Computer Science and Application 计算机科学与应用, 2026, 16(8), 149-159
Published Online August 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.168270

Hans汉斯

基于嵌入空间差异学习生成文本检测方法研究

裴超凡，张　娟*，李子臣

北京印刷学院数字版权保护技术研究中心，北京

收稿日期：2026年7月6日；录用日期：2026年8月10日；发布日期：2026年8月18日

摘　要

随着大语言模型生成能力的持续增强，机器生成文本检测成为保障信息可靠性的关键技术。现有研究普遍采用因果语言模型在概率空间中建模，存在判别力与计算成本紧密均衡的问题，且对输入形态变化的鲁棒性仍有待增强。本文提出一种基于嵌入空间差异学习的生成文本检测方法，将差异学习的边际约束由概率空间迁移至嵌入空间，以人类文本嵌入质心为参照点、采用欧氏距离构造差异函数，有效区分人机文本；同时辅以双态训练机制以增强模型对不同输入的检测鲁棒性。在MIRAGE-DIG与自建DeepSeek-Bench数据集上实验结果表明，仅使用约12.4%骨干参数量，取得与因果语言模型评分方法相当的检测精度。研究结果为生成文本检测提供了一条新技术路径。

关键词

机器生成文本检测，大语言模型，自然语言处理，鲁棒性，嵌入空间差异学习

Research on a Generated-Text Detection Method Based on Embedding-Space Difference Learning

Chaofan Pei, Juan Zhang*, Zichen Li

Digital Copyright Protection Technology Research Center, Beijing Institute of Graphic Communication, Beijing

Received: July 6, 2026; accepted: August 10, 2026; published: August 18, 2026

Abstract

With the continuous improvement of the generative capabilities of large language models, detecting machine-generated text has become a key technology for ensuring information reliability. Existing studies typically model text in the probability space using causal language models, which tightly couples discriminative power with computational cost and offers limited robustness to changes in

*通讯作者。

文章引用: 裴超凡, 张娟, 李子臣. 基于嵌入空间差异学习生成文本检测方法研究[J]. 计算机科学与应用, 2026, 16(8): 149-159. DOI: 10.12677/csa.2026.168270

input form. This paper proposes a generated-text detection method based on embedding-space difference learning, transferring the margin constraint of difference learning from the probability space to the embedding space. Using the centroid of human-text embeddings as the reference point and the Euclidean distance to construct the difference function, the method enables a bidirectional encoder to effectively distinguish between human- and machine-generated text; a dual-state training mechanism is further introduced to improve robustness to different inputs. Experiments on the MIRAGE-DIG benchmark and a self-built DeepSeek-Bench dataset show that, using only about 12.4% of the backbone parameters, the method achieves detection accuracy comparable to causal-language-model scoring methods. These results indicate that the discriminative mechanism of difference learning can be effectively established in the embedding space, providing a new technical approach for generated-text detection.

Keywords

Machine-Generated Text Detection, Large Language Model, Natural Language Processing, Robustness, Embedding Space Difference Learning

1. 引言

随着大语言模型生成能力的持续增强，以 ChatGPT [1]、DeepSeek-R1 [2]、QwQ-32B [3]为代表的大语言模型在语法规范性、语义连贯性以及内容组织能力等方面已接近人类写作水平，并被广泛应用于内容创作、智能问答和文本改写等场景。同时，机器生成文本的滥用也对新闻真实性、学术诚信和网络舆论治理构成潜在威胁[4]。因此机器生成文本检测作为一项判别给定文本是由人类撰写还是由机器生成的二分类任务[4] [5]，已成为自然语言处理与人工智能安全领域的重要研究方向。

现有检测方法中，以直接差异学习[4]为代表的一类因果语言模型评分方法，通过构建显式间隔约束对人机文本进行判别，并取得了不错的检测性能。然而该类方法构建的差异函数需要对待检测文本逐词元计算概率分布；其骨干模型参数规模达数十亿量级，推理延迟与显存开销十分显著，难以适应大规模在线检测场景[4] [6] [7]。值得注意的是，直接差异学习的判别力来源于人机文本在差异得分上的可分离间隔结构，而非概率空间本身；这意味着若能在更紧凑的语义表征空间中重建等价的间隔结构，检测性能与计算成本之间的联系可以解除。此外推理型大语言模型输出的显式推理链标签在实际传播中常被删除或截断，进一步对检测方法的鲁棒性提出要求。基于此，如何将差异学习机制的判别能力有效迁移至嵌入空间，并增强对不同输入的检测鲁棒性，构成了本文的核心研究问题。

基于以上问题，现有嵌入空间生成文本检测方法虽借助双向编码器降低计算成本，但判别方法多以相似性对比或类别分类为目标[8]，且传统交叉熵分类目标容易使模型依赖浅层特征而非深层语义差异[9] [10]，因而既难以复刻概率空间下的判别力，也无法消除格式捷径学习。基于此，本文提出一种基于嵌入空间差异学习的生成文本检测方法，引入直接差异学习的间隔优化思想，将其差异计算方式由概率空间迁移至嵌入空间，并以人类文本质心为参照、通过欧氏距离构造可优化差异得分，并引入基于指数移动平均的动态质心更新机制保持语义参照点的稳定性；同时辅以双态训练机制增强模型对不同输入的检测鲁棒性。本文的主要贡献概括如下：

(1) 提出嵌入空间差异学习方法，将直接差异学习的间隔约束由概率空间迁移至嵌入空间，以人类文

本质心为参照构造欧氏差异得分，并通过指数移动平均机制实现质心的动态更新。实验表明，所提方法以约 12.4%的骨干参数量取得与因果语言模型评分方法相当的检测精度，验证了差异学习的间隔判别方法在嵌入空间中的有效性。

(2) 针对推理型大语言模型输出中推理链被截断的问题，引入双态训练机制作为辅助机制，增强检测方法对不同输入的检测鲁棒性。

2. 生成文本检测相关研究

现有机器生成文本检测研究可从检测方式上概括为零样本检测与基于训练的检测两类[4]。本节在梳理这两类方法研究现状基础上，将进一步聚焦差异学习这一技术路径，剖析其从概率空间迁移至嵌入空间的有效性与不足，为本文方法奠定基础。

2.1. 机器生成文本检测研究现状

零样本检测方法利用语言模型的概率评分、词元排序以及扰动前后的统计差异实现判别，无需针对特定大语言模型生成内容进行训练。早期工作多采用对数似然、困惑度、词元秩等统计指标判定文本来源[5]。在此基础上，Gehrmann 等人提出 GLTR，通过分析词元的概率排序分布并加以可视化，辅助人工判别文本来源[11]；Su 等人提出 DetectLLM 引入对数秩等特征以增强零样本检测方法性能[12]；Venkatraman 等人提出的 GPT-Who 则从信息密度层面刻画人机文本的分布差异[13]；陈鹏飞等人通过融合自一致性与对数秩进一步改进了零样本黑盒检测[14]。然而，单一概率指标易受文本长度和评分模型选择等因素影响。为提升鲁棒性，Mitchell 等人提出了 DetectGPT，通过比较原始文本与其扰动版本的对数概率差异进行判别[6]；Bao 等人在此基础上提出 Fast-DetectGPT，利用条件概率曲率显著降低了计算复杂度[7]；Hans 等人提出的 Binoculars 通过对比两个相近语言模型的交叉困惑度构造判别指标，在零样本设定下取得了较强的检测性能[15]。部分研究进一步将概率差异从后验指标转化为可优化的训练目标。Chen 等人提出的 ImBD 在 Fast-DetectGPT 上引入直接偏好优化[16]；Fu 等人提出的 DetectAnyLLM 设计了直接差异学习方法，将差异得分作为优化对象，使人类撰写与机器生成文本在差异空间上形成可控间隔[4]。

基于训练的检测方法则利用预训练编码器提取文本语义表征，在表征空间中学习判别边界。BERT 等双向编码器在文本分类任务上展现出良好的上下文建模能力[9]，因而被广泛用于生成文本检测；Verma 等人提出的 Ghostbuster 通过多个小语言模型提取概率特征并训练分类器[10]。LaCava 等人提出的 WhosAI 结合 Transformer 与对比学习，采用三元组损失学习文本相似性，并以样本嵌入到类别质心的距离判定文本来源[8]，验证了“文本嵌入–类别质心距离”范式在生成文本检测中的有效性。

总的来说，零样本检测方法在判别效果上不断取得进展，但普遍依赖因果语言模型进行逐词元概率计算，骨干参数规模常达数十亿级，计算成本显著；基于训练的方法借助双向编码器降低了计算开销，但其训练目标多服务于相似性对比或类别分类，缺乏对差异得分的显式间隔约束。上述现状表明，如何在保留差异学习间隔判别能力的同时摆脱对大规模因果语言模型的依赖，是当前生成文本检测研究需要解决的问题。

2.2. 差异学习与嵌入空间检测

在上述方法中，Fu 等人提出的直接差异学习[4]不再将检测视为简单的分类任务，而是通过计算差异得分建立显式间隔约束，使人类撰写与机器生成文本在差异空间上形成分离结构。这一间隔优化思想为检测方法提供了明确的判别依据，是当前生成文本检测方法中具有代表性的训练目标设计。

然而，该方法在原始形式中将差异函数实现为因果语言模型的概率曲率，致使判别力的提升与计算成本的上升始终紧密耦合。值得注意的是，间隔优化目标本身对差异函数的具体形式并无严格限定，其

判别力本质上源于人机文本在差异得分上的可分离间隔结构，而非概率空间本身。这为将差异学习从概率空间迁移至嵌入空间提供了理论可能。

另一方面，嵌入空间中已有的检测工作为上述迁移奠定了部分基础。WhosAI [8]验证了以类别质心距离作为判别依据的可行性，表明双向编码器的嵌入空间具备承载人机差异信息的能力。然而，现有嵌入空间检测方法多以相似性对比或类别分类为目标，并未引入显式间隔约束，且常用的交叉熵损失易使模型关注表层格式特征而忽视深层语义差异[9] [10]。因此，这些方法尚未建立起与概率空间下直接差异学习等价的间隔判别结构。

综上，直接差异学习的间隔优化思想具备向嵌入空间迁移的理论基础，嵌入空间中的质心距离范式则为之提供了结构参照；但如何在嵌入空间中构造显式间隔约束以兼顾判别力与计算效率，仍是当前研究尚未解决的问题。本文的方法设计正是围绕这一问题展开，详见第 3 节。

3. 基于嵌入空间差异学习生成文本检测方法

上一节分析表明，直接差异学习的间隔优化思想是概率空间检测方法的优势，而嵌入空间检测方法虽然在计算效率上占优，却缺乏与之等价的间隔约束。将差异学习迁移至嵌入空间，方法设计上需解决两个问题。其一，如何在嵌入空间中构造与概率空间等价的差异函数与间隔约束；其二，在双向编码器持续微调、嵌入分布不断变化的条件下，如何保证差异度量的语义参照稳定。本节将围绕这两个问题展开。3.1 节阐明将差异学习从概率空间迁移至嵌入空间的理论基础；3.2 节给出嵌入空间差异学习的具体构造，包括差异函数定义、质心动态更新机制，以及面向不同输入的辅助训练机制。方法整体框架如图 1 所示。

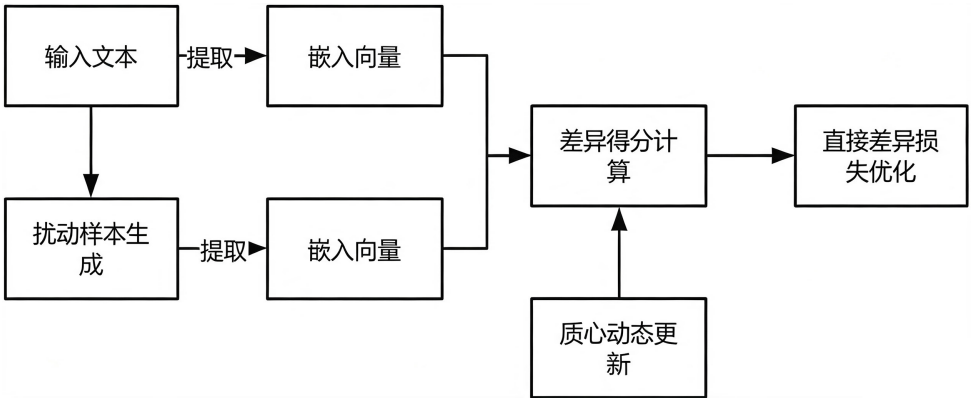

Figure 1. Overall framework of the embedding space difference learning method for text generation detection
图 1. 嵌入空间差异学习生成文本检测方法总体框架

3.1. 从概率空间到嵌入空间的差异学习

机器生成文本检测通常被建模为二分类任务。传统做法以交叉熵损失最小化为目标，仅约束模型对类别标签的拟合，并不约束人机文本在判别空间中的相对位置。这一目标形式在监督信号充足时可获得较高的训练集精度，但其判别边界完全由数据分布隐式决定，模型缺乏人机文本特征信号的显式参照。

直接差异学习[4]则通过为人机文本设置不同的目标得分，将检测任务转换为对差异度量本身的优化。给定人类文本 x_h 与机器文本 x_m，直接差异学习构造差异函数 $d(\cdot)$，并使二者分别趋近于不同目标得分。

$$\ell_{DDL} = \left|d(x_h) - 0\right| + \left|d(x_m) - \gamma\right| \tag{1}$$

其中，γ 表示机器生成文本目标差异距离。该目标使人类文本差异得分趋近于 0，同时使机器生成文本差异得分趋近于 γ，从而在差异得分空间上形成可分离结构。这一性质使直接差异学习相较普通分类目标

具备两点优势：其一，差异得分可直接作为人机判别依据，避免了类别概率在阈值附近的不稳定性；其二，间隔 γ 为模型提供了明确的优化目标，使训练过程对样本质量和数据规模具有更好的鲁棒性。

然而，直接差异学习在原始形式中将差异函数实现为因果语言模型的概率曲率，检测计算需要逐词元调用大规模评分模型。事实上，直接差异学习的间隔优化目标对差异函数的具体形式并无要求，只要差异函数能够稳定刻画文本与人类语言分布的偏离程度，间隔结构就可以在任意可微的语义表征空间中构造。基于这一观察，本文将差异函数由概率空间迁移至嵌入空间。该迁移保留了直接差异学习的间隔优化思想，同时使其计算复杂度由逐词元概率计算降为一次前向传播加一次向量距离计算，从结构上摆脱了对因果语言模型的依赖。

3.2. 生成文本检测方法设计

将差异学习迁移至嵌入空间，需要回答两个子问题：差异函数在嵌入空间中如何构造，以及该构造在编码器持续微调的过程中如何保持稳定。

3.2.1. 差异函数构造

本文采用 BERT-large [9]作为基础编码器。给定输入文本 x，首先将其转换为词元序列并输入 BERT 编码器，随后取最后一层隐藏状态中 $[CLS]$ 位置对应的向量作为文本语义表示：

$$e_x = f_\theta(x)_{[CLS]}, e_x \in R^d \tag{2}$$

其中，$f_\theta(\cdot)$ 表示参数为 θ 的 BERT 编码器，d 为隐藏层维度。设 $c \in R^d$ 为人类文本在嵌入空间中的中心向量，则文本 x 的嵌入空间差异得分定义为：

$$d_e(x,c) = \|e_x - c\|_2 \tag{3}$$

人机文本在嵌入空间中的差异同时体现于向量的方向和模长，两者共同携带判别信息。本文采用欧氏距离作为差异度量，使其同时利用方向与模长两类信号，在保留向量模长信息的同时为优化过程提供更大的几何自由度。

将差异函数代入直接差异学习的间隔目标，得到嵌入空间差异学习的训练损失：

$$\ell_{ES-DDL} = \frac{1}{|B_h|}\sum_{x_h \in B_h} |d_e(x_h) - 0| + \frac{1}{|B_m|}\sum_{x_m \in B_m} |d_e(x_m) - \gamma| \tag{4}$$

其中，B_h 和 B_m 分别表示一个训练批次中的人类文本集合与机器生成文本集合。该损失迫使人类文本表征在方向上靠近人类质心，同时将机器生成文本表征推向远离质心的区域，从而在嵌入空间中重建出与原始直接差异学习等价的间隔结构。目标间隔 γ 控制人机文本在差异得分上的分离程度，其合理取值应与所选距离度量的尺度相匹配。过小则人机文本区域重叠，过大则超出当前嵌入空间的优化范围。

3.2.2. 质心动态更新机制

嵌入空间差异学习以人类质心 c 作为差异得分的参照点，人类质心能否准确刻画当前编码器输出的人类文本分布，直接决定差异得分的语义有效性。已有“文本嵌入–类别质心距离”范式[8]通常将人类质心设为训练前一次性估计的静态向量，其隐含前提是编码器的特征空间在训练全程保持稳定。然而，本文采用 LoRA [17]对编码器进行微调，适配参数的迭代更新会使编码器输出的嵌入分布持续偏移，训练前估计的静态质心将逐渐脱离当前人类文本的实际分布中心，导致差异得分的参照失效。因此，如何在编码器持续微调的过程中维持人类质心的稳定，是差异学习由概率空间迁移至嵌入空间后必须解决的问题。

针对这一问题，本文引入基于指数移动平均的质心动态更新机制，使人类质心的更新速率显著低于编码器参数的更新速率，从而在跟踪嵌入分布整体偏移的同时抑制单批次估计带来的噪声。质心的计算

分为初始化与动态更新两个阶段。

初始化阶段。从训练集中随机选取 N 条人类文本样本，将其 $[CLS]$ 表征均值作为初始参照点：

$$c_0 = \frac{1}{N}\sum_{i=1}^{N} e_{x_h^i} \tag{5}$$

此处使用人类文本的统计均值而非零向量或随机向量进行初始化，使自训练第一步起，差异得分便能承载有效的语义信息，避免训练初期因参照点无效而出现优化方向震荡。

动态更新阶段。在每个训练批次上以批内人类样本表征均值 $\bar{e}_h$ 作为对真实质心的当前观测，并通过指数移动平均将 c 缓慢更新：

$$c \leftarrow \alpha c + (1-\alpha)\bar{e}_h, \bar{e}_h = \frac{1}{|B_h|}\sum_{x_h \in B_h} e_{x_h} \tag{6}$$

其中，动量衰减系数控制质心的更新速率。动量衰减系数取值越大，质心对历史估计的依赖越强、对单批次波动越不敏感。本文取较大值，使动量衰减系数在跟随嵌入分布缓慢演化的同时保持稳定，与 LoRA 微调下表示空间渐变的特性相匹配。

为避免测试信息泄露并保证训练推理一致性，质心在实现上被定义为模型缓冲区，不参与反向传播；动态更新仅基于训练批次中的人类样本进行，验证集与测试集均不参与质心估计；训练完成后质心被冻结，测试阶段统一使用最后一次更新后的固定质心计算差异得分。

3.2.3. 双态训练机制

至此，嵌入空间差异学习已在嵌入空间内重建出直接差异学习的间隔判别能力。然而，当训练数据中机器文本携带显式推理链标签时，该格式特征同样会被编码器映射进向量，模型有动机优先学习这一更易拟合的表层信号。一旦推理链在传播过程中被删除或截断，模型所依赖的格式信号消失，检测性能随之下降。

要削弱这一捷径，关键在于强制模型对同一机器回答的不同输出形态产生一致的高差异响应。基于此，本文引入双态训练策略作为辅助机制。在数据预处理阶段，对训练集中每条包含推理链的机器生成文本同时构造两种形态：

$$x_m^{full} = x_m \tag{7}$$

$$x_m^{strip} = R(x_m) \tag{8}$$

其中，x_m^{full} 表示保留完整链的原始输出，x_m^{strip} 表示经过规则函数 $R(\cdot)$ 删除推理链后的文本。$R(\cdot)$ 通过正则表达式匹配并删除显式推理链标签及其前后的多余空白，仅保留最终回答。具体而言，剥离函数 $R(\cdot)$ 通过正则表达式匹配完整的显式推理链片段，并将其连同前后多余空白一并删除；对于不含<think>标记的机器样本，则视为已处于去链形态，不作处理。对于不包含推理链的机器样本，不进行扩展；人类文本同样不进行扩展，以避免破坏人类样本的自然分布。

经过双态扩展后，机器生成文本集合由 B_m 扩展为 B_m^{dual}。相应地，训练目标可写为：

$$\ell_{ES-DDL+DST} = \frac{1}{|B_h|}\sum_{x_h \epsilon B_h} |d_e(x_h) - 0| + \frac{1}{|B_m^{dual}|}\sum_{x_m \in B_m^{dual}} |d_e(x_m) - \gamma| \tag{9}$$

该损失将完整链与去链文本共同约束到同一目标差异距离 γ。由于推理链标签仅存在于完整链形态，而两种形态被约束到相同的高差异得分，依赖该标签的捷径无法稳定成立，模型转而学习两种形态共享的深层语义特征。双态训练策略并未引入额外对齐损失项或修改模型结构，可无缝叠加在嵌入空间差异学习之上。第 4.4.1 节将通过完整链与去链测试场景的对比验证该策略的实际效果。

4. 实验结果与分析

4.1. 实验设置

数据集与实验划分。本文采用公开基准与自建数据集相结合的实验设置。基准选用 MIRAGE-DIG，用于验证检测精度与计算效率；自建 DeepSeek-Bench 数据集用于验证检测方法对不同输入的鲁棒性。MIRAGE-DIG 是面向机器生成文本检测的大规模公开基准数据集，覆盖多领域、多语料来源和多大语言模型。本文使用 14,776 对润色样本进行训练，并在 15,735 对改写样本上进行测试。

由于现有公开基准较少针对推理型大语言模型的显式推理链形态进行鲁棒性测试，本文进一步构建 DeepSeek-Bench 数据集。该数据集基于 DeepSeek-R1 的推理型输出构建，采用与 MIRAGE-DIG 一致的样本组织格式，共包含 1000 对训练样本与 500 对独立测试样本；每条样本由一段人类撰写文本与对应的机器生成文本配对组成。对每条机器生成文本，同时保留其完整推理链形态(含推理链标签)与去链形态(删除该标签后仅保留最终回答)，以模拟显式推理过程在传播中被删除的检测场景。在数据构建过程中，对生成调用失败或被模型拒绝作答(如触发内容安全策略)的少量样本予以剔除，保留样本均包含完整的推理链。

对比方法与模型配置。为评估本文方法在检测性能与计算效率方面的表现，选取 DetectAnyLLM [4] 作为主要方法对比。该方法以因果语言模型为评分模型，并以直接差异学习构造检测得分，是当前概率空间检测中的代表性工作。本文报告 DetectAnyLLM-500 与 DetectAnyLLM-MIRAGE 两种设置：前者沿用原方法 500 对训练样本的配置，后者使用与本文方法相同的 14,776 对 MIRAGE-DIG 训练样本，以控制训练数据规模差异对结果的影响。

本文方法采用 BERT-large-uncased [9]作为基础编码器，并仅在注意力模块的 Query 和 Value 投影矩阵中引入 LoRA 适配参数。LoRA 秩为 8、缩放系数为 32、Dropout 比例为 0.1。训练批次大小为 16，目标间隔 γ 设为 20，动量衰减系数 α_{EMA} 设为 0.99。

评估指标与实验环境。本文采用 AUROC 作为主要评价指标，并同时报告 AUPR、Matthews 相关系数、平衡准确率以及 5%误报率下的真正率。为量化模型对显式推理链格式的依赖程度，本文定义 $\Delta_{state} = \left| AUROC_{full} - AUROC_{strip} \right|$，其值越小表示模型在两种输入形态下的检测能力越接近。所有实验均在单张 NVIDIA A40 GPU (48 GB 显存)上完成，软件环境为 Python 3.10、PyTorch 2.7.1 和 CUDA 12.1。训练阶段采用 FP16 混合精度，评估阶段采用 FP32 精度。

4.2. 检测精度与计算效率对比

本节在 MIRAGE-DIG Rewrite 跨任务设置下，从检测精度与推理效率两个维度评估本文方法相对于概率空间检测方法[4]的性能与计算开销。

4.2.1. 检测精度对比

Table 1. MIRAGE-DIG Rewrite cross-task evaluation results
表 1. MIRAGE-DIG Rewrite 跨任务评估结果

方法	类型	参数量	AUROC	AUPR	MCC	BalAcc
DetectAnyLLM-500 [4]	Decoder	2.7 B	0.9192	0.9327	0.7233	0.8599
DetectAnyLLM-MIRAGE [4]	Decoder	2.7 B	0.9888	0.9873	0.9575	0.9787
BERT ES-DDL (L2)	**Encoder**	**335 M**	**0.9791**	**0.9515**	**0.9216**	**0.9606**

各方法的检测精度对比结果见表 1。

DetectAnyLLM 在训练规模由 500 对扩展为 14,776 对后性能大幅提升，表明该方法对训练数据规模较敏感，后续以 DetectAnyLLM-MIRAGE 作为更公平的参考基线。在相同训练数据条件下，本文方法取得 0.9791 的 AUROC，与 DetectAnyLLM-MIRAGE 仅相差 0.97 个百分点，但其骨干参数量仅为后者的约 12.4%。该结果验证了直接差异学习的判别性能主要取决于差异得分上的可分离间隔结构，等价的间隔结构在更紧凑的嵌入空间中同样可以构造。

4.2.2. 推理效率对比

Table 2. Inference efficiency comparison
表 2. 推理效率对比

指标	模型参数量	可训练参数	单条推理时间(ms)	GPU 显存峰值(GB)
DetectAnyLLM	2.7 B	2 M	21.0	5.47
BERT ES-DDL	**335 M**	**786 K**	**8.4**	**1.27**

各方法的推理效率对比结果见表 2。

与 DetectAnyLLM 相比，本文方法的单条推理时间降至原方法的约 40%，显存峰值降至原方法的约 23%；批次大小为 32 时吞吐量达到 133.7 条/秒。该效率差异源于计算路径的变化。因果语言模型评分需逐词元执行自回归概率计算，复杂度与文本长度强相关；而嵌入空间差异学习仅需一次前向传播加一次向量距离计算，复杂度由逐词元降为逐文本，并天然支持批处理。

4.3. 方法稳定性验证

为排除单次训练随机性的影响，本文使用 5 个随机种子$(0, 42, 123, 456, 789)$独立训练，并分别在完整链和去链测试集上评估，结果见表 3。

Table 3. ES-DDL + DST multi-seed statistical results
表 3. ES-DDL + DST 多种子统计结果

评估场景	AUROC	MCC	Bal Acc
完整链	0.9936 ± 0.0053	0.9806 ± 0.0094	0.9902 ± 0.0048
去链	**0.9857 ± 0.0020**	**0.9070 ± 0.0060**	**0.9534 ± 0.0029**

所提方法在两种场景下均保持稳定，完整链 AUROC 为 0.9936 ± 0.0053，去链 AUROC 为 0.9857 ± 0.0020，标准差较小，表明所提方法的检测性能不依赖于特定的随机初始化。两种场景下 AUROC 差距为 0.79 个百分点，说明检测器在不同输入形态下也能保持基本一致的判别能力。

4.4. 消融实验

本节通过消融实验分析方法中关键设计选择对最终性能的贡献，依次考察双态训练、目标间隔与超参数取值的影响。所有实验均在 DeepSeek-R1 测试集上进行，除消融项外其余设置保持一致。

4.4.1. 双态训练消融

为评估双态训练机制的贡献，本节比较启用与不启用双态训练机制两种设置下的检测性能，结果见表 4。

如表所示，未启用双态训练时，模型在完整链场景下取得 1.0000 的 AUROC，而去链场景下 AUROC 仅为 0.7609，Δ_{state} 达 23.91 个百分点，表明模型在训练阶段过度依赖了完整链所特有的推理链格式特征。启用双态训练后，去链 AUROC 提升至 0.9870，Δ_{state} 收窄至 1.01 个百分点，完整链 AUROC 则由 1.0000

轻微回落至 0.9971。完整链性能的小幅下降与去链性能的显著提升表明，双态训练通过约束模型在两种形态下产生一致响应，有效抑制了对推理链格式的捷径学习。

Table 4. Dual-state training mechanism ablation (DeepSeek-R1 test set)
表 4. 双态训练机制消融(DeepSeek-R1 测试集)

配置	场景	AUROC	MCC	Bal Acc	TPR@5%
ES-DDL noDST	完整链	1.0000	1.0000	1.0000	1.0000
ES-DDL noDST	去链	0.7609	0.4367	0.7080	0.3020
ES-DDL + DST	完整链	0.9971	0.9900	0.9950	1.0000
ES-DDL + DST	**去链**	**0.9870**	**0.9080**	**0.9540**	**0.9580**

4.4.2. 目标间隔消融

目标间隔 γ 控制机器文本相对于人类质心的目标差异得分，由于不同距离度量的取值范围不同，γ 的合理取值随度量方式变化。本节以欧氏距离为载体扫描 $\gamma \in \{5,10,15,20,25,30\}$，结果如表 5 所示。

Table 5. Target interval ablation
表 5. 目标间隔消融

γ	AUROC	AUPR	MCC	Bal Acc
5	0.7796	0.7906	0.4245	0.7070
10	0.9742	0.9451	0.8677	0.9330
15	0.9739	0.9647	0.8761	0.9380
20	**0.9887**	**0.9783**	**0.9246**	**0.9620**
25	0.9817	0.9661	0.8923	0.9460
30	0.9628	0.9567	0.8303	0.9130

γ 取值对检测性能的影响呈现先升后降的趋势。γ 过小时，机器文本与人类文本在差异得分上的可分离间隔不足以支撑判别，AUROC 仅为 0.7796；γ 过大时，优化目标超出当前嵌入空间所能承载的尺度，AUROC 回落至 0.9628；$\gamma = 20$ 时 AUROC 达到 0.9887。

4.4.3. 超参数敏感性分析

本节评估质心更新速率 α_{EMA} 与 LoRA 秩对方法性能的影响，结果见表 6。

Table 6. Hyperparameter ablation
表 6. 超参数消融

超参数	取值	AUROC	MCC	Bal Acc
α_{EMA}	0.9	0.9791	0.8921	0.9460
α_{EMA}	0.95	0.9758	0.8785	0.9390
$\boldsymbol{\alpha_{EMA}}$	**0.99**	**0.9887**	**0.9246**	**0.9620**
α_{EMA}	0.999	0.9850	0.9080	0.9540
LoRA r	4	0.9829	0.9041	0.9520
LoRA r	**8**	**0.9887**	**0.9246**	**0.9620**
LoRA r	16	0.9829	0.9000	0.9500
LoRA r	32	0.9861	0.8858	0.9420

α_{EMA} 在 0.9 至 0.999 区间变化时，AUROC 波动幅度约 1.3%，表明质心更新速率在该区间内对最终性能不敏感。这一结果表明，当 α_{EMA} 落于合理区间时，指数移动平均更新机制可同时缓和单批次估计噪声与表示空间位移对质心的影响。LoRA 秩在 $r=8$ 时性能达到最高，$r=16$ 与 $r=32$ 下检测性能未进一步提升，表明 $r=8$ 已为本任务提供了充分的可训练参数容量。本文主实验默认采用 $\alpha_{EMA}=0.99$ 与 $r=8$。

4.4.4. 距离度量消融

为剖析所学嵌入空间的判别能力是否依赖于评估阶段所采用的距离度量方式，本节固定以欧氏距离训练所得的检查点(seed = 42，与其余消融实验保持一致)，在评估阶段分别改用欧氏、余弦、马氏三种距离度量重新计算差异得分并评估检测性能，结果如表 7 所示。

Table 7. Distance metric ablation
表 7. 距离度量消融

距离度量	完整链 AUROC	完整链 AUPR	完整链 TPR@5%	去链 AUROC	去链 AUPR	去链 TPR@5%
欧氏距离	0.9971	0.9947	1.0000	0.9870	0.9843	0.9580
余弦距离	0.9972	0.9941	1.0000	0.9875	0.9845	0.9580
马氏距离	0.7791	0.6430	0.0160	0.8613	0.7983	0.2280

如表 7 所示，欧氏距离在两种场景下均取得稳定的高 AUROC，余弦距离在同一嵌入空间上评估时表现几乎一致，二者差距小于 0.1 个百分点；马氏距离因依赖协方差矩阵估计，在 1024 维嵌入空间下有限的训练样本不足以稳定估计协方差结构，性能明显落后。这表明，本文所学的差异表征对评估阶段的距离度量并不敏感，只要度量本身能够稳定刻画文本相对人类质心的偏离方向，即可在差异空间中完成有效判别；而依赖二阶统计量的度量在高维小样本条件下反而不够稳健。该结果与第 3 节将差异学习迁移至嵌入空间的设计思路一致。

5. 结论

本文针对如何将差异学习机制的判别能力有效迁移至嵌入空间，并增强对不同输入的检测鲁棒性，提出了一种基于嵌入空间差异学习的生成文本检测方法。该方法通过将直接差异学习的间隔优化思想由概率空间迁移至嵌入空间，以人类文本质心为参照、通过欧氏距离构造差异得分，并引入指数移动平均保持语义参照的稳定性，同时辅以双态训练机制增强模型对不同输入的检测鲁棒性。实验结果表明，所提方法在 MIRAGE-DIG 公开基准上以约 12.4%的骨干参数量取得与因果语言模型判别方法相当的检测精度，推理时间和显存占用相比原有方法分别降低至 40%和 23%，说明差异学习方法可在嵌入空间中重建出等价结构，而不必依赖因果语言模型的概率空间。下一步工作将探索所提方法在更多大语言模型及隐式推理链场景下的扩展，并在更大规模训练数据上评估其性能上限。

基金项目

北京市高等教育学会(20240004)；基于人工智能大数据的大学生双语故事加工及研究(No.22150124045)；国家自然科学基金(62472040)。

参考文献

[1] Achiam, J., Adler, S., Agarwal, S., *et al*. (2023) GPT-4 Technical Report. arXiv:2303.08774.

[2] Guo, D., Yang, D., Zhang, H., *et al*. (2025) DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, **645**, 633-638.

[3] Qwen Team (2025) QwQ-32B: Embracing the Power of Reinforcement Learning. Qwen. https://qwen.ai/blog?id=qwq-32b

[4] Fu, J., Guo, C. and Li, C. (2025) DetectAnyLLM: Towards Generalizable and Robust Detection of Machine-Generated Text across Domains and Models. *Proceedings of the* 33*rd ACM International Conference on Multimedia*, Dublin, 27-31 October 2025, 11229-11238. https://doi.org/10.1145/3746027.3754862

[5] Fariello, S., Fenza, G., Forte, F., Gallo, M. and Marotta, M. (2025) Distinguishing Human from Machine: A Review of Advances and Challenges in AI-Generated Text Detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, **9**, 6-18. https://doi.org/10.9781/ijimai.2024.12.002

[6] Mitchell, E., Lee, Y., Khazatsky, A., *et al*. (2023) DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. *International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 24950-24962.

[7] Bao, G., Zhao, Y., Teng, Z., *et al*. (2023) Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130.

[8] La Cava, L., Costa, D. and Tagarelli, A. (2024) Is Contrasting All You Need? Contrastive Learning for the Detection and Attribution of AI-Generated Text. In: *Frontiers in Artificial Intelligence and Applications*, IOS Press, 3179-3186. https://doi.org/10.3233/faia240862

[9] Devlin, J., Chang, M.-W., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, June 2 - June 7 2019, 4171-4186. https://doi.org/10.18653/v1/n19-1423

[10] Verma, V., Fleisig, E., Tomlin, N. and Klein, D. (2024) Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *Proceedings of the* 2024 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Volume 1: Long Papers), Mexico City, 16-21 June 2024, 1702-1717. https://doi.org/10.18653/v1/2024.naacl-long.95

[11] Gehrmann, S., Strobelt, H. and Rush, A. (2019) GLTR: Statistical Detection and Visualization of Generated Text. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Florence, 28 July-2 August 2019, 111-116. https://doi.org/10.18653/v1/p19-3019

[12] Su, J., Zhuo, T., Wang, D. and Nakov, P. (2023) DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 12395-12412. https://doi.org/10.18653/v1/2023.findings-emnlp.827

[13] Venkatraman, S., Uchendu, A. and Lee, D. (2024) GPT-Who: An Information Density-Based Machine-Generated Text Detector. *Findings of the Association for Computational Linguistics*: *NAACL* 2024, Mexico City, 16-21 June 2024, 103-115. https://doi.org/10.18653/v1/2024.findings-naacl.8

[14] 陈鹏飞, 孙更新, 宾晟. 融合自一致性与对数秩的大语言模型生成文本检测[J/OL]. 计算机工程与应用: 1-15. https://link.cnki.net/urlid/11.2127.TP.20251230.1417.002, 2026-04-27.

[15] Hans, A., Schwarzschild, A., Cherepanova, V., *et al*. (2024) Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. *Proceedings of the* 41*st International Conference on Machine Learning* (*ICML*), Vienna, 21-27 July 2024, 17519-17537.

[16] Chen, J., Zhu, X., Liu, T., Chen, Y., Xinhui, C., Yuan, Y., *et al*. (2025) Imitate before Detect: Aligning Machine Stylistic Preference for Machine-Revised Text Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 23559-23567. https://doi.org/10.1609/aaai.v39i22.34525

[17] Hu, E.J., Shen, Y., Wallis, P., *et al*. (2022) LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations* (*ICLR*), Virtual Conference, 25-29 April 2022. https://openreview.net/forum?id=nZeVKeeFYf9