [1] **Guangsheng Bao** [1] **Han Cui** [1] **Yue Zhang*** [1]

摘要

零样本方法通过代理模型计算统计特征来检测大语言模型生成文本。现有方法通常对所有输入采用固定代理模型，而不考虑未知的源模型。我们系统地考察了这一设计，发现检测性能会因代理模型与源模型的匹配程度而显著变化。我们观察到，虽然没有任何单一代理模型能始终实现最佳性能，但对于任意给定的输入，多样化的模型池中通常都存在一个匹配良好的代理模型。这一发现将稳健检测转化为一个路由问题：为每个输入选择最合适的代理模型。我们提出 DetectRouter，这是一种基于原型的框架，通过两阶段训练学习文本检测器亲和度。第一阶段从白盒模型中构建具有判别性的原型；第二阶段则通过将几何距离与观测到的检测分数对齐，泛化到黑盒源模型。在 EvoBench 和 MAGE 基准上进行的实验表明，该方法在多种检测标准和模型家族上均能持续提升性能。

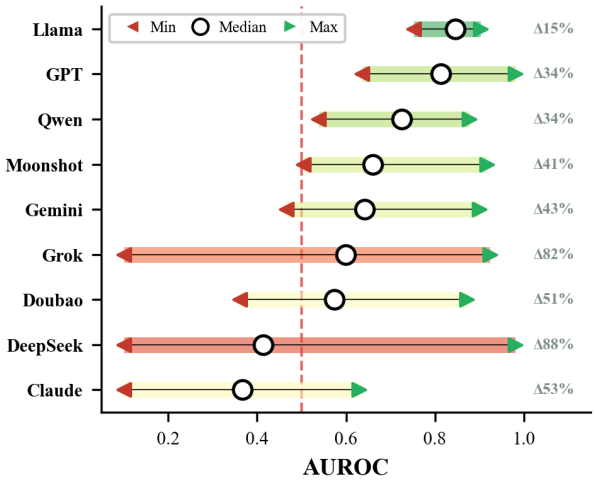

图 *1*。 MIRAGE 基准上九个黑盒大语言模型家族中 Fast-DetectGPT 的检测方差。每个柱状条表示不同代理模型所取得的 AUROC 最小值至最大值范围，圆圈表示中位数。右侧的百分比表示每个源模型中最佳代理模型与最差代理模型之间的性能差距。

1. 引言

大语言模型的迅速普及彻底改变了自然语言生成，使创意写作、代码合成等应用成为可能。然而，这种能力也对信息完整性构成了重大风险，包括大规模制造虚假信息、学术不端，以及污染用于训练未来模型的网络语料库（Milano 等人，2023年；Demszky 等人，2023年；Doshi 和 Hauser，2024）。因此，区分人类撰写文本与机器生成文本已成为一项迫切的安全挑战（Yang 等人，2024；Wu 等人，2025年）。

在现有的检测方法中，零样本方法因无需训练而日益受到关注。这些方法利用替代大语言模型计算候选文本的词元级统计量，例如条件概率或似然曲率，再将这些统计量汇总为检测分数，以区分机器生成内容与人类撰写文本（Mitchell 等人，2023年；Bao 等人，2024；Sun 等人，2026）。这类方法在可访问源模型的白盒设置下表现出色，但在生成模型未知的现实黑盒场景中，其效能会显著下降（Fu 等人，2025a；Bao 等人，2025年）。这一差距带来了亟待系统研究的紧迫挑战。

造成这一局限的一个关键因素是，所有现有的零样本方法都使用一个固定代理模型，来检测来自不同文本来源的文本；其依据是分布匹配假设，即不同大语言模型之间的概率分布相似（Mitchell 等人，2023年；Bao 等人，

[1]西湖大学工程学院，中国。通信作者：Yue Zhang <zhangyue@westlake.edu.cn>。

2024）。我们的初步实验对这一假设提出了质疑。如图1所示，在 MIRAGE 基准（Fu 等人，2025a）上使用九个开源代理模型评估 Fast-DetectGPT，结果显示其性能差异巨大：对于 DeepSeek 和 Grok 等具有挑战性的源模型，最佳与最差代理模型之间的差距超过 80%，检测表现从近乎随机到效果显著不等。我们在第 3 节进一步量化了这一现象，并推导出失配风险界，该界表明，性能差距的上界为源模型分布与代理分布之间 KL散度的平方根。这一界将检测问题转化为路由问题：给定一段输入文本，我们必须找出与未知源模型的分布差异最小的代理模型。

基于这一洞见，我们提出 DetectRouter，一种面向零样本大语言模型检测的基于原型的路由框架。给定一篇来源未知的文本，DetectRouter 首先将其编码到学习得到的度量空间中，然后在一组检测器特定表示中识别最近原型，最后将该文本路由到相应的最优检测器进行评分。由于黑盒模型的真实来源分布不可获取，我们从几何上近似分布亲和度：我们学习一个嵌入空间，其中到检测器原型的距离反映潜在的分布邻近性。DetectRouter 采用两阶段训练策略。阶段 1 通过对比学习，利用具有已知源标签的白盒模型构建判别性原型。第二阶段通过将来自替代模型池的检测分数作为监督信号，并最小化预测亲和度与观测分数分布之间的 KL散度，从而泛化到黑盒来源。在推理时，路由器通过最近原型匹配将每个输入路由到最兼容的检测器。

在 EvoBench 和 MAGE 基准上的大量实验表明，DetectRouter 持续提升全部六项检测标准，相对于固定代理基线的相对增益范围为 5.4% 至 139.4%，平均为 36.1%。DetectRouter 取得了最先进的性能，在 EvoBench 上的平均曲线下面积为 90.85%，在 MAGE 上为 77.92%，分别比最佳基线高出 9.84% 和 4 个百分点。

我们的贡献总结如下：

- 我们首次对零样本大语言模型检测中的代理选择进行了系统分析，发现代理模型与源模型之间的亲和度是决定黑盒性能的关键因素。
- 我们提出 DetectRouter，这是一种基于原型的路由框架，通过结合判别式原型构建与分布对齐的两阶段训练，学习将输入文本匹配到最优检测器。
- 在两个基准测试上进行的大量实验表明，DetectRouter 达到了最先进的性能，并可作为适用于所有零样本检测方法的通用增强层。

2. 相关工作

文本生成检测。现有方法分为两类。基于训练的方法通过集成、对比或对抗目标微调分类器（Ab- buri 等，2023；Hu 等，2023；Guo 等，2024；Chen 等，2025；Su 等，2023b），但在未见过的模型或提示上往往性能下降（Yang 等，2024；Wu 等，2025；Sun 等，2025）。基于统计的方法则将替代大语言模型视为评分函数，并根据词元概率设计检测规则，从而无需特定任务数据即可实现零样本检测器。GLTR 使用对数似然、熵和排名直方图等词元级统计量（Gehrmann 等，2019）；DetectLLM 利用对数排名分数（Su 等，2023a）；DetectGPT 衡量扰动下的对数似然曲率（Mitchell 等人，2023）；Fast-DetectGPT 则以高效的解析近似取代采样（Bao 等人，2024）。近期的扩展包括基于核的检验（Yu 等人，2024；Song 等，2025）和基于记忆的检测器（Guo 等，2024）。尽管取得了这些进展，大多数方法都假设单个代理模型已足够；而我们的工作会将每个输入路由至池中最合适的检测器。

大语言模型的路由与来源追踪。模型路由通过学习将输入分配给模型池中准确率与成本权衡各异的合适模型，从而在预算限制下实现高效推理（Hu 等人，2024；Huang 等人，2025年；Jitkrittum 等人，2025年；Chuang 等人，2024）。然而，现有方法主要将大语言模型视为问答或推理任务的求解器，而非生成文本的检测器。Li 等人（2023年）和 Fu 等人（2025b）等来源追踪方法构建分类器，在黑盒场景下识别给定文本由哪个大语言模型生成，但其目标是进行来源归因，而非利用检测器互补性来提升检测效果。Sun 等人（2025年）表明，不同大语言模型呈现出高度可区分的特异性模式，从而能够高准确率地恢复源模型。

3. 分析：错配成本

在介绍我们的方法之前，我们先考察支撑当前零样本检测方法的分布匹配假设是否成立。标准方法使用固定代理模型来估计统计伪迹，并假设其概率分布景观能够充分近似多种来源分布（Mitchell 等人，2023年；Bao等人，2024）。这种假设看似简洁，却忽略了

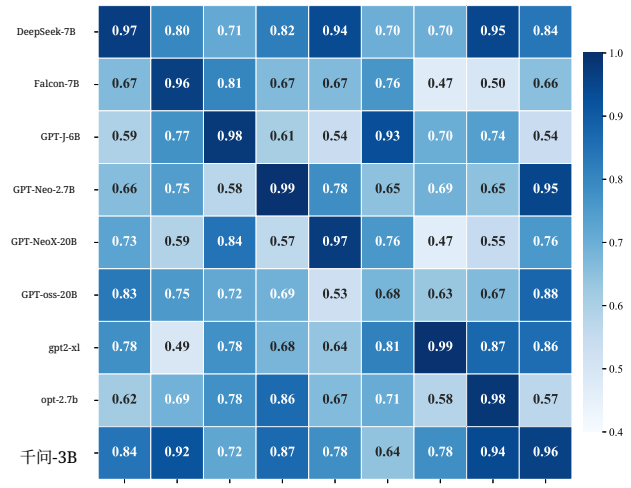

图 *2.* 使用 Fast-DetectGPT 的检测性能亲和矩阵。九个开源模型同时作为生成器（行）和代理模型（列）。对角占优证实，最优性能需要源模型与代理模型对齐，而非对角线模式则揭示了由架构相似性支配的结构化亲和度。

模型模型特异性——不同架构和预训练语料库带来的独特统计指纹（Sun 等人，2025年；Yang 等人，2024）。

3.1. 实验设置

我们在白盒设置和黑盒设置下设计了交叉评估实验。对于白盒设置，我们遵循 Fast-DetectGPT（Bao 等人，2024）的协议，并使用 XSum（Narayan 等人，2018）作为种子语料库。九个参数规模从 2.7B 到 20B 的开源大语言模型各自通过核采样生成 500 条文本续写。随后，我们将每个模型作为代理检测器，在全部九个语料库上进行评估，从而得到一个 9×9 交叉评估矩阵。对于黑盒设置，我们使用 MIRAGE 基准（Fu 等人，2025年a），其中包含 GPT-4、Claude、Gemini 和 DeepSeek 等专有模型。完整的实验细节见附录E。

3.2. 实证研究

白盒设置。图2 展示了 Fast-DetectGPT 的两两AUC分数。热力图揭示了两种关键模式。首先，我们观察到明显的对角占优，这证实了只有当源模型与代理模型相同时，检测器才能达到最佳性能。其次，非对角线性能下降具有结构性：共享架构谱系的代理模型（例如 GPT-Neo 和 GPT-J 系列）保持较高的相互亲和性，而结构迥异的模型对则会出现显著的性能崩溃。值得注意的是，即使在规模相当的模型之间，仅架构差异就可能导致曲线下面积下降超过 30 个百分点。我们在附录C.1中提供了全部五种检测标准的亲和度矩阵。

表1. 量化五种检测标准下源模型与代理模型的错位程度。匹配：源模型 = 代理模型；交叉：源模型 $\neq$ 代理模型。在黑盒设置中，Max/Min 表示对所有生成器系列取平均后表现最佳/最差的代理模型。

	白盒 / XSum		黑盒 / MIRAGE		
准则	匹配	交叉	Mean	Max	Min
似然	0.81	0.42	0.49	0.76	0.35
Rank	0.76	0.60	0.52	0.64	0.43
LogRank	0.85	0.46	0.48	0.73	0.35
熵	0.55	0.71	0.59	0.74	0.41
Fast-DetectGPT	0.94	0.71	0.65	0.93	0.47
平均值	0.78	0.58	0.55	0.76	0.40

黑盒设置。 如图 1所示，不同代理模型选择所导致的性能方差极大。对于 Claude 和 DeepSeek 等难以检测的源模型，曲线下面积的范围从接近随机水平的 0.2 到超过 0.9、接近完美检测的水平，差异完全取决于代理模型的选择。尽管性能中位数徘徊在随机基线附近，但最高性能始终保持在较高水平——这证实开源社区中存在检测能力，但这些能力分散在不同模型中。这种分散性意味着没有任何单一代理模型能够普遍占优，因此有必要根据输入特征进行自适应选择。附录 C.2中提供了更多可视化结果，包括各源模型的分布。

3.3. 定量分析

表 1 量化了五种检测标准下源模型与代理模型错位所造成的成本。在白盒设置中，平均曲线下面积从匹配条件下的 0.78 降至跨模型条件下的 0.58，相对下降了 26%。对于 Fast-DetectGPT，性能从 0.94 降至 0.71——这一差距使近乎完美的检测器沦为表现平庸的检测器。

在黑盒设置中，不同代理模型选择之间的方差同样显著。尽管平均性能仅为 0.55，表现最佳的代理模型平均得分却达到 0.76，这证实每个源模型都存在最优代理模型。对于 Fast-DetectGPT，差距尤为明显：最佳代理模型的得分为 0.93，而最差的则降至 0.47。这些发现表明，固定代理选择会带来显著损失，但在多样化的模型池中确实存在最优选择——这也说明了采用路由机制进行动态代理选择的必要性。

3.4. 理论论证

实证结果揭示了一种一致的模式：检测可靠性与代理模型和源模型之间的分布距离呈负相关。为将这一点形式化，我们建立如下符号体系。令 P_{src} 表示生成文本的源模型的概率分布，令 P_{sur} 表示用于检测的代理模型的概率分布。给定有界评分函数$T(x)$：$\mathcal{X} \to \mathbb{R}$，满足 $|T(x)| \leq B$，最优检测信号为 $\mu^* = \mathbb{E}_{P_{src}}[T(x)]$,，而使用代理模型获得的代理信号为 $\mu_{proxy} = \mathbb{E}_{P_{sur}}[T(x)]$。

命题 3.1 (失配风险界)。对于任意有界的零样本检测统计量 T，满足 $|T(x)| \leq B$:

$$|\mu^* - \mu_{proxy}| \leq B \cdot \sqrt{2D_{KL}(P_{src} \| P_{sur})} \tag{1}$$

其中 D_{KL} 表示 *Kullback-Leibler* 散度。

证明通过全变差距离和 Pinsker 不等式进行；完整推导见附录 A。该界严格解释了我们的实证发现：当架构不同时，跨模型性能下降反映了 KL 散度的增大，而最优代理模型恰好是使该散度最小化的模型。因此，稳健的零样本检测可转化为一个路由问题：给定一个代理模型池 $\{M_1, \ldots, M_K\}$及其对应的分布 $\{P_{M_1}, \ldots, P_{M_K}\}$，最优策略是针对每个输入选择使 M_k 最小化$D_{KL}(P_{src} \| P_{M_k})$ 的代理模型。这一洞见构成了 DetectRouter 的基础。

4. 方法

4.1. 问题表述

第 3 节的分析表明，最优策略是动态选择一个代理模型 M_k 以最小化$D_{KL}(P_{src} \| P_{M_k})$。如图 3 所示，这将检测重新表述为一个路由问题：给定文本 x，其源模型 M_{src}未知，需要从模型池中确定最优检测器$\mathcal{D} = \{D_1, \ldots, D_N\}$，其中每个 D_k 由一个代理模型 M_k和一个检测准则 $F(\cdot;\ M_k)$ 配对组成，例如似然或 Fast-DetectGPT:

$$k^* = \underset{k \in \{1,\ldots,N\}}{\arg\min}\ D_{KL}(P_{src} \| P_{M_k}) \tag{2}$$

由于黑盒模型无法访问 P_{src}，我们采用几何方式近似分布亲和性。我们学习一个编码器 E_θ：$\mathcal{X} \to \mathbb{R}^d$，并用可学习原型 $\mathcal{P}_k = \{\mathbf{p}_{k,\ 1},\ \ldots,\ \mathbf{p}_{k,\ K}\}$ 表示每个检测器 D_k，以捕捉模型家族 k的特征。路由决策转化为最近原型查询:

$$k^* = \arg\min_k \min_{\mathbf{p} \in \mathcal{P}_k} \|E_\theta(x) - \mathbf{p}\|_2 \tag{3}$$

我们采用两阶段训练：阶段 1 从白盒模型中构建判别式原型，第二阶段则通过分布对齐来对齐黑盒来源。

4.2. 阶段 1：判别性原型构建

第一阶段使用模型池 $\mathcal{M}$ 中具有真实源标签的 N 白盒模型，构建结构良好的度量空间。一个关键的设计考量是确保原型捕捉内在的风格特征，而非特定任务伪影。我们通过多任务生成构建多样化的锚点数据集 $\mathcal{T}_1 = \{(x_j,\ y_j)\}_{j=1}^{|\mathcal{T}_1|}$：让每个模型 M_i 执行以下任务：(1). 根据种子提示直接生成文本；(2). 润色人工草稿；(3). 改写现有段落。这种多样性促使编码器识别在不同条件下依然存在的模型特定模式，从而生成不局限于单一任务、具有良好泛化能力的原型。

基于距离的分类。 我们通过输入 x（嵌入为 $\mathbf{z} = E_\theta(x)$）到其最近原型的距离，量化该输入与检测器类别 i之间的亲和度:

$$d_i(\mathbf{z}) = \min_{k \in \{1,\ldots,K\}} \|\mathbf{z} - \mathbf{p}_{i,k}\|_2 \tag{4}$$

这些几何距离通过温度缩放的 softmax 转换为分类概率，其中距离越小表示亲和度越高:

$$P(y = i \mid \mathbf{z}) = \frac{\exp(-d_i(\mathbf{z})/\tau)}{\sum_{j=1}^{N} \exp(-d_j(\mathbf{z})/\tau)} \tag{5}$$

训练目标。 编码器与原型通过复合目标函数联合优化。我们最小化交叉熵损失 $\mathcal{L}_{\text{CE}}$，以使嵌入聚集在其真实标签原型周围:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{T}_1|} \sum_{(x,y) \in \mathcal{T}_1} \log P(y \mid E_\theta(x)) \tag{6}$$

为进一步构建度量空间，我们引入基于间隔的分离损失 $\mathcal{L}_{\text{sep}}$，通过对来自不同类别且距离小于间隔 m的原型对施加惩罚，来强制实现类间分离:

$$\mathcal{L}_{\text{sep}} = \frac{1}{Z} \sum_{i \neq j} \sum_{k,l} \max\left(0, m - \|\mathbf{p}_{i,k} - \mathbf{p}_{j,l}\|_2\right)^2 \tag{7}$$

归一化项 $\mathcal{L}_{\text{norm}} = \sum_{i,k} \|\mathbf{p}_{i,k}\|_2^2$ 约束原型的幅度以保障训练稳定性。完整的阶段 1 目标为:

$$\mathcal{L}_1 = \mathcal{L}_{\text{CE}} + \lambda_{\text{sep}} \mathcal{L}_{\text{sep}} + \lambda_{\text{norm}} \mathcal{L}_{\text{norm}} \tag{8}$$

收敛后，这将形成一个判别性嵌入空间，其特点是类内紧凑性高、类间分离清晰。

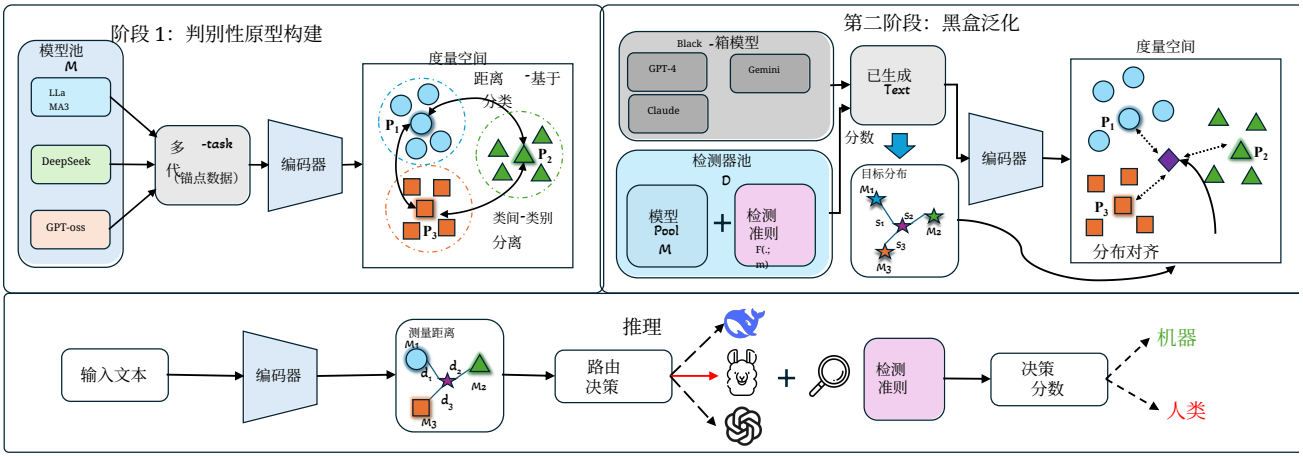

图 *3.* DetectRouter 概述。阶段 1 通过多任务生成与基于距离的分类，从白盒模型构建判别式原型。第二阶段通过将几何距离与检测分数分布进行对齐，泛化至黑盒来源。在推理时，路由器依据最近原型亲和度选择最优检测器。

4.3. 第二阶段：通过分布对齐实现黑盒泛化

阶段 1 构建了坐标系，但依赖显式标签。现实场景中存在 GPT-4 等未知的黑盒模型，其标签无法获取。为弥合这一差距，我们利用检测器池 $\mathcal{D}$ 生成的检测分数向量，作为分布对齐的经验性代理。

目标分布构建。 对于黑盒文本x，我们从检测器池中观测到分数向量 $\mathbf{s} = [s_1, \ldots, s_N]$。这些分数反映了该文本与每个已知模型家族之间的统计接近程度。由于尺度差异会使绝对分数的回归不稳定，我们不直接回归绝对分数，而是通过温度缩放 Softmax 将其归一化为目标分布 Q：

$$Q_i = \frac{\exp(s_i/T)}{\sum_{j=1}^{N} \exp(s_j/T)} \tag{9}$$

同时，路由器输出一个预测亲和度分布 P，该分布由原型空间中的几何距离导出，如式5。

KL 散度优化。 我们将学习目标表述为分布匹配问题，最小化目标分布 Q 与预测几何分布 P 之间的 Kullback-Leibler 散度：

$$\mathcal{L}_{\text{KL}} = D_{KL}(Q\|P) = \sum_{i=1}^{N} Q_i \log \frac{Q_i}{P_i} \tag{10}$$

这一目标明确使路由器的几何逻辑与最小化不匹配风险这一理论目标保持一致：路由器学会对黑盒嵌入进行定位，使其与原型之间的距离能够反映实际的检测亲和度。

原型锚定。 无监督适应存在灾难性遗忘的风险，即原型可能发生漂移并破坏阶段 1 所学到的判别性结构。为防止这种情况，我们采用原型锚定：冻结一份已收敛的阶段 1 原型副本 $\{\mathbf{p}^{\text{anchor}}\}$，并对偏差施加惩罚：

$$\mathcal{L}_{\text{anchor}} = \sum_{i,k} \|\mathbf{p}_{i,k} - \mathbf{p}_{i,k}^{\text{anchor}}\|_2^2 \tag{11}$$

完整的第二阶段目标将分布对齐与正则化相结合：

$$\mathcal{L}_2 = \mathcal{L}_{\text{KL}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} \tag{12}$$

这确保路由器能够泛化至未见来源，同时保留已知模型的稳健拓扑结构。

4.4. 推理

在推理过程中，给定输入文本 x，编码器会将其投影到度量空间中，以计算其与所有检测器原型之间的距离$\{d_1, \ldots, d_N\}$。路由决策通过求解 $k^* = \arg\min_k d_k$，确定亲和度最高的检测器。随后，输入会被路由至选定的检测器 D_{k^*}，由该检测器计算最终的检测分数，以将文本分类为人类撰写或机器生成。

5. 实验

5.1. 实验设置

训练数据。我们在 MIRAGE 基准上训练 DetectRouter（Fu 等人，2025a），这是一个全面的多任务

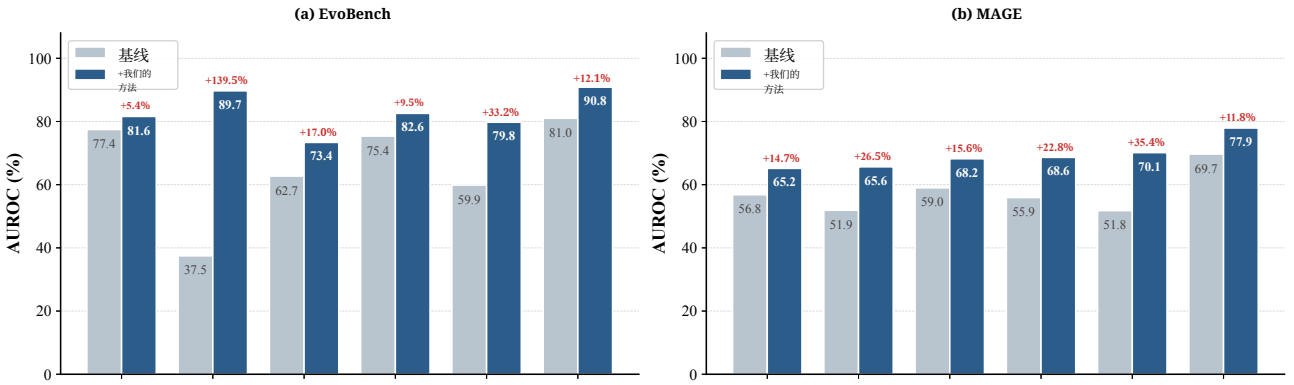

图4. 将 DetectRouter 应用于各零样本检测方法所带来的性能提升。柱形表示在 EvoBench 和 MAGE 上的平均曲线下面积；数字表示相对提升幅度。

数据集，涵盖五个领域——学术、新闻、评论、电子邮件和网站——以及三种生成任务：生成、润色和改写。对于阶段 1，我们通过提示本地缓存的代理模型生成符合 MIRAGE 架构的文本，合成了 3 万个训练样本。对于第二阶段，我们直接利用 MIRAGE，并使用标注了检测器池中全部检测器检测分数的数据，为学习文本与检测器亲和模式提供监督信号。

代理模型与检测方法。我们的代理模型池包含十个开源大语言模型，参数量从 1.5B 到 20B 不等：GPT-J-6B（Wang 和 Komatsuzaki，2021年）、GPT-NeoX-20B（Biderman 等人，2023年）、GPT-Neo-2.7B（Black 等人，2021年）、GPT2-XL（Radford 等人，2019年）、GPT-oss-20B（OpenAI，2025年）、LLaMA3-8B（Dubey 等人，2024年）、Falcon-7B（Almazrouei 等人，2023年）、DeepSeek-7B（DeepSeek-AI，2023年）、OPT-2.7B（Zhang 等人，2022）和 Qwen-3B（Bai 等人，2023年）。这一选择确保了架构和训练语料库的多样性。我们评估六种零样本检测准则——似然、熵、排名、LogRank、对数似然比和 Fast-DetectGPT（Bao 等人，2024）——并将每种固定代理基线与我们的路由变体进行比较。我们首先在全部十个代理模型上对所有检测方法进行基准测试，并确定 LLaMA3-8B 始终表现最佳；因此，我们将所有基线方法固定为使用 LLaMA3-8B，以确保比较具有竞争力，而非人为削弱基线。

路由器配置。主实验中，路由器采用 LLM2Vec-LLaMA3-8B（BehnamGhader et al.，2024）作为编码器，并在消融研究中探索轻量级替代方案。每个类别由 10 个可学习原型表示，采用余弦相似度和温度 τ=10。两个训练阶段均运行 8 轮，批次大小为 32；阶段 1 的学习率为 2×10^{-5}，第二阶段则将其降至 10^{-5} 并平衡交叉熵、距离回归和原型锚定损失。

基线。我们将多种具有代表性的零样本检测方法作为对比，包括似然（Solaiman 等人，2019年）、熵和排名（Gehrmann 等人，2019年），以及 LogRank（Mitchell 等人，2023年）；基于对数似然比（LLR）的相对评分（Su 等人，2023a）；以及基于扰动的 Fast-DetectGPT（Bao 等人，2024）。我们还纳入了近期方法：Lastde（Xu 等人，2024b）、FourierGPT（Xu 等人，2024a）、Diveye（Basani 和 Chen，2025年）以及 UCE（Hou 等人，2025年）。为将路由框架的贡献与训练数据的影响区分开来，我们还使用相同的第二阶段数据训练了监督分类器。所有基线均采用相同的代理模型，以确保公平比较。

评估基准。我们在两个互补的真实场景基准上进行评估。MAGE（Li 等人，2024）覆盖范围广泛，包含来自八个模型家族的 30,265 个测试对——OpenAI、OPT、FLAN、LLaMA、BigScience、EleutherAI、GLM 和人工释义——涵盖了成熟的生成器和多样化领域。EvoBench（Yu 等人，2025年）专门针对大语言模型不断演化的特性，包含来自七个模型家族和 29 个演化版本的 32,993 个测试对，涵盖发布者更新以及微调和剪枝等开发者修改。遵循先前研究（Mitchell 等人，2023年；Bao 等人，2024），我们将曲线下面积（AUROC）作为主要指标。关于数据构建、超参数和基准统计信息的完整细节见附录 D。

5.2. 主要结果

DetectRouter 的一项核心贡献是，它能够作为通用增强层，服务于任何零样本检测

表2。 检测表现 EvoBench 和 MAGE 上的表现。固定代理基线使用 LLaMA3-8B。曲线下面积以百分比表示 e.

	EvoBench								MAGE
方法	LLaMA3	LLaMA2	Claude	GPT-4o	Gemini	GPT-4	Qwen	Avg	Avg
似然	88.84	61.68	79.70	77.80	78.89	77.80	77.40	77.44	56.81
熵	28.85	57.58	33.61	35.43	38.34	38.10	30.34	37.46	51.90
Rank	68.49	58.38	64.50	64.64	64.46	63.43	55.06	62.71	58.99
LogRank	86.47	59.99	77.37	76.80	75.99	74.87	76.10	75.37	55.91
LLR	66.46	49.61	61.73	60.57	59.12	58.43	63.08	59.86	51.78
Fast-DetectGPT	94.17	80.27	81.57	74.89	80.48	75.58	80.10	81.01	69.69
Lastde	90.80	80.54	76.75	71.34	76.50	71.51	69.20	76.66	70.71
FourierGPT	63.99	57.23	58.20	55.52	56.69	55.16	64.39	58.74	60.34
Diveye	84.65	73.46	74.21	73.32	74.06	72.28	68.85	74.40	69.60
UCE	75.62	61.35	68.72	62.04	61.28	59.38	63.63	64.57	54.17
监督式	74.96	66.88	73.38	77.66	62.07	76.16	66.49	71.26	73.92
DetectRouter	**95.66**	**80.82**	**94.22**	**94.29**	**93.79**	**94.16**	**83.01**	**90.85**	**77.92**

表3. 对训练阶段和损失分量的消融。平均曲线下面积以百分比报告。 Δ 表示相对于完整方法的相对下降。

设置	EvoBench	MAGE	Δ
固定代理	81.01	69.69	-10.7%
直接分类	83.08	75.34	-5.9%
仅阶段1	72.87	75.67	-11.3%
仅阶段2	87.39	74.49	-4.1%
完整（不含 $\mathcal{L}_{KL}$）	87.07	75.56	-3.6%
完整（不含 $\mathcal{L}_{sep}$）	88.53	74.41	-3.5%
完整（不含 $\mathcal{L}_{anchor}$）	89.56	74.90	-2.6%
Full	90.85	77.92	–

准则。在介绍最先进方法比较之前，我们先展示通用增强特性。

路由作为通用增强。图4 展示了将我们的路由机制应用于六种检测方法后取得的性能提升。在 EvoBench 上，路由带来的相对提升幅度从似然的 5.4% 到熵的 139.4% 不等。在 MAGE 上，提升幅度从 Fast-DetectGPT 的 11.8% 到对数似然比的 35.4% 不等。基于熵的检测受益最为显著：固定代理基线在 EvoBench 上仅达到 37.46%，在若干模型家族上甚至低于随机猜测；而路由将性能提升至 89.70%。这表明，我们的框架并不局限于任何特定的检测准则，而是可与现有方法无缝集成的通用增强层。所有方法组合按模型家族划分的详细结果见附录B.2。

与基线方法比较。 表2 展示了 EvoBench 上按模型家族划分的曲线下面积（AUROC）分数，以及 MAGE

表4。路由器编码器架构消融实验。报告平均 AUROC 百分比，以及模型参数量和单样本推理延迟。

编码器	EvoBench	MAGE	参数	延迟
固定代理	82.53	69.68	–	–
BERT-base	86.41	71.07	109M	4.87毫秒
RoBERTa-base	88.34	74.10	125M	4.87毫秒
RoBERTa-large	88.18	75.59	355M	12.70毫秒
LLM2Vec-LLaMA3	90.85	77.92	8B	166.2 毫秒

平均值。DetectRouter 在 EvoBench 和 MAGE 上分别取得 90.85% 和 77.92% 的最佳总体表现，显著优于所有基线。在固定代理方法中，Fast-DetectGPT 表现最强，在 EvoBench 和 MAGE 上分别达到 81.01% 和 69.69%；我们的路由机制较该基线分别提升了 9.84 和 8.23 个百分点。在分布差距最大的专有模型上，提升尤为显著：Claude 上的表现从 81.57% 提升至 94.22%；GPT-4o 上从 74.89% 提升至 94.29%；GPT-4 上从 75.58% 提升至 94.16%。在近期的零样本方法中，Lastde 在 EvoBench 上的平均值最高，为 76.66%，但 DetectRouter 仍高出 14.19 个百分点。FourierGPT、Diveye 和 UCE 的表现均明显逊于我们的方法，这表明自适应路由能为检测准则设计带来正交收益。MAGE 上按模型家族划分的完整结果见附录 B.1。

为了将我们路由框架的贡献与训练数据分离开来，我们在相同的第二阶段数据上训练监督分类器。使用 LLama3-8b 表征的最佳有监督基线在 EvoBench 上达到 71.26%，在 MAGE 上达到 73.92%。DetectRouter 在 EvoBench 上比该基线高出 12.93 个百分点，同时匹配 perfor-

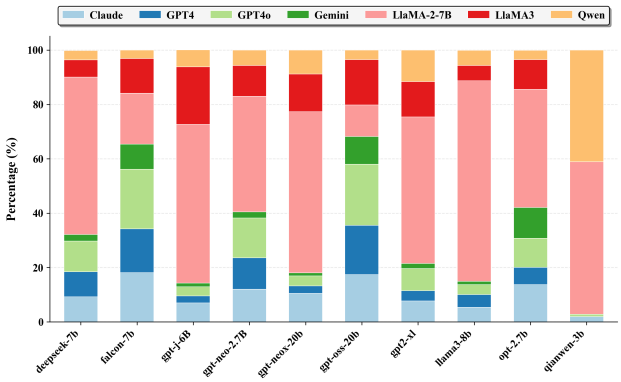

图 *5.* 各源模型家族所选代理模型的分布，证实了源特定路由偏好。

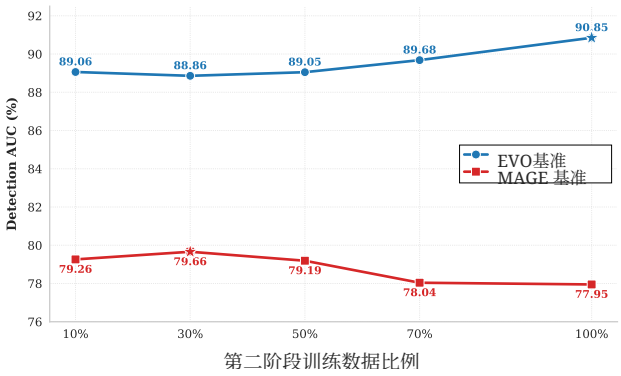

图*6*。以第二阶段所用训练数据的比例为横轴，展示 EvoBench 和 MAGE 上的检测曲线下面积分数。

MAGE 上的表现，这表明提升源于学得的文本与检测器亲和模式，而非仅仅使用更多训练数据。全部十种代理配置的完整结果见附录B.3。

5.3. 消融与分析

训练阶段与损失分量。 表 3 展示了消融实验结果。两个训练阶段都至关重要：仅阶段1的性能下降 -11.3%，因为仅在白盒模型上训练的原型无法泛化；而仅阶段2则因缺少经过良好初始化的原型而出现 -4.1% 的差距。每个损失分量都有显著贡献——将 KL散度替换为 $\mathcal{L}_1$ 回归会导致 -3.6% 的性能下降，而移除 $\mathcal{L}_{sep}$ 会导致 -3.5% 的性能下降。直接分类绕过了原型框架，并出现 -5.9% 的差距，证实了度量空间路由的优势。

编码器架构。表4 比较了路由网络采用的不同编码器架构。在轻量编码器中，RoBERTa-base 实现了最佳平衡：在 EvoBench 和 MAGE 上的表现分别比 BERT-base 高出 1.93% 和 3.03%，同时每个样本的推理延迟保持不变，为 4.87 毫秒。LLM2Vec-LLaMA3 的总体表现最佳，在 EvoBench 和 MAGE 上分别达到 90.85% 和 77.92%，但每个样本需要 166.2 毫秒。这种权衡表明，对于延迟敏感应用，RoBERTa-base 以极低的计算开销实现了出色性能，是一个颇具吸引力的选择。

路由模式解读。 为验证我们的性能提升源于自适应对齐，而非简单的集成，我们在图 5 中展示了不同来源类别所选代理模型的分布。堆叠条形图显示，路由器并不会对所有输入都简单地默认使用单一通用检测器。相反，它表现出明显的来源特定偏好。例如，在检测来自 DeepSeek-7B 的文本时，路由器会频繁选择结构相似的代理模型；而对于 Falcon-7B，其偏好则显著转向另一组检测器。这证实了我们的假设：亲和度是一种潜在的连续属性，路由器能够成功识别并利用每个源模型独特的统计特征。

数据效率。黑盒适配的一项关键实际考量是收集检测分数的成本。我们通过调整阶段二训练数据的比例（从 10% 到 100%），考察 DetectRouter 的数据效率。如图 6所示，该模型表现出显著的稳定性。在 EvoBench 上，仅使用 10% 的数据，路由器就达到了 89.06% 的曲线下面积，与使用完整数据集时的表现相差不到 1.8 个百分点。这表明，阶段 1 构建的原型空间提供了强有力的正则化先验，使 DetectRouter 能够以极高的成本效益适配新模型。关于原型空间几何的更多分析见附录 B.4。

6. 结论

本研究揭示，零样本大语言模型检测对代理模型与未知来源之间的分布对齐高度敏感。通过系统性实验和理论分析，我们证明，代理模型与来源不匹配会导致性能显著下降，并推导出一个界，表明这种差距由分布之间的 KL散度决定。基于这一发现，我们提出了 DetectRouter，这是首个利用多个代理模型作为互补专家，并将每个输入动态路由至最匹配检测器的框架。实验结果证实，这种自适应路由策略在所有检测标准上均能持续提升性能，并在两个基准测试上达到最先进性能。

影响声明

本研究旨在推动对机器生成文本的可靠检测，这对于在大语言模型时代维护信息完整性具有广泛意义。有效的检测方法有助于遏制人工智能生成的虚假信息传播、维护学术诚信，并防止训练语料库受到合成内容污染。我们的路由框架提升了在多种不断演进的大语言模型环境下的检测鲁棒性，使其更适用于源模型未知的现实部署场景。

我们承认，检测技术与生成方法处于持续的对抗动态之中。尽管我们的工作增强了防御能力，但有决心的对手仍可能开发对策。我们还注意到，不完美的检测器有对人类作者作出错误指控的风险，这在教育或职业情境中可能造成严重后果。我们鼓励从业者将检测工具作为众多信号之一，而非作为确定性判断，并针对其具体应用设定适当阈值，以平衡假阳性率与假阴性率。

参考文献

Abburi, H., Roy, K., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E.与Bhattacharya, S. 一种简单而高效的人工智能生成文本检测集成方法。载于Gehrmann, S., Wang, A., Sedoc, J., Clark, E., Dhole, K., Chandu, K. R., Santus, E.与Sedghamiz, H.（编），第三届自然语言生成、评估与度量研讨会（*GEM*）论文集，第413–421页，新加坡，2023年12月。计算语言学协会。Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A.,´Cojocaru, R., Debbah, M., Goffinet, E., Hesslow, D., Launay, J., Malartic, Q.等人。开放语言模型的Falcon系列。*arXiv*预印本*arXiv:2311.16867*，2023年。Bai, J., Yang, A., Li, B.等人。通义千问技术报告。*arXiv*预印本*arXiv:2309.16609*，2023年。网址 https://arxiv.org/abs/2309.16609。

Bao, G., Zhao, Y., Teng, Z., Yang, L., 和 Zhang, Y. Fast-DetectGPT：通过条件概率曲率实现机器生成文本的高效零样本检测。见第十二届学习表征国际会议，2024。

Bao, G., Zhao, Y., He, J., and Zhang, Y. Glimpse：让白盒方法能够使用专有模型进行零样本大语言模型生成文本检测。载于第十三届国际学习表征会议，2025年。

Basani, A. R. and Chen, P.-Y. 多样性提升人工智能生成文本检测效果。*arXiv* 预印本 *arXiv:2509.18880*，2025年。

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., 和 Reddy, S. LLM2Vec：大语言模型其实是强大的文本编码器。*arXiv*预印本 *arXiv:2404.05961*，2024。

Biderman, S., Black, S., Gao, L. 等. GPT-NeoX-20B：一种开源自回归语言模型。载于*BigScience*第*5*期研讨会论文集——创建大语言模型的挑战与展望，2023年。

Black, S., Gao, L., Wang, P., Leahy, C., 和 Biderman, S. GPT-Neo：使用Mesh-TensorFlow进行大规模自回归语言建模。https://doi.org/10.5281/zenodo.5297715，2021年。Zenodo。

Chen, J., Zhu, X., Liu, T., Chen, Y., Xinhui, C., Yuan, Y., Leong, C. T., Li, Z., Tang, L., Zhang, L., 等。先模仿后检测：对齐机器风格偏好以进行机器修订文本检测。载于*AAAI*人工智能会议论文集，第39卷，第23559–23567页，2025年。

Chuang, Y.-N., Sarma, P. K., Gopalan, P., Boccio, J., Bolouki, S., Hu, X. 和 Zhou, H. 使用置信度标记学习如何路由大语言模型。*arXiv* 预印本 *arXiv:2410.13284*，2024。

DeepSeek-AI。DeepSeek 大语言模型：答案由此而来。https://github.com/deepseek-ai/DeepSeek-LLM，2023年。DeepSeek LLM 7B/67B 基础模型和聊天模型。

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., 等。将大语言模型应用于心理学。《自然综述：心理学》，2(11):688–701，2023年。

Doshi, A. R. 和 Hauser, O. P. 生成式人工智能提升个体创造力，但会降低新颖内容的集体多样性。《科学进展》，10(28):eadn5290，2024。

Dubey, A., Jauhri, A., Pandey, A., 等。Llama 3 模型系列，2024年。网址：https://arxiv.org/abs/2407.21783。

Fu, J.、Guo, C.-L.与Li, C. DetectAnyLLM：面向跨领域与跨模型的机器生成文本的可泛化且稳健检测。载于第*33*届*ACM*国际多媒体会议论文集，第11229–11238页，2025a。

Fu, Z.、Chen, J.、Sun, H.、Yang, T.、Li, R.与Zhang, Y. Fdllm：一种面向多语言、多领域黑盒环境中大语言模型的文本指纹检测方法。*arXiv*预印本，第arXiv–2501页，2025b。

Gehrmann, S., Strobelt, H., and Rush, A. GLTR：生成文本的统计检测与可视化。载于 Costa- juss`a, M. R. and Alfonseca, E. (编)，计算语言学协会第*57*届年会论文集：系统演示，第111–116页，佛罗伦萨，意大利，2019年7月。计算语言学协会。

Guo, H., Cheng, S., Jin, X., Zhang, Z., Zhang, K., Tao, G., Shen, G., and Zhang, X. Biscope：通过检查前序词元的记忆情况检测人工智能生成文本。神经信息处理系统进展，37：104065–104090，2024。Hou, J., Zhao, J., Zhang, S., and Wang, P. 检测大模型生成时间序列的理论分析。*arXiv*预印本 *arXiv:2511.07104*，2025。Hu, Q. J., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ran- ganath, G., Keutzer, K., and Upadhyay, S. K. Router- bench：多大语言模型路由系统的基准。*arXiv*预印本 *arXiv:2403.12031*，2024。Hu, X., Chen, P.-Y., a nd Ho, T.-Y. Radar：通过对抗学习实现鲁棒的人工智能文本检测。神经信息处理系统进展，36：15077–15095，2023。Huang, Z., Ling, G., Lin, Y., Chen, Y., Zhong, S., Wu, H., and Lin, L. Routereval：用于探索大语言模型模型级扩展的大语言模型路由综合基准。*arXiv*预印本 *arXiv:2503.10657*，2025。Jitkrittum, W., Narasimhan, H., Rawat, A. S., Juneja, J., Wang, C., Wang, Z., Go, A., Lee, C.-Y., Shenoy, P., Pani- grahy, R., et al. 面向高效大语言模型推理的通用模型路由。*arXiv*预印本 *arXiv:2502.08773*，2025。Li, L., Wang, P., Ren, K., Sun, T., and Qiu, X. 大语言模型的来源追踪与检测。*arXiv*预印本 *arXiv:2304.14072*，2023。Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., and Zhang, Y. Mage：真实环境中的机器生成文本检测。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第36–53页，2024。Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*，2019。

Milano, S., McGrane, J. A., 和 Leonelli, S. 大语言模型对高等教育的未来构成挑战。*NatureMachine Intelligence*，5(4):333–334，2023年。Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., 和 Finn, C. 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，第 24950–24962 页。PMLR，2023年。Narayan, S., Cohen, S. B., 和 Lapata, M. 不要给我细节，只要摘要！用于极限摘要的主题感知卷积神经网络。*arXiv*，abs/1808.08745，2018年。OpenAI。GPT-oss-20B。https://huggingface.co/openai/gpt-oss-20b，2025年。开源的 20B 参数 GPT 模型。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等。语言模型是无监督的多任务学习器。*OpenAI*博客，1(8):9，2019年。Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Ki m, J. W., Kreps, S., 等。语言模型的发布策略及其社会影响。*arXiv* 预印本 *arXiv:1908.09203*，2019年。Song, Y., Yuan, Z., Zhang, S., Fang, Z., Yu, J., 和 Liu, F. 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，2025年。Su, J., Zhuo, T., Wang, D., 和 Nakov, P. DetectLLM：利用对数秩信息对机器生成文本进行零样本检测。载于 Bouamor, H., Pino, J., 和 Bali, K.（编），计算语言学协会会议论文集：*EMNLP 2023*年研究成果，第 12395–12412 页，新加坡，2023年12月。计算语言学协会。Su, Z., Wu, X., Zhou, W., Ma, G., 和 Hu, S. HC3 Plus：语义不变的人类与 ChatGPT 对比语料库。*arXiv* 预印本 *arXiv:2309.02731*，2023年。Sun, K., Bao, G., Cui, H., 和 Zhang, Y. 当人工智能趋于稳定：后期稳定性作为人工智能生成文本检测的标志。*arXiv* 预印本 *arXiv:2601.04833*，2026年。Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., 和 Liu, Z. 大语言模型的特异性。*arXiv* 预印本 *arXiv:2502.12150*，2025年。Wang, B. 和 Komatsuzaki, A. GPT-J-6B：一个拥有 60 亿参数的自回归语言模型。
h
t
t
p
s://
github.com/kingoflolz/mesh-transformer-jax，2021年。访问日期：2025年11月29日。

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., 和 Wong, D. F. 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，2025年。Xu, Y., Wang, Y., An, H., Liu, Z., 和 Li, Y. 利用相对似然谱检测人类语言与模型语言之间的细微差异。*arXiv* 预印本 *arXiv:2406.19874*，2024a。Xu, Y., Wang, Y., Bi, Y., Cao, H., Lin, Z., Zhao, Y., 和 Wu, F. 通过挖掘词元概率序列实现无需训练的大语言模型生成文本检测。*arXiv* 预印本 *arXiv:2410.06072*，2024b。Yang, X., Pan, L., Zhao, X., Chen, H., Petzold, L. R., Wang, W. Y., 和 Cheng, W. 大语言模型生成内容检测综述。载于 Al-Onaizan, Y., Bansal, M., 和 Chen, Y.-N.（编），计算语言学协会会议论文集：*EMNLP 2024*年研究成果，第 9786–9805 页，美国佛罗里达州迈阿密，2024年11月。计算语言学协会。Yu, X., Qi, Y., Chen, K., Chen, G., Yang, X., Zhu, P., Shang, X., Zhang, W., 和 Yu, N. DPIC：通过解耦提示与内在特征检测大语言模型生成文本。载于 Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., 和 Zhang, C.（编），神经信息处理系统进展，第 37 卷，第 16194–16212 页。Curran Associates, Inc.，2024年。Yu, X., Yu, Y., Liu, D., Chen, K., Zhang, W., Yu, N., 和 Shao, J. EvoBench：面向不断演进的大语言模型的真实场景大语言模型生成文本检测基准测试。载于计算语言学协会会议论文集：*ACL 2025*，第 14605–14620 页，2025年。Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., 和 Zettlemoyer, L. OPT：开放式预训练变换器语言模型，2022。

A. 理论分析

本节给出命题 3.1 的完整证明，说明零样本检测中的性能差距从根本上受源模型与代理模型之间的分布散度所限。

A.1. 预备知识

我们介绍概率散度所需的背景知识。

定义 A.1 （全变差距离）。对于定义在可测空间 $(\mathcal{X}, \mathcal{F})$ 上的两个概率分布 P 和 Q，全变差距离为

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx \tag{13}$$

其中，p 和 q 分别表示 P 和 Q 的概率密度函数。

定义 A.2 (Kullback-Leibler 散度). 对于两个概率分布 P 与 Q，且 P 关于 Q 绝对连续，KL散度为

$$D_{KL}(P \| Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{14}$$

引理 A.3 (Pinsker 不等式). 对于任意两个概率分布 P 与 Q:

$$d_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)} \tag{15}$$

这一经典结果将全变差距离与KL散度联系起来：前者可直接用于界定期望差异，后者具有信息论解释，且通常更易于估计。

A.2. 失配风险界的证明

证明. 设 $T(x)$: $\mathcal{X} \to \mathbb{R}$ 为零样本评分函数，且对所有 $x \in \mathcal{X}$ 有 $|T(x)| \leq B$。设 P_{src} 表示源分布， P_{sur} 表示代理分布，其对应的密度函数分别为 p_{src} 和 p_{sur}。

最优检测信号为

$$\mu^* = \mathbb{E}_{x \sim P_{src}}[T(x)] = \int_{\mathcal{X}} T(x) p_{src}(x) \, dx \tag{16}$$

使用代理模型获得的代理信号为

$$\mu_{proxy} = \mathbb{E}_{x \sim P_{sur}}[T(x)] = \int_{\mathcal{X}} T(x) p_{sur}(x) \, dx \tag{17}$$

步骤 *1*：将差距表示为积分。性能差距可以写成

$$|\mu^* - \mu_{proxy}| = \left| \int_{\mathcal{X}} T(x) p_{src}(x) \, dx - \int_{\mathcal{X}} T(x) p_{sur}(x) \, dx \right| \tag{18}$$

$$= \left| \int_{\mathcal{X}} T(x) \left(p_{src}(x) - p_{sur}(x) \right) dx \right| \tag{19}$$

步骤 *2*：应用有界性条件。由于 $|T(x)| \leq B$，我们有

$$\left| \int_{\mathcal{X}} T(x) \left(p_{src}(x) - p_{sur}(x) \right) dx \right| \leq \int_{\mathcal{X}} |T(x)| \cdot |p_{src}(x) - p_{sur}(x)| \, dx \tag{20}$$

$$\leq B \int_{\mathcal{X}} |p_{src}(x) - p_{sur}(x)| \, dx \tag{21}$$

步骤 *3*：联系全变差距离。根据定义：

$$\int_{\mathcal{X}} |p_{src}(x) - p_{sur}(x)|\, dx = 2 \cdot d_{TV}(P_{src}, P_{sur}) \tag{22}$$

因此：

$$|\mu^* - \mu_{proxy}| \leq 2B \cdot d_{TV}(P_{src}, P_{sur}) \tag{23}$$

步骤 *4*：应用 *Pinsker* 不等式。根据引理A.3：

$$|\mu^* - \mu_{proxy}| \leq 2B \cdot d_{TV}(P_{src}, P_{sur}) \tag{24}$$

$$\leq 2B \cdot \sqrt{\frac{1}{2} D_{KL}(P_{src} \| P_{sur})} \tag{25}$$

$$= B \cdot \sqrt{2 D_{KL}(P_{src} \| P_{sur})} \tag{26}$$

至此，证明完毕。 □

A.3. 对零样本检测的启示

失配风险界对零样本大语言模型检测有几项重要启示。

散度单调性。该界随 $D_{KL}(P_{src} \| P_{sur})$单调递增。源模型与代理模型之间的分布差异越大，潜在的性能下降也越严重。这一理论预测与我们在表1中的实证观察完全一致：与匹配设置相比，跨模型性能显著下降。

路由最优性。 给定一个由候选代理模型 $\{M_1, M_2, \ldots, M_K\}$及其对应分布$\{P_{M_1}, P_{M_2}, \ldots, P_{M_K}\}$组成的池，误差界表明，对于源模型 P_{src}，最优代理模型为

$$M^* = \arg\min_{k \in [K]} D_{KL}(P_{src} \| P_{M_k}) \tag{27}$$

这种最小化收紧了误差界，并在替代模型池中实现了最佳检测性能，从而直接为我们的路由方法提供了依据。

架构亲和性。 图 2 中观察到的结构化模式（同一模型家族中的模型具有更高的相互亲和性）可以通过该框架加以解释。共享架构设计选择和预训练数据的模型，其KL散度往往较小，因此误差界更紧，跨模型检测性能也更好。

黑盒设置中的方差。 图 1 中观察到的极端性能方差，对应于不同代理模型 $D_{KL}(P_{src} \| P_{M_k})$ 之间不同的 M_k 值。对于固定的黑盒源模型，有些代理模型的分布恰好与源模型更接近，因此检测AUC较高；另一些代理模型则存在较大的散度，导致性能接近随机水平。

A.4. 有限样本扩展

在实践中，我们处理的是有限样本而非真实分布。该界可以扩展以考虑估
计误差，使用标准的集不等式。

推论 A.4（有限样本界）。 设 $\hat{}\mu^*$ 与 $\hat{}\mu_{proxy}$ 为基于 n 个独立同分布样本得到的 μ^* 与 μ_{proxy} 的经验估计。则至少以概率 $1 - \delta$：

$$|\hat{\mu}^* - \hat{\mu}_{proxy}| \leq B \cdot \sqrt{2 D_{KL}(P_{src} \| P_{sur})} + 2B \sqrt{\frac{\log(2/\delta)}{2n}} \tag{28}$$

该加法 额外项反映了有限抽样带来的统计不确定性，随着 $n \to \infty$ 的增加而消失 .

表 *5.* 在八个模型家族上对 MAGE 的检测性能。曲线下面积以百分比表示。每列中的最佳结果以粗体突出显示。

方法	OpenAI	GLM	Human-Para	FLAN	LLaMA	BigScience	OPT	EleutherAI	Avg
似然	89.12	91.30	60.61	30.98	**92.96**	37.51	31.21	0.26	56.81
熵	23.08	17.98	62.83	66.65	20.30	66.22	76.39	81.76	51.90
Rank	73.74	83.12	53.58	34.93	81.74	45.49	52.75	46.63	58.99
LogRank	87.05	90.28	53.81	31.56	91.69	37.38	32.39	23.10	55.91
LLR	72.90	74.24	36.21	36.17	78.09	39.97	40.17	36.52	51.78
Fast-DetectGPT	76.40	80.41	61.92	**76.29**	67.35	72.40	64.61	58.12	69.69
Lastde	74.47	83.87	62.96	66.35	72.01	64.63	72.35	69.07	70.71
FourierGPT	59.31	84.88	68.32	4.72	84.39	39.60	57.55	63.97	60.34
Diveye	87.89	**92.61**	62.10	34.65	93.80	52.00	67.60	66.11	69.60
UCE	18.81	20.02	70.60	67.05	23.48	62.86	**82.81**	**87.73**	54.17
DetectRouter	**93.24**	90.37	**88.52**	76.66	72.55	**74.65**	67.56	59.81	**77.92**

A.5. 与信息论的联系

界中出现 KL散度并非偶然。从信息论的角度来看，$D_{KL}(P_{src}\|P_{sur})$量化了使用 P_{sur} 近似 P_{src}时所损失的信息。在大语言模型检测中，这种信息损失会直接导致检测性能下降，因为代理模型无法捕捉源模型独特的统计模式。

这种联系表明，可以将有效路由视为一个信息论信道匹配问题：我们旨在将每个输入路由至信息损失最小的代理模型，从而最大限度地提高检测信号的保真度。

B. 补充实验结果

本节提供全面的实验结果，以补充论文正文。

B.1. MAGE 基准完整结果

表 5 展示了 MAGE 基准上按模型家族划分的完整结果。结果显示，不同模型家族的基线性能存在显著差异。例如，似然方法在 LLaMA 家族文本上的准确率达到 92.96%，但在 EleutherAI 上仅为 0.26%，这反映出架构差异较大时，代理模型与来源不匹配的问题十分严重。我们的路由机制通过动态选择合适的代理模型，有效缓解了这种方差。

B.2. 完整路由增强结果

表格 6 和 7 分别展示了在 EvoBench 和 MAGE 上，将 DetectRouter 应用于各零样本检测方法后，按模型家族划分的完整结果。

B.3. 监督式基线结果

我们使用在第二阶段数据上训练的各代理模型表征进行有监督检测，以完成二元分类。表8 和9 分别给出了在EvoBench和MAGE上的完整结果。不同代理模型的性能差异显著，表明监督式方法对表征模型的选择高度敏感。例如，DeepSeek-7B在EvoBench上达到79.30%，而Falcon-7B仅为48.27%，这凸显了即便在监督式设置中，代理选择同样至关重要。DetectRouter始终优于所有单一代理模型的有监督基线。

表 6. EvoBench 上的路由增强。报告了各检测方法在使用和不使用路由时按模型系列划分的曲线下面积（百分比）。 Δ 表示应用 DetectRouter 后的相对提升。

方法	LLaMA3	LLaMA2	Claude	GPT-4o	Gemini	GPT-4	Qwen	Avg	Δ
似然	88.84	61.68	79.70	77.80	78.89	77.80	77.40	77.44	+5.4%
似然+我们的方法	90.58	66.12	84.56	85.92	84.20	84.44	75.31	81.59	
熵	28.85	57.58	33.61	35.43	38.34	38.10	30.34	37.46	+139.4%
熵+我们的方法	94.78	80.20	92.75	92.24	92.36	92.59	83.01	89.70	
Rank	68.49	58.38	64.50	64.64	64.46	63.43	55.06	62.71	+17.0%
排名+我们的方法	78.13	69.24	75.01	76.35	74.34	75.06	65.48	73.37	
LogRank	86.47	59.99	77.37	76.80	75.99	74.87	76.10	75.37	+9.5%
LogRank+我们的方法	90.26	70.31	83.82	86.07	84.35	84.47	78.67	82.56	
LLR	66.46	49.61	61.73	60.57	59.12	58.43	63.08	59.86	+33.2%
对数似然比+我们的方法	85.97	71.96	80.05	81.57	81.46	80.66	76.68	79.76	
Fast-DetectGPT	94.17	80.27	81.57	74.89	80.48	75.58	80.10	81.01	+12.1%
Fast-DetectGPT+我们的方法	95.66	80.82	94.22	94.29	93.79	94.16	83.01	90.85	

表 7. MAGE 上的路由增强。报告了各检测方法在启用和未启用路由时按类别计算的曲线下面积（百分比）。 Δ表示应用 DetectRouter 后的相对提升。

方法	OpenAI	GLM	Human-Para	FLAN	LLaMA	BigScience	OPT	EleutherAI	Avg	Δ
似然	89.12	91.30	60.61	30.98	92.96	37.51	31.21	0.26	56.81	+14.7%
似然+我们的方法	91.67	85.68	83.46	54.41	72.62	55.02	47.54	30.85	65.16	
熵	23.08	17.98	62.83	66.65	20.30	66.22	76.39	81.76	51.90	+26.5%
熵+我们的方法	76.16	42.35	86.80	68.37	33.29	68.62	73.12	76.37	65.64	
Rank	73.74	83.12	53.58	34.93	81.74	45.49	52.75	46.63	58.99	+15.6%
排名+我们的方法	83.54	85.74	79.94	53.48	74.40	58.91	61.61	47.90	68.19	
LogRank	87.05	90.28	53.81	31.56	91.69	37.38	32.39	23.10	55.91	+22.8%
LogRank+我们的方法	91.94	90.39	80.81	54.65	79.04	58.98	54.19	39.07	68.63	
LLR	72.90	74.24	36.21	36.17	78.09	39.97	40.17	36.52	51.78	+35.4%
对数似然比+我们的方法	89.40	90.17	72.46	59.63	81.28	61.35	59.98	46.77	70.13	
Fast-DetectGPT	76.40	80.41	61.92	76.29	67.35	72.40	64.61	58.12	69.69	+11.8%
Fast-DetectGPT+我们的方法	93.24	90.37	88.52	76.66	72.55	74.65	67.56	59.81	77.92	

B.4. 原型空间几何

为验证所学度量空间的质量，我们使用t-SNE对测试集嵌入进行可视化，如图7所示。该可视化揭示了两个关键的几何特性。首先，不同模型家族生成的文本形成了彼此分离、界限清晰的簇，验证了我们的对比学习目标能够捕捉模型特定伪影。其次，学习得到的原型被准确地定位在各自源簇的质心处。GPT-J与GPT-Neo的簇彼此相邻，反映出它们共享的架构谱系，但仍可区分。这种连贯结构证实，DetectRouter学习到了一个语义流形，其中邻近程度与检测器亲和性相关，从而实现稳健的泛化。

C. 补充分析可视化

本节提供补充可视化，用以补充第3节中的分析，并阐明促使我们采用路由方法的根本性挑战。

C.1. 匹配设置与不匹配设置之间的性能差距。

图 8 展示了从白盒检测设置转向跨模型检测设置时的性能下降。对于五种零样本检测方法，我们比较了代理模型与源模型一致和不一致时的曲线下面积。

表*8*。 使用基于第二阶段数据训练的不同代理表示，在 EvoBench 上进行的有监督检测结果。曲线下面积以百分比表示。

代理模型	LLaMA3	LLaMA2	Claude	GPT-4o	Gemini	GPT-4	Qwen	Avg
LLaMA3-8B	74.96	66.88	73.38	77.66	62.07	76.16	66.49	71.09
DeepSeek-7B	85.71	68.16	83.89	84.64	74.05	84.85	73.81	79.30
Falcon-7B	43.47	56.67	42.32	46.58	50.48	47.86	50.54	48.27
GPT2-XL	61.16	61.81	65.08	62.51	74.42	62.08	61.03	64.01
GPT-J-6B	73.92	61.72	72.85	73.74	54.55	72.66	68.78	68.32
GPT-Neo-2.7B	69.42	62.62	73.42	75.20	69.41	73.81	66.46	70.05
OPT-2.7B	72.79	64.32	68.75	72.12	68.17	72.16	65.64	69.14
Qwen-3B	82.22	64.96	81.18	80.59	76.22	80.81	69.56	76.51
GPT-oss-20B	50.07	57.80	47.90	51.15	50.72	51.93	60.73	52.90
GPT-NeoX-20B	57.93	56.44	58.39	57.56	57.37	57.74	55.89	57.33
DetectRouter	**95.66**	**80.82**	**94.22**	**94.29**	**93.79**	**94.16**	**83.01**	**90.85**

表*9.* 使用第二阶段数据训练的不同代理表示在 MAGE 上的有监督检测结果。曲线下面积以百分比报告。

代理模型	OpenAI	GLM	Human-Para	FLAN	LLaMA	BigScience	OPT	EleutherAI	Avg
LLaMA3-8B	83.11	81.82	82.15	75.26	72.13	74.51	64.41	64.15	74.69
DeepSeek-7B	89.38	82.26	86.84	74.23	71.53	75.80	71.21	69.78	77.63
Falcon-7B	48.88	77.60	72.32	39.83	75.61	47.72	59.56	61.87	60.42
GPT2-XL	52.34	51.10	49.01	48.28	50.41	46.60	39.82	34.43	46.50
GPT-J-6B	81.61	77.60	78.14	63.61	70.17	62.57	50.07	49.31	66.64
GPT-Neo-2.7B	60.85	75.08	60.78	48.37	71.03	53.22	51.28	48.25	58.61
OPT-2.7B	78.13	74.96	74.66	67.43	67.32	62.51	52.10	48.35	65.68
Qwen-3B	88.66	81.14	80.99	68.58	68.20	65.34	62.09	56.71	71.46
GPT-oss-20B	53.76	61.59	63.87	44.68	59.36	51.26	56.76	63.29	56.82
GPT-NeoX-20B	71.42	68.24	69.67	64.51	65.65	64.96	52.24	52.18	63.61
DetectRouter	**93.24**	**90.37**	**88.52**	**76.66**	72.55	**74.65**	**67.56**	59.81	**77.92**

青色柱表示白盒性能，即检测器可以访问真实的源模型；粉色柱表示跨模型性能的平均值，误差条则显示不同代理模型选择带来的方差。结果表明，所有方法的性能都出现了持续且显著的下降。基于似然的检测从 0.81 降至 0.47，低于随机猜测水平。LogRank 的性能下降最为显著，从 0.85 暴跌至 0.50。即使是最稳健的方法 Fast-DetectGPT，也下降了 20 个百分点，从 0.94 降至 0.74。跨模型设置下较大的误差条进一步凸显了固定代理模型方法的不稳定性：性能会因代理模型选择的随意性而大幅波动。

C.2. 黑盒检测性能的方差。

图9详细考察了针对黑盒源模型时检测性能的方差。我们在 MIRAGE 基准上使用 Fast-DetectGPT 准则，将十个开源代理模型分别用于检测由九个专有黑盒模型系列生成的文本：Claude、DeepSeek、豆包、GPT、LLaMA、通义千问、Gemini、Grok 和 Moonshot。每一列代表一个不同的黑盒源模型，各个数据点则显示每个代理模型取得的曲线下面积。箱线图汇总了针对每个源模型的代理模型性能分布。从这一可视化中可以得出两个重要观察结果。首先，每个黑盒源模型都存在最优代理模型，最高曲线下面积从 Claude 的 0.92 到 LLaMA 的 1.00 不等，这表明选对代理模型即可实现有效检测。其次，性能方差极大，多个源模型的曲线下面积跨度超过 0.70。

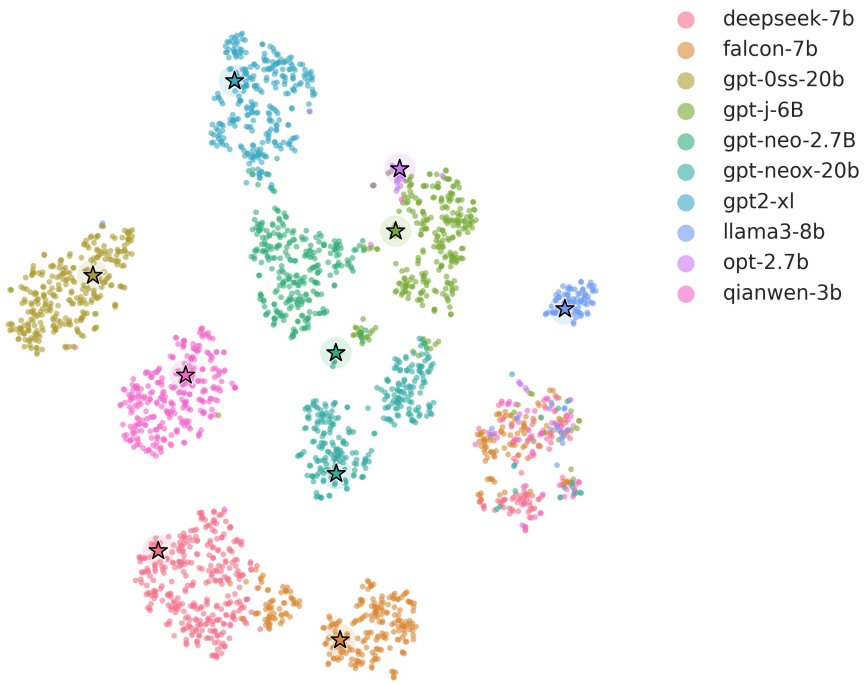

图7. t-SNE 投影的嵌入空间，其中展示了来自不同源模型的文本样本及其对应的原型。

DeepSeek 的范围最广，从 0.15 到 0.99；而 GPT 的范围则从 0.30 到 0.96。这种方差证实，代理选择并非仅仅是优化手段，而是实现可靠检测的基本要求。随机选择代理模型往往会导致接近随机猜测的性能，其数值接近红色虚线所示的 0.5 基线。

D. 实验设置详情

本节提供了训练数据构建模型配置和评估协议的全面详情 s.

D.1. 训练数据构建

数据集概述。 我们的训练基于 MIRAGE 基准（Fu 等人，2025年a），这是迄今为止最全面的多任务机器生成文本检测评估框架。MIRAGE 从五个领域的十个语料库中采集人类撰写文本：学术领域包括 ArXiv 和 NeurIPS，新闻领域包括 CNN DailyMail 和 XSum，评论领域包括 Amazon Review，电子邮件领域包括 Enron Emails，网站领域包括 OpenWebText。该基准采用了 17 个主流大语言模型，其中包括 13 个专有模型和 4 个开源模型，并定义了三种生成任务：生成基于人类撰写文本的前 30 个词元创作新文本，润色在保留细节和含义的同时改进现有文本，以及改写在不改变语义内容的情况下进行释义。所有文本均通过移除特殊字符进行清理，并筛选为 90–220 个单词，以控制基于长度的检测偏差。

阶段 1：原型训练数据。 为进行判别性原型构建，我们基于 MIRAGE-DIG子集，通过不相交输入生成来合成路由器对齐训练数据。对于每条记录，我们将原始字段视为真实人工文本，并提示本地缓存的语言模型为生成、改写和润色任务产出输出。所得记录遵循与源数据相同的 JSON架构，以便无缝集成到现有流水线中。我们为每个任务类别生成 10,000 个样本，共计得到 30,000 个训练实例。

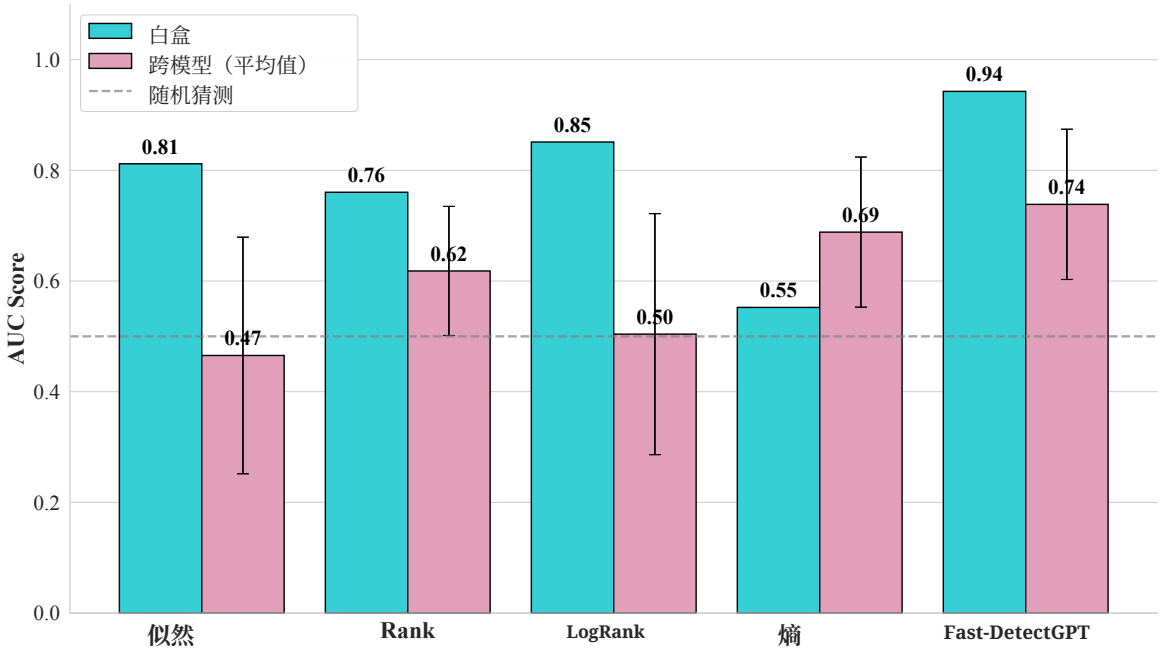

图*8.* 各种检测方法的性能差距。对比了 XSum 上五种零样本指标的白盒检测与跨模型检测性能。跨模型设置下，平均曲线下面积大幅下降且方差增加，表明固定代理模型无法系统性地泛化。此可视化对应表1中的定量分析。

表 *10.* MAGE 基准测试集按模型家族划分的统计数据。

模型家族	测试对
OpenAI	8,180
OPT	7,956
FLAN	4,629
LLaMA	3,684
BigScience	2,672
EleutherAI	1,436
GLM	913
人工释义	795
总计	30,265

阶段 2：距离回归数据。 在黑盒泛化训练中，我们直接使用 MIRAGE 基准，并通过将每段文本输入检测器池来标注检测分数。每个样本由代理模型与检测标准的所有组合进行评分，从而生成作为回归目标的分数向量，用于学习文本与检测器亲和模式。

D.2. 评估基准

我们在两个具有代表性的真实场景基准上进行评估，二者共同考察跨模型泛化与时间鲁棒性。

MAGE 基准。 MAGE（Li 等人，2024）覆盖了包括问答、评论、新闻、学术和故事在内的五个领域，并包含多样化的来源模型，代表了机器生成文本检测中既有的评估协议。表 10 按来源模型家族总结了测试集的组织方式。该基准包含 30,265 个测试对，涵盖八个模型家族，从而能够对跨模型泛化进行全面评估。

EvoBench 基准。 EvoBench（Yu 等人，2025年）专门关注大语言模型的演化特性，既涵盖发布方因版本变更所进行的更新，也包括开发者通过微调和

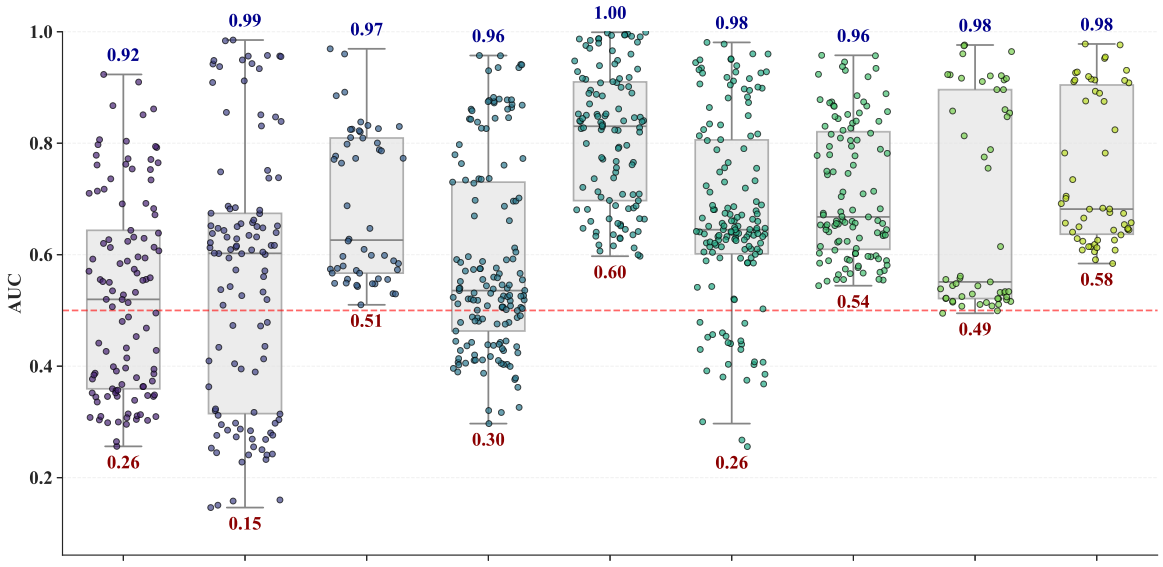

图*9*。 黑盒性能方差详析。在 MIRAGE 基准（Fu 等人，2025a）上使用 Fast-DetectGPT 准则进行评估。每个数据点表示特定开源代理模型在检测来自某个黑盒生成器家族的文本时所达到的曲线下面积。每一列中较大的垂直分布范围表明，对于任何未知来源，都存在一个池中的最优代理模型能够取得较高的曲线下面积，但随机选择很可能会失败，其曲线下面积接近红色虚线基线。这种极端方差凸显了自适应代理模型选择的必要性。

剪枝所带来的发展。例如，GPT-4o 大约每三个月更新一次。该基准涵盖七个大语言模型家族的 29 个演化版本，以及五个领域数据集：XSum、Writing、PubMed、SocialMedia 和 PeerRead，覆盖基于续写、问答和释义生成的范式。该基准直接检验路由机制能否在源模型演化时保持稳健性，这是实际部署的一项关键要求。表11展示了测试集统计数据。

评估指标。 按照机器生成文本检测领域的标准做法（Mitchell 等人，2023年；Bao 等人，2024；Fu 等人，2025a），我们采用受试者工作特征曲线下面积作为主要评估指标。曲线下面积衡量随机选取的一篇机器生成文本的检测分数高于随机选取的一篇人工撰写文本的概率，从而提供对检测器判别能力的阈值无关评估。该指标的取值范围为 0.5（随机猜测）至 1.0（完美区分）。

D.3. 代理模型池

我们的白盒代理模型池由十种开源语言模型组成，涵盖多种架构、训练语料库和参数规模，参数量从 1.5B 到 20B 不等。

DeepSeek-7B（DeepSeek-AI，2023年）是一种拥有 7B 参数、针对高效推理进行优化的稠密型 Transformer。Falcon- 7B（Almazrouei 等，2023年）拥有 7B 参数，使用 RefinedWeb 进行训练，并采用多查询注意力。GPT- oss-20B（OpenAI，2025年）是一种拥有 20B 参数、采用标准架构的开源 GPT 变体。GPT-J-6B（Wang 和 Komatsuzaki，2021年）是一种拥有 6B 参数的自回归 Transformer，由 EleutherAI 使用 The Pile 进行训练。GPT-Neo- 2.7B（Black 等，2021年）是 EleutherAI 开发的模型，拥有 2.7B 参数，并采用局部注意力模式。GPT-NeoX-20B（Biderman 等，2023年）拥有 20B 参数，采用旋转位置嵌入和并行化注意力。GPT2-XL（Radford 等，2019年）是 OpenAI 拥有 1.5B 参数、使用 WebText 训练的基线模型。LLaMA3-8B（Dubey 等，2024）是 Meta 的指令微调模型，拥有 8B 参数，并采用分组查询注意力。OPT-2.7B（Zhang 等，2022）是 Meta 开发的开放式预训练 Transformer，拥有 2.7B 参数。Qwen-3B（Bai 等，2023年）是阿里巴巴开发的多语言模型，拥有 3B 参数。

这种多样性确保了 l　　学习得到的原型涵盖了不同模型系列的多种生成模式　　s,

表 *11.* EvoBench 基准按模型家族划分的测试集统计数据。数据集表示模型版本和领域组合的数量。

模型家族	数据集	测试对
LLaMA-2-7B	74	11,100
GPT4o	35	5,250
Claude	30	4,500
GPT4	28	4,200
LLaMA3	28	4,200
Gemini	15	2,243
Qwen	10	1,500
总计	220	32,993

规模和架构选择。

D.4 检测方法

我们评估了六种零样本检测准则，每种准则都代表了一种从不同统计视角区分机器生成文本与人类撰写文本的方法。

似然（Solaiman 等人，2019年）根据代理模型下词元的平均对数概率进行计算，其依据是这样一种直觉：机器生成文本在语言模型下往往具有更高的似然。熵用于衡量词元级预测不确定性；由于人类文本更难预测，其熵通常更高。排名（Gehrmann 等人，2019年）计算生成词元的平均词汇排名，机器生成文本往往使用排名更高的词元。LogRank（Mitchell 等人，2023年）对词元排名进行对数变换，以提高稳定性。对数似然比（Su 等人，2023a）计算代理模型与参考模型之间的对数似然比，以捕捉相对概率差异。Fast-DetectGPT（Bao 等人，2024）可近似计算条件概率曲率，而无需进行耗费大量计算的扰动。

对于每种准则，我们都将表现最佳的固定代理基线与路由变体进行比较；后者会针对每个输入动态选择最优代理模型。

D.5 路由器编码器架构

我们探索了三种路由网络编码器架构，在计算效率与表征能力之间进行权衡。

RoBERTa-base（Liu et al.，2019年）是我们的默认选择，采用 12 层 Transformer 架构，包含 1.25 亿个参数，每个样本的推理耗时为 4.87 毫秒，效率出色。RoBERTa-large 采用 24 层 Transformer，包含 3.55 亿个参数，能够提供更强的表征。LLM2Vec-LLaMA3-8B（BehnamGhader etal.，2024）通过掩码下一词元预测和无监督对比学习，将 LLaMA3-8B 这一仅解码器大语言模型改造为用于生成稠密嵌入的模型。它取得了总体最佳性能，但每个样本的耗时为 166.2 毫秒。

D.6. 超参数配置

原型设置。我们为每个类别配置 10 个可学习原型，采用余弦相似度，温度为 $\tau = 10$，并使用 L2 归一化特征。原型通过 Xavier 初始化进行初始化。我们采用间隔为 0.5、权重为 10^{-3} 的类间分离，并使用权重为 10^{-4} 的原型范数正则化。

阶段一训练。我们训练 8 轮，批次大小为 32，学习率为 2×10^{-5}，权重衰减为 0.01，预热步数为 200。损失仅由交叉熵分类损失构成。

阶段二训练。 我们训练 8 轮次，批次大小为 32，降低学习率至 10^{-5}，权重衰减为 0.01，预热步数为 200。总损失由交叉熵损失（权重为 1.0）、距离回归损失（权重为 1.0）和原型锚定损失（权重为 1.0）组成，以防止迁移期间发生流形坍塌。

E. 分析实验设置

本节详细介绍　　第 4 节所述代理源亲和度分析的详细实验配置　　3.

E.1. 白盒交叉评估协议

数据集构建。遵循 Fast-DetectGPT（Bao 等人，2024）所确立的协议，我们使用 XSum 数据集（Narayan 等人，2018）作为种子语料库。XSum 包含来自 BBC 的 226,711 篇新闻文章，每篇均配有单句摘要。我们从测试集划分中随机抽取 500 篇文章，并将每篇截取至前 30 个词元，作为生成提示。随后，九个代理模型分别使用核采样生成 150–200 个词元的续写，参数设置为$p = 0.96$，温度为 $T = 1.0$，与 Fast-DetectGPT 的原始配置一致。由此得到九个机器生成文本平行语料库，每个语料库包含 500 个样本。

代理模型池。白盒分析采用了九种开源大语言模型，涵盖多种架构和规模：GPT-Neo-2.7B（Black 等人，2021年）、GPT-J-6B（Wang 和 Komatsuzaki，2021年）、GPT-NeoX-20B（Biderman 等人，2023年）、OPT-2.7B（Zhang 等人，2022）、LLaMA3-8B（Dubey 等人，2024）、Falcon-7B（Almazrouei 等人，2023年）、DeepSeek-7B（DeepSeek-AI，2023年）、Qwen-3B（Bai 等人，2023年）以及 GPT2-XL（Radford 等人，2019年）。参数量从 1.5B 到 20B 不等，涵盖 GPT-2 和 GPT-Neo 系列等早期架构，也包括 LLaMA3 和 DeepSeek 等较新的模型。

交叉评估矩阵。 我们构建一个 9×9 评估矩阵，其中每一行代表一个源模型，每一列代表一个代理模型。对于每个源模型-代理模型组合，我们使用代理模型对源模型生成的 500 个机器生成样本进行检测，并配以取自原始 XSum 文章的 500 个人类撰写样本，据此计算检测分数。我们报告每个单元格的曲线下面积，共得到 81 个评估点，从而展现跨模型检测性能的全貌。

检测准则。 白盒分析评估五种零样本检测准则：似然、熵、排名、LogRank 和 Fast-DetectGPT。对于每项准则，我们构建完整的 9×9 亲和矩阵。图 2 在论文正文中展示了 Fast-DetectGPT 矩阵；其他准则对应的矩阵也呈现出类似的对角线占优模式，但各自具有特定差异。

E.2. 黑盒评估协议

基准选择。 对于黑盒分析，我们采用 MIRAGE 基准（Fu 等人， 2025a），该基准提供由 GPT-4、GPT-4o、Claude、Gemini、DeepSeek、Grok、LLaMA、通义千问和 GLM 等主流商业大语言模型生成的文本。该基准涵盖五个领域和三种生成任务，全面覆盖了现实世界中的黑盒场景。

代理模型遍历。 我们在 MIRAGE 中的所有黑盒源模型系列上，逐一评估九个开源代理模型。对于每个源模型-代理模型组合，我们使用 Fast-DetectGPT 准则计算曲线下面积。由此得到每个来源类别的性能分布，并从中提取各代理模型的最小、中位数和最大曲线下面积。

性能差距计算。 图 1 中报告的性能差距，按每个来源类别最佳与最差代理模型之间的相对差异计算：差距 $=$（最大 $-$最小）/最大 $\times 100\%$。该指标量化了与不幸的随机选择相比，最优代理选择可能带来的提升幅度。

E.3. 定量汇总统计

表 1汇总了正文中的交叉评估结果，并给出了总结性统计数据。对于白盒设置，我们将对角线 AUROC 的平均值作为匹配条件，将非对角线 AUROC 的平均值作为交叉条件。对于黑盒设置，我们报告所有源模型-代理模型对的平均 AUROC，以及各源模型最大和最小 AUROC 值的平均值。我们针对每项检测准则分别计算这些统计数据，以揭示对代理选择的准则特定敏感性。

F. 局限性

我们讨论了所提出方法的若干局限性。

首先，DetectRouter 在推理过程中会带来额外的计算开销。路由器需要对每条输入文本进行编码，并计算其与所有原型之间的距离，然后才能选择最优检测器。虽然 RoBERTa-base 等轻量编码器的延迟为每个样本 4.87ms，尚可接受，但性能最佳的 LLM2Vec 编码器每个样本需要 166.2ms，这可能会限制其在高吞吐量场景中的适用性。

其次，我们的路由框架的有效性取决于代理模型池的覆盖范围。当遇到来自某个源模型的文本，且该源模型的分布与池中所有代理模型都相距甚远时，路由器可能无法找到匹配良好的检测器。扩大代理模型池可以缓解这一问题，但会增加训练和推理成本。

第三，阶段二训练需要将来自多种黑盒来源的检测分数作为监督信号。收集这类标注需要在大规模数据集上运行多个检测器，这会在数据准备阶段产生大量计算开销。

最后，我们当前的评估聚焦于英文文本。DetectRouter 在多语言环境中的泛化能力尚未得到探索，可能需要特定语言的代理模型池或跨语言适应策略。