

DivScore: 在特定领域中零样本检测 LLM 生成文本

Zhihui Chen^{1*}, Kai He^{1*}, Yucheng Huang², Yunxiao Zhu³, Mengling Feng^{1†}

¹Saw Swee Hock School of Public Health, National University of Singapore

²Interactive Entertainment Group, Tencent,

³Faculty of Law, The University of Hong Kong

zhihui.chen@u.nus.edu, huangyucheng.andrew@gmail.com,

talia_zhu@connect.hku.hk, { kai_he, ephfm } @nus.edu.sg

Abstract

在医学和法律等专业性和高风险领域中，检测由 LLM 生成的文本对于打击错误信息和确保真实性至关重要。然而，目前的零样本检测器虽然在一般文本上有效，但由于领域转移，在应用于专业内容时常常失效。我们提供了一个理论分析，显示这种失败从根本上与人类、检测器和来源文本分布之间的 KL 散度相关。为了解决这一问题，我们提出了 DivScore，一种使用归一化熵评分和领域知识蒸馏的零样本检测框架，以稳健地识别专业领域中 LLM 生成的文本。我们还发布了专门针对医学和法律领域的 LLM 生成文本检测领域特定基准。我们的基准实验表明，DivScore 始终优于最先进的检测器，AUROC 提升了 14.4%，召回率提升了 64.0%（假阳性率阈值为 0.1%）。在对抗性环境中，DivScore 表现出比其他基线更优越的鲁棒性，AUROC 平均优势为 22.8%，召回率为 29.5%。代码和数据已公开提供¹。

1 介绍

随着大型语言模型（LLMs）生成的文本在流畅性和真实性方面不断提升，对可靠的机器生成文本检测的需求变得至关重要。尤其是在医疗和法律等高风险领域，未被检测到的 AI 撰写内容可能会破坏信任和安全。

最初用于检测 LLM 生成文本的方法采用了有监督学习，如 GPT-2 探测器（Solaiman et al., 2019）和 GPT-3.5 探测器（Guo et al., 2023）。虽然这些有监督探测器在它们所训练的数据上表现有效，但在面对新领域或更先进的语言模型时显示出明显的不足。例如，在法律文章上训练的探测器可能在医疗文本上失效，而调整为检测 GPT-2 输出的分类器可能会错误分类来自新 LLM 的文本。此外，每一个新 LLM 或领域都需要创建一个新的标记数据集并重新训练探测器，这使得该过程资源密集且效率低下。

*Equal contribution

†Corresponding author

¹<https://github.com/richardChenzhihui/DivScore>

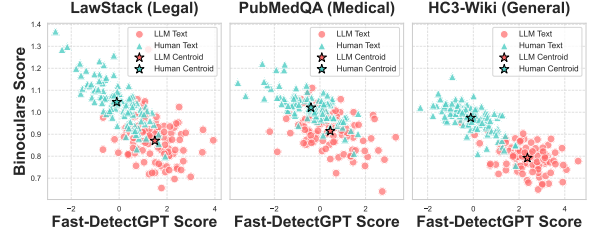


Figure 1: 在专业（法律、医疗）和一般领域中，Fast-DetectGPT (Bao et al., 2023) 和 Binoculars (Hans et al., 2024) 检测器对 LLM/人类文本的评分。在专业领域中，这两个检测器都未能创造出检测 LLM 生成文本所需的评分差距。

作为对策，研究人员开发了不需要训练样本的检测器，而是利用大语言模型（LLM）生成文本的统计不规则性。例如，DetectGPT (Mitchell et al., 2023) 使用 LLM 的概率曲率来标记生成的段落，而无需任何微调，而 Binoculars (Hans et al., 2024) 则对比两个语言模型的输出，达到无需训练数据的最先进的检测准确性。

然而，大多数现有的零样本检测研究主要集中在普通领域数据。像医疗文献或法律文件这样的专业领域提出了几个额外的挑战，使得 LLM 生成文本的检测更加困难。这些领域通常具有：(1) 稀疏的标注数据。用于训练的 LLM 生成文本的注释示例很少或没有。(2) 复杂的语言。特定领域的术语和复杂的写作风格与普通文本显著不同。(3) 不熟悉的内容。即使是高级 LLM 在预训练期间也可能不常见的主题领域和事实。

因此，在一般文本上表现良好的检测方法在这种领域转变下往往会遇到困难。如图 1 所示，我们观察到 Fast-DetectGPT (Bao et al., 2023) 和 Binoculars (Hans et al., 2024) 方法在专业领域中都有压缩倾向：无论是人类撰写的文本还是 LLM 生成的文本都产生了相似的检测分数，削弱了可靠区分所需的差距。为了解决这些差距，我们提出了 DivScore——一种基于归一化熵的散度度量，用于在专业领域进行 LLM 生成文本的无须预训练 (domain-specific,

zero-shot) 检测。我们的方法专门设计为能够在不同领域中保持鲁棒性。

本质上, DivScore 通过针对估计的基线对候选文本的领域特定熵进行归一化, 从而衡量与参考语言模型对该领域的预期相比, 文本的惊讶程度。我们的方法首先强调领域特定的匹配, 然后应用熵归一化来突出人类撰写的文本与 LLM 生成的文本之间的差异, 从而增强检测所需的对比度。我们采用无监督的领域知识蒸馏框架来实现 DivScore, 保持零样本设定, 不需要标记数据。在我们的评估中, DivScore 在显著的领域转换情况下优于现有的最先进 (SOTA) 零样本基线, 同时在一般领域设置中保持竞争力。我们做出以下关键贡献:

- 我们介绍了一种名为 DivScore 的方法。据我们所知, 这是第一个专为识别特定领域的 LLM 生成文本设计的零样本检测方法, 灵感来自领域对齐和熵归一化。
- 我们发布了一个高质量的基准, 用于检测专门领域中由 LLM 生成的文本。它包括专业的法律和医学语料库, 以及知识蒸馏数据集和用于检测器评估和微调的适配流程。
- 我们进行了广泛的实验和分析。结果表明, DivScore 在 AUROC 中比 SOTA 基线高出 14.4 %, 在召回率 (0.1 % 假阳性率) 中高出 64.0 %, 同时在对抗条件下也表现出强大的鲁棒性。

2 相关工作

机器生成文本检测通常被定义为一个二分类问题 (Su et al., 2023; Bao et al., 2023; Mitchell et al., 2023; Solaiman et al., 2019; Hans et al., 2024)。基于监督学习的检测器, 例如基于 RoBERTa 的模型 (Solaiman et al., 2019), 在领域内数据上表现良好, 但需要额外的人为标注, 因而在可扩展性和灵活性上有限 (Bakhtin et al., 2019; Su et al., 2023; Mitchell et al., 2023)。

另一组研究旨在构建基于大型语言模型 (LLM) 测量的人类和机器生成文本之间统计特征差异的零样本检测器。基于文本统计的检测器如 LLM 文本熵 (Lavergne et al., 2008)、对数概率 (Solaiman et al., 2019) 和每个标记排名 (Gehrmann et al., 2019) 已被证明在一般 LLM 文本检测中有效。Mitchell et al. (2023) 提出了 DetectGPT, 通过在受控文本扰动下的 LLM 文本对数概率函数的曲率来检测 LLM 生成的段落。Fast-DetectGPT (Bao et al., 2023) 使用更高效的 LLM 采样步骤代替 DetectGPT 中

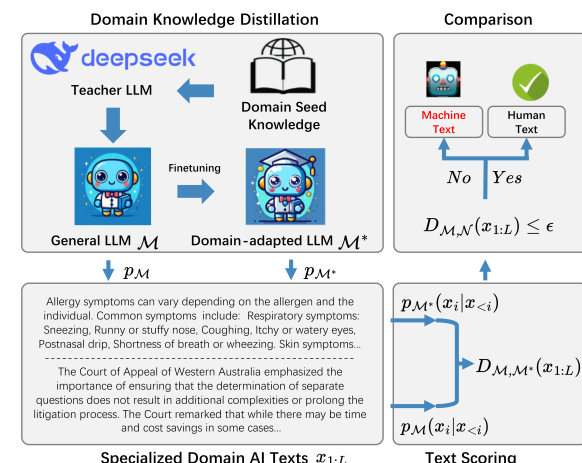


Figure 2: DivScore 的整体框架

的扰动方法。Hans et al. (2024) 进一步发现, 对比两个密切相关的语言模型的 LLM 文本困惑度可以有效地区分 LLM 生成的内容。

然而, 零样本特性以牺牲在专业领域的适应能力为代价。当前的零样本检测器, 包括 Binoculars (Hans et al., 2024), 主要关注普通领域内容, 如新闻、文章和社交媒体帖子 (Narayan et al., 2018; Guo et al., 2023; Fan et al., 2018; Wu et al., 2025a; Dugan et al., 2024)。依赖于普通领域的 LLMs 进行统计特征评分, 当前方法隐式地假设检测器具备充分的领域熟悉度。在具有稀缺知识的专业领域 (例如医学或法律), 由人类和 LLM 生成的文本可能对于这些检测器不熟悉, 导致统计差异缩小并降低检测性能。我们的 DivScore 旨在避免此类问题。

3 方法论

在本节中, 我们首先形式化了为什么当文本分布转移到专业领域时, 零样本检测器会失效, 为 DivScore 奠定了理论基础。在这一分析的指导下, 我们推导出我们的核心检测器, 然后引入一个领域知识蒸馏步骤, 以缩小检测器与专业语言之间的差距, 显著增强鲁棒性和准确性。DivScore 框架如图 2 所示。

3.1 理论分析: 为什么零样本检测器在领域转移下会失败?

大多数零次检测器, 例如基于熵的评分 (Lavergne et al., 2008) 或基于困惑度的望远镜 (Hans et al., 2024), 假设由大型语言模型 (LLM) 生成的文本表现出统计特征 (例如, 较低的熵、独特的困惑度), 这些特征可以与人类撰写的文本区分开。

然而, 当应用于专业领域的语料库时 (见图 1), 这些信号常常会失效: 对于一般领域的

LLM 来说，人工生成和机器生成的文本可能显得同样“令人惊讶”或“熟悉”。为了形式化这个现象，我们首先分析一个基于检测器 LLM 的可能性 $p_Q(X)$ 对文本 X 进行评分的一般零样本检测器的理论性能。具体来说，我们提出了以下定理，该定理描述了检测器性能与两个特定的 Kullback-Leibler (KL) 散度之间的关系。

Theorem 1. 给定一个零样本 LLM 生成的文本检测器 D ，其检测分数由任何严格单调函数 $f: \mathbb{R} \rightarrow \mathbb{R}$ 映射自检测器 LLM 的文本似然性 $p_Q(X)$ ：

$$D(X) = f(p_Q(X)), \quad (1)$$

，其中 Q 表示从 D 计算出的文本分布。 D 的接收者操作特征曲线下的面积 (AUROC)，记作 $\text{AUROC}(D(X))$ ，满足：

$$\text{AUROC}(D(X)) \propto^+ \delta_{KL}, \quad (2)$$

，其中 $\propto^+$ 表示正单调相关， δ_{KL} ，有效的 KL 散度，给出为：

$$\delta_{KL} = D_{KL}(P \parallel Q') - D_{KL}(Q \parallel Q'), \quad (3)$$

，其中 Q' 表示真实来源的文本分布。 P 表示来自人类的文本分布。 $D_{KL}(\cdot \parallel \cdot)$ 表示 KL 散度。

定理 1 显示，无需示例的探测器的性能取决于人类文本分布 P 相较于真实来源分布 Q' 的偏离程度是否大于与探测器分布 Q 的偏离。在具有独特语言特征的专业领域（例如医学或法律）中，来源分布 Q' 可能更接近于 P 而不是通用领域的 LLM 分布 Q 。因此， δ_{KL} 的减少会降低探测器在这些领域中区分机器生成文本的能力。详见附录 ?? 中的详细定理证明。

在理论分析的基础上，我们设计了一个探测器，利用领域适应的 LLM $\mathcal{M}^*$ 的熵（见小节 3.2）来量化一个句子 $x_{1:L}$ 在其分布下的可预测性或熟悉度。给定一个 $x_{1:L} = (x_1, \dots, x_L)$ ，平均 token 熵定义为：理想情况下，如果 $\mathcal{M}^*$ 与指定领域的 LLM 生成文本的分布良好对齐，那么 LLM 生成的文本应当比人工撰写的文本产生更低的熵。然而，在实际中，由于细微的领域不匹配或者存在高度可预测（即低熵）的人工文本，这一假设不总是成立。

为了解决这一问题，我们引入了一种基于通用 LLM $\mathcal{M}$ 与领域自适应 LLM $\mathcal{M}^*$ 之间的交叉熵的归一化方案。交叉熵定义为：

$$CE_{\mathcal{M}, \mathcal{M}^*}(x_{1:L}) = -\frac{1}{L} \sum_{i=1}^L p_{\mathcal{M}}(x_i \mid x_{<i}) \cdot \log p_{\mathcal{M}^*}(x_i \mid x_{<i}) \quad (4)$$

上述交叉熵量化了两个模型对于文本的概率估计之间的分歧程度。对于在专业领域中由 LLM 生成的文本，预期领域自适应的 $\mathcal{M}^*$ 将会明显偏离通用 LLM $\mathcal{M}$ ，从而导致较高的交叉熵得分。相反，对于对两个模型都不熟悉的人类编写的文本，分歧通常较小，从而导致较低的交叉熵值。这一直觉在第 ?? 节的详细统计分析中得到了进一步支持。最终的 DivScore 计算如下：

$$D_{\mathcal{M}, \mathcal{M}^*}(x_{1:L}) = \frac{H_{\mathcal{M}^*}(x_{1:L})}{CE_{\mathcal{M}, \mathcal{M}^*}(x_{1:L})} \quad (5)$$

由高熵（不熟悉度）和低交叉熵（低模型分歧）驱动的较高 $D_{\mathcal{M}, \mathcal{M}^*}(x_{1:L})$ 表明 $x_{1:L}$ 更有可能是人工撰写的。相反，较低的得分则表明文本是机器生成的。这种 DivScore 归一化增强了人类和 LLM 生成文本之间的分离，特别是在仅靠原始熵信号不足的专业领域中。

3.2 通过无监督领域知识蒸馏增强检测器

DivScore 使用一对 LLM ($\mathcal{M}, \mathcal{M}^*$) 来评分，其中包括一个通用和领域适应的模型。在基本的 DivScore 设置中， $\mathcal{M}^*$ 可以是针对目标领域的任何 LLM。此外，由于定理 1 的启发，并旨在提高在高度专门化领域的检测性能，我们引入了一个量身定制的改进：通过将领域特定知识提炼到通用模型 $\mathcal{M}$ 中来构建 $\mathcal{M}^*$ 。这一改进的目标是通过使检测器的分布与领域特定 LLM 生成的文本对齐来减少 KL 散度 $D_{KL}(Q \parallel Q')$ 。过程如下：

1. 教师模型选择：选择一个具有广泛领域专业知识和大量参数的大型语言模型 (LLM) 作为教师模型（例如，DeepSeek-R1）。
2. 领域种子知识构建：收集一组代表该专业领域的问题。利用种子知识，提示教师 LLM 基于初始输入生成详细和深入的问题。
3. 蒸馏知识生成：使用教师 LLM 生成种子问题的高质量答案和叙述性解释，形成由教师 LLM 输出构成的领域特定的蒸馏知识库。
4. 学生学习：使用生成的特定领域问答对微调通用 LLM $\mathcal{M}$ 的 DivScore，生成增强后的模型 $\mathcal{M}^*$ 以适应这个专业领域。

用于构建 $\mathcal{M}^*$ 的这种蒸馏过程将领域知识从教师 LLM 转移到 DivScore 的一般 LLM $\mathcal{M}$ ，缩小了检测器和 LLM 文本分布之间的差距，从而增加了定理 1 中的 δ_{KL} ，提高了检测性

能。整个方法在零样本的方式下运行，不需要人为标注的文本用于训练和推断，使其在专业领域的实际部署中极具实用性。

我们将上述所有资源作为一个基准发布，其中包括 (1) 医疗和法律领域的专门知识蒸馏数据集和适应流程，以及 (2) 4,000 份高质量的专业法律和医疗文档，并配有用于检测器评估的 LLM 生成对应部分。

在本节中，我们评估了 DivScore 在检测法律和医学领域机器生成文本方面的有效性。我们进一步评估了其试图绕过检测的对抗性攻击的鲁棒性。此外，我们分析了领域知识蒸馏对文本评分的影响，并通过消融研究检验了检测器 LLM 选择、知识蒸馏和评分组件对整体性能的影响。

3.3 实现细节

本研究调查了零样本黑箱检测设置，这些设置现实地反映了实际部署场景。在零样本情况下，检测器必须识别 LLM 生成的文本，而没有任何任务特定的标记示例。黑箱约束进一步限制了关于源 LLM 及其 logits 的信息，因此检测器必须依赖替代模型进行评分。这些约束共同反映了实际情况，其中既没有领域内参考数据，也没有生成 LLM 的身份信息可用。

DivScore 检测器 DivScore 使用 Mistral-7B-v0.2 (Jiang et al., 2023) 作为文本评分和领域适应的一般大模型。领域适应的大模型通过领域知识蒸馏的过程开发，通过在 Mistral-7B-v0.2 上微调进行。在医学和法律领域同时进行的综合知识蒸馏优化了检测器在这些专业领域内的能力。在主实验中，我们还评估了两种变体的 DivScore，它们仅在医学或法律领域进行了部分知识蒸馏，分别称为 DivScore (med) 和 DivScore (law)。

领域知识蒸馏

我们选择 DeepSeek-R1 (DeepSeek-AI, 2025) 作为教师大型语言模型，由于其强大的医疗和法律专业能力 (Tordjman et al., 2025; Yu et al., 2025) 以及其宽松的 MIT 许可证，这允许进行全面的知识蒸馏。医学种子知识来源于 22,000 个由 GPT-4o 生成的、可验证的医学问题 Chen et al. (2024)。法律种子知识来源于：(1) EQUALS (Chen et al., 2023)，包含 6,909 个来自专业法律咨询的法律问题；(2) Open Legal QA (Butler, 2023)，包含 2,124 个基于澳大利亚法律案件的 GPT-4 合成问题；及 (3) Pile-of-Law (Henderson et al., 2022)，包含 15,393 个来自 Reddit 论坛的法律问题。蒸馏知识随后由 DeepSeek-R1 以问答的形式生成。生成的输出用于通过低秩适配 (LoRA) 来微调探测器的一般大型语言模型。关于种子知识构建和 LoRA

微调的细节在附录 5.1 中提供。

数据集我们选择了四个医疗和法律领域的代表性数据集来评估 DivScore 和基线：(1) MIMIC-IV-Note (MIMIC) (Johnson et al., 2023)，患者的去识别化医院记录，(2) PubMedQA (Jin et al., 2019)，从 PubMed 收集的生物医学研究论文的问答，(3) Open Australia Legal Corpus (OALC) (Butler, 2023)，澳大利亚立法和司法文件的集合，(4) LawStack (Moslem, 2025)，从 Stack Exchange 收集的法律问答。在每个数据集中，我们选择了 1000 对文本，每对由人工撰写的文本及其 LLM 生成的对比文本组成。用于生成 LLM 文本的源 LLM 选自两个模型系列：GPT-4o (OpenAI, 2024)、O3-mini (OpenAI, 2025)、DeepSeek-V3 (DeepSeek-AI, 2024) 和 DeepSeek-R1 (DeepSeek-AI, 2025)，代表了当前面向公众的 SOTA LLM。在附录 ?? 中提供了详细设置。

基线检测器我们对 DivScore 与一套全面的零样本和有监督检测方法进行了基准测试。对于零样本检测器，我们包括了 Log-Likelihood (Solaiman et al., 2019)，Entropy (Lavergne et al., 2008)，Rank, Log-Rank (Gehrmann et al., 2019)，DetectGPT (Mitchell et al., 2023)，Fast-DetectGPT (Bao et al., 2023)，和 Binoculars (Hans et al., 2024)。对于有监督检测器，我们评估了 OpenAI (Solaiman et al., 2019) 发布的两个基于 RoBERTa 的分类器 (RoB-base 和 RoB-large)。在零样本实验设置中，有监督检测器在当前领域检测任务中未经过进一步训练就被评估。详细的基线定义见附录 ??。

评价指标先前的研究 (Su et al., 2023; Yu et al., 2024; Wu et al., 2025b) 主要使用 AUROC 来评估检测器的性能。然而，作为一个汇总指标，AUROC 可能无法反映在低误报率情况下的性能，这在误将人类撰写的文本判定为 LLM 生成的文本会造成很大危害的特定领域中至关重要。因此，除了 AUROC，我们还报告了在误报率 (FPR) 为 0.1 % 时的真正率 (TPR)，这符合 Hans et al. (2024) 的评估协议。这个指标反映了在 1,000 个人工撰写的文本中，至多允许一个误报时的 TPR。

表 1 展示了在由 GPT-4o、O3-mini、DeepSeek-V3 和 DeepSeek-R1 分别生成的四个医疗和法律数据集上，基准方法和 DivScore 的检测性能。如所示，DivScore 及其部分知识蒸馏的变体——DivScore (med) 和 DivScore (law)——在 Binoculars 基准上表现出显著的性能提升。对于 DivScore (med)，平均而言，它在 AUROC 上获得 14.3% 的提升，在 TPR 上获得 62.1 % 的提升。对于 DivScore (law)，它取得了类似的性能，平

Text Source Datasets	GPT-4o				O3-mini				DeepSeek-V3				DeepSeek-R1			
	MIMIC	PubMed	OALC	LawStack	MIMIC	PubMed	OALC	LawStack	MIMIC	PubMed	OALC	LawStack	MIMIC	PubMed	OALC	LawStack
AUROC																
RoB-base	66.79 _(31.55)	85.79 _(39.63)	58.33 _(19.86)	65.10 _(29.98)	67.98 _(26.71)	98.70 _(14.13)	58.07 _(10.67)	73.35 _(9.02)	69.11 _(30.89)	82.39 _(17.33)	23.20 _(175.07)	56.54 _(43.33)	68.99 _(20.11)	99.63 _(9.46)	65.24 _(10.69)	74.73 _(13.16)
RoB-large	93.05 _(5.29)	91.69 _(3.73)	81.23 _(42.76)	72.36 _(22.72)	94.21 _(8.48)	99.66 _(15.09)	77.29 _(19.89)	85.55 _(3.18)	91.10 _(8.39)	92.14 _(7.58)	52.31 _(45.96)	67.82 _(32.05)	87.35 _(1.75)	99.84 _(9.67)	84.22 _(29.67)	86.36 _(11.53)
Entropy	68.52 _(29.82)	98.29 _(12.87)	33.30 _(5.17)	91.85 _(3.21)	84.73 _(9.96)	96.95 _(12.88)	38.23 _(19.17)	83.53 _(1.16)	94.40 _(5.46)	99.87 _(10.15)	87.14 _(11.13)	98.22 _(1.65)	68.21 _(20.89)	99.39 _(9.22)	58.92 _(14.37)	94.70 _(9.81)
Rank	71.97 _(26.37)	87.54 _(7.88)	29.97 _(8.5)	85.73 _(3.35)	98.66 _(13.97)	96.36 _(11.79)	58.52 _(11.12)	92.58 _(10.21)	86.86 _(13.14)	94.56 _(15.16)	68.35 _(29.92)	93.33 _(6.54)	92.34 _(13.24)	92.49 _(12.32)	65.19 _(10.64)	87.16 _(9.73)
Log-Rank	83.37 _(14.97)	98.75 _(13.33)	32.40 _(6.07)	91.19 _(3.89)	95.07 _(9.28)	98.89 _(14.32)	47.05 _(10.35)	88.34 _(3.97)	99.57 _(10.43)	99.96 _(10.24)	94.09 _(14.18)	99.03 _(0.84)	87.92 _(11.18)	99.31 _(9.14)	67.70 _(13.15)	95.02 _(7.13)
Log-Likelihood	81.29 _(17.05)	98.30 _(12.33)	32.32 _(6.15)	90.15 _(4.93)	94.08 _(10.61)	98.20 _(13.63)	47.32 _(10.08)	86.77 _(4.4)	99.37 _(10.63)	99.94 _(10.22)	93.79 _(14.48)	98.70 _(11.17)	88.18 _(10.92)	99.14 _(9.97)	70.36 _(15.81)	94.73 _(9.84)
DetectGPT	53.71 _(44.63)	42.39 _(53.03)	31.05 _(7.42)	57.51 _(37.57)	63.15 _(21.54)	54.10 _(30.47)	32.55 _(24.85)	86.05 _(3.68)	78.68 _(21.32)	73.45 _(26.27)	47.36 _(130.91)	89.30 _(10.57)	87.74 _(11.96)	80.95 _(19.22)	43.36 _(11.19)	97.07 _(9.18)
Fast-DetectGPT	96.23 _(12.11)	87.79 _(17.63)	46.32 _(7.85)	69.19 _(25.9)	90.38 _(14.31)	77.55 _(7.02)	71.91 _(14.51)	58.81 _(23.56)	99.89 _(10.11)	99.65 _(10.07)	98.22 _(10.05)	98.50 _(11.28)	89.63 _(10.53)	74.86 _(15.31)	65.24 _(10.9)	53.50 _(34.3)
Binoculars	98.34	95.42	38.47	95.08	94.69	84.57	57.40	82.37	100.00	99.72	98.27	99.87	89.10	90.17	54.55	87.89
DivScore (med)	<u>99.86</u> (11.32)	99.07 (13.65)	99.17 (16.07)	99.27 (14.19)	<u>99.92</u> (15.3)	<u>99.87</u> (115.3)	99.45 (142.05)	<u>99.76</u> (117.39)	99.97 (10.03)	<u>99.87</u> (10.15)	98.51 (10.24)	99.93 (10.06)	<u>99.98</u> (110.88)	<u>99.98</u> (19.81)	99.28 (144.73)	<u>99.99</u> (112.1)
DivScore (law)	<u>99.86</u> (11.32)	<u>99.61</u> (14.46)	98.65 (100.18)	<u>99.54</u> (14.46)	<u>99.92</u> (15.3)	99.81 (115.24)	<u>99.25</u> (141.85)	99.87 (117.5)	99.97 (10.03)	99.81 (10.09)	98.61 (10.34)	<u>99.94</u> (10.07)	99.94 (10.84)	99.93 (19.76)	<u>99.37</u> (144.82)	<u>99.92</u> (112.1)
DivScore	<u>99.87</u> (11.33)	<u>99.90</u> (14.48)	<u>98.88</u> (10.41)	<u>99.84</u> (14.76)	100.00 (15.31)	<u>99.96</u> (15.39)	99.20 (141.8)	<u>99.87</u> (117.5)	<u>99.99</u> (10.01)	<u>99.99</u> (10.27)	<u>98.54</u> (10.27)	<u>99.99</u> (10.12)	<u>99.99</u> (110.89)	100.00 (19.81)	<u>99.56</u> (145.01)	100.00 (112.11)
TPR @ 0.1 % FPR																
RoB-base	0.00 _(0.68)	35.40 _(10.3)	0.10 _(0.4)	0.00 _(0.27)	0.00 _(14.3)	62.60 _(762.6)	0.10 _(0.1)	0.10 _(0.3)	0.00 _(0.97)	36.90 _(11.8)	0.00 _(174.5)	0.20 _(0.84)	0.04 _(0.36)	59.20 _(19.2)	0.20 _(70.2)	0.10 _(0.2)
RoB-large	9.80 _(15.6)	26.70 _(12.1)	1.90 _(1.4)	0.00 _(0.27)	5.08 _(10.22)	78.80 _(178.8)	0.90 _(0.7)	0.30 _(0.1)	0.94 _(0.76)	23.20 _(11.9)	0.10 _(174.4)	0.30 _(0.84)	0.62 _(1.68)	82.30 _(192.3)	2.80 _(2.8)	1.30 _(1.0)
Entropy	0.00 _(0.68)	28.80 _(12.37)	0.10 _(0.4)	0.00 _(0.27)	0.60 _(14.17)	10.20 _(10.2)	0.00 _(0.2)	0.00 _(0.4)	5.90 _(0.89)	89.20 _(154.1)	4.40 _(170.1)	0.00 _(0.65)	0.00 _(0.3)	3.70 _(17.1)	0.10 _(0.1)	0.00 _(0.3)
Rank	13.50 _(15.49)	9.80 _(14.7)	0.00 _(0.5)	5.10 _(12.6)	34.30 _(120.6)	14.20 _(114.2)	0.00 _(0.2)	5.30 _(1.9)	33.30 _(166.4)	28.90 _(14.2)	2.80 _(171.7)	19.50 _(165.3)	0.40 _(0.4)	3.20 _(1.2)	0.00 _(0.0)	0.40 _(0.1)
Log-Rank	0.90 _(0.75)	59.90 _(154.8)	0.10 _(0.4)	0.00 _(0.27)	4.30 _(10.69)	48.70 _(148.7)	0.00 _(0.2)	0.00 _(0.4)	66.30 _(133.4)	98.00 _(162.9)	4.40 _(170.1)	0.00 _(0.65)	1.10 _(1.7)	63.20 _(193.2)	0.10 _(0.1)	0.00 _(0.3)
Log-Likelihood	0.00 _(0.68)	42.00 _(156.9)	0.10 _(0.4)	0.00 _(0.27)	1.50 _(12.28)	23.40 _(123.4)	0.00 _(0.2)	0.00 _(0.4)	40.20 _(159.5)	96.90 _(162.8)	4.40 _(170.1)	0.00 _(0.65)	0.30 _(0.4)	48.60 _(186.6)	0.10 _(0.1)	0.00 _(0.3)
DetectGPT	0.00 _(0.68)	0.00 _(12.1)	0.00 _(0.5)	0.00 _(0.27)	0.00 _(14.13)	0.00 _(0.0)	0.00 _(0.2)	0.20 _(0.2)	0.00 _(0.97)	0.00 _(125.1)	0.00 _(174.5)	0.10 _(0.4)	0.00 _(0.3)	0.00 _(0.0)	0.00 _(0.0)	2.70 _(12.4)
Fast-DetectGPT	42.20 _(126.2)	7.30 _(12.1)	0.90 _(0.4)	1.70 _(26.0)	12.50 _(1.8)	1.50 _(11.5)	0.80 _(0.6)	0.00 _(0.4)	92.70 _(17.0)	82.20 _(147.1)	73.80 _(16.7)	60.90 _(124.1)	7.70 _(0.36)	1.60 _(1.6)	0.50 _(0.5)	0.00 _(0.3)
Binoculars	68.40	5.10	0.50	27.70	14.30	0.00	0.20	0.40	99.70	35.10	74.50	85.00	8.30	0.00	0.00	0.30
DivScore (med)	94.10 (125.7)	<u>71.80</u> (106.7)	81.00 (180.5)	72.00 (144.3)	<u>99.20</u> (184.9)	<u>95.90</u> (195.9)	71.10 (170.9)	83.70 (183.3)	98.30 (11.4)	<u>91.10</u> (156.0)	77.00 (12.5)	94.70 (19.7)	<u>97.60</u> (189.3)	<u>98.70</u> (198.7)	88.30 (188.3)	99.20 (198.9)
DivScore (law)	<u>91.20</u> (122.8)	51.40 (146.3)	66.00 (166.1)	<u>80.20</u> (152.5)	97.80 (183.5)	83.90 (183.9)	<u>48.10</u> (147.5)	<u>93.50</u> (193.1)	97.70 (12.0)	76.20 (141.1)	78.00 (13.5)	<u>97.30</u> (112.8)	94.20 (185.9)	92.60 (192.6)	<u>89.70</u> (198.7)	<u>99.70</u> (199.4)
DivScore	90.40 (122.0)	94.50 (189.4)	<u>69.60</u> (169.1)	88.00 (160.3)	<u>99.60</u> (185.3)	99.50 (199.5)	42.80 (142.6)	96.50 (196.1)	<u>98.80</u> (10.9)	99.10 (164.0)	<u>77.30</u> (12.8)	98.10 (113.1)	98.80 (190.5)	99.90 (199.9)	89.90 (199.9)	99.90 (199.6)

Table 1: 主要实验——基线方法和 DivScore 在医学 (MIMIC, PubMedQA) 和法律数据集 (OALC, LawStack) 上的检测 AUROC 分数和 TPR (在 0.1 % FPR 处)。方法与 Binoculars 检测器进行基准测试，箭头表明相对于 Binoculars 的性能差异 (/)。加粗表示最佳结果，underline 表示第二佳。

均在 AUROC 上获得 14.3 % 的提升，在 TPR 上获得 57.4 % 的提升。通过从医疗和法律领域的蒸馏增强的完整 DivScore 进一步将平均改进推进到 AUROC 上的 14.4 % 和 TPR 上的 64.0 %。这些结果强调了 DivScore 在检测专业 LLM 生成内容时具有少量误报的能力。此外，DivScore 相较于其特定领域变体的改进表明其跨不同专业领域的广泛适用性，显示出其在跨领域检测专业 LLM 生成文本的潜力。附录 6.1 和图 5 对所有检测器的检测性能以及真实与误报的权衡进行了更详细的分析。

3.4 检测器鲁棒性

最近的研究表明，改写 (Sadasivan et al., 2023; Krishna et al., 2023) 和词级扰动 (Peng et al., 2023; Liang et al., 2023) 可以通过改变词汇和句法特征有效规避当前的 AI 文本检测器。为了评估 DivScore 在现实应用中的鲁棒性，我们对 GPT-4o 文本进行改写 (Krishna et al., 2023) 和词级替换 Peng et al. (2023) 攻击。正如表 ?? 所示，这两种攻击大大降低了零样本和监督基线的性能。相比之下，DivScore 保持了鲁棒性，平均在 AUROC 上比 Binoculars 高出 22.8 %，在 TPR 上高出 29.5 %。这种鲁棒性可能是由于提出的知识蒸馏，它将检测器与特定领域的 LLM 知识对齐。由于对抗性攻击必须保留语义和领域特定内容以确保可用性，DivScore 仍然能够有效识别 LLM 生成的文本，即使在强扰动下也能保持高检测精度。详细设置请见附录 ??。

在本节中，我们研究领域知识蒸馏如何影响 DivScore 使用的统计特征。我们分析了 DivS-

Mean $\pm$ Std	Mean Token Cross-Entropy		Mean Token Entropy		Detector Score	
	LLM	Human	LLM	Human	LLM	Human
w/o Distillation						
MIMIC	1.81 $\pm$ 0.17	1.97 $\pm$ 0.21	0.98 $\pm$ 0.11	1.24 $\pm$ 0.15	0.54 $\pm$ 0.04	0.63 $\pm$ 0.04
PubMed	1.70 $\pm$ 0.25	2.74 $\pm$ 0.47	0.99 $\pm$ 0.17	2.07 $\pm$ 0.40	0.58 $\pm$ 0.07	0.75 $\pm$ 0.04
OALC	1.81 $\pm$ 0.30	1.80 $\pm$ 0.38	1.12 $\pm$ 0.22	1.35 $\pm$ 0.29	0.61 $\pm$ 0.06	0.75 $\pm$ 0.05
LawStack	1.87 $\pm$ 0.21	2.42 $\pm$ 0.37	1.13 $\pm$ 0.14	1.72 $\pm$ 0.31	0.61 $\pm$ 0.05	0.71 $\pm$ 0.07
Avg.	1.79 $\pm$ 0.23	2.23 $\pm$ 0.36	1.05 $\pm$ 0.16	1.60 $\pm$ 0.29	0.59 $\pm$ 0.06	0.71 $\pm$ 0.05
w/ Distillation						
MIMIC	3.56 $\pm$ 0.29	2.80 $\pm$ 0.30	0.46 $\pm$ 0.07	0.88 $\pm$ 0.13	0.13 $\pm$ 0.03	0.31 $\pm$ 0.03
PubMed	4.56 $\pm$ 0.16	4.42 $\pm$ 0.14	0.37 $\pm$ 0.08	1.21 $\pm$ 0.32	0.08 $\pm$ 0.02	0.28 $\pm$ 0.04
OALC	4.03 $\pm$ 0.37	3.04 $\pm$ 0.31	0.44 $\pm$ 0.09	0.68 $\pm$ 0.18	0.11 $\pm$ 0.02	0.23 $\pm$ 0.04
LawStack	5.19 $\pm$ 0.27	4.32 $\pm$ 0.32	0.36 $\pm$ 0.06	0.90 $\pm$ 0.20	0.07 $\pm$ 0.01	0.21 $\pm$ 0.05
Avg.	4.33 $\pm$ 0.27	3.64 $\pm$ 0.27	0.41 $\pm$ 0.07	0.92 $\pm$ 0.21	0.10 $\pm$ 0.02	0.26 $\pm$ 0.03

Table 2: DivScore 及其组成部分的统计数据 (平均值 $\pm$ 标准差)：针对每个数据集的交叉熵、熵。统计结果是在四个文本来源的 LLM 上计算的，有和没有领域知识蒸馏的情况下。"Avg." 行显示数据集平均值。

core 及其组成部分的统计数据——均值符号熵和带有或不带蒸馏的均值符号交叉熵。仅在一般指令上微调的 Mistral-7B-v0.2-Instruct 作为“无蒸馏”基线的 $\mathcal{M}^*$ 。正如图 ?? 所示，知识蒸馏增强了大语言模型生成文本和人类编写文本之间的可分性，这对于精确检测是必不可少的。表 2 还报告了这两种文本类型的特征的均值和标准差，并给出了以下发现。

(1) 交叉熵 & 熵。知识蒸馏减少了熵的标准差 (LLM 文本减少了 56.3 %，人工文本减少了 27.6 %)，使得两个簇更加紧凑。它还增加了交叉熵的质心 (平均值) 距离 36.2 %，扩大了 LLM 与人工文本簇之间的差距。这些结果如表 2 所示，证实了知识蒸馏可以改善 LLM 和人工生成文本的可分性。交叉熵统计进一步支持了第 ?? 节中提出的假设，即由于领域适应的 LLM 与通用领域 LLM 的分歧，LLM 生成的文本产生更高的交叉熵。

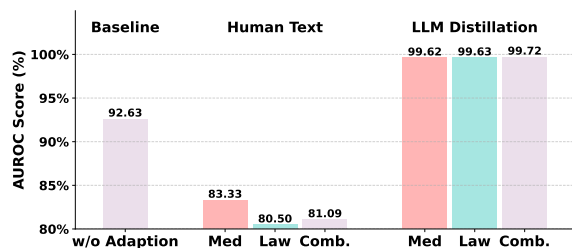


Figure 3: 在三种设置下的检测器 AUROC 分数比较：基线（无域适应）、人工域文本微调 and LLM 域知识蒸馏。对于后两种设置，包含三种变化：适应医疗领域、法律领域及其组合。AUROC 得分在所有评估数据集中取平均值。

(2) 检测器评分。所提出的检测器评分是上述统计特征的比率，因知识蒸馏而产生的集体影响。如表 2 所示，蒸馏导致组间中心（平均）距离整体增加 33.3%，LLM 评分标准差减少 66.7%，人类评分标准差减少 40%。这些结果表明，领域知识蒸馏通过增加 LLM 与人类文本间的差距（中心距离）和减少组内变异（标准差），增强了 LLM 生成文本的可区分性。

3.5 消融研究

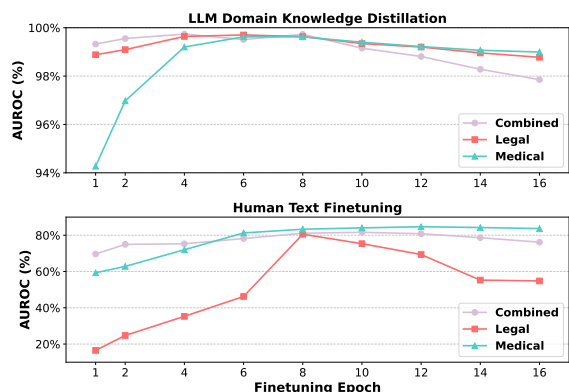


Figure 4: 检测器性能（AUROC %）在整个微调纪元中，LLM 蒸馏（提出的方法）与人类文本微调的比较。AUROC 分数在所有评估数据集上取平均值。

领域知识蒸馏的消融。在第 3.2 节中，我们分析了引入 LLM 领域知识蒸馏以提高 DivScore 性能的动机。我们进一步实证评估其对检测的影响。具体来说，我们将 DivScore 与以下内容进行比较：(1) 三种变体，其中对一般 LLM $\mathcal{M}$ 的领域适应是通过医疗、法律领域及其组合的人类撰写文本进行微调构建的，(2) 在第 ?? 节中提到的没有领域适应的基线 DivScore。

图 3 展示了所有数据集和文本来源 LLM 的 AUROC 平均值。结合 LLM 领域知识蒸馏显著提升了检测性能，与从人类文本语料库和基线适配的 DivScore 变体相比，分别实现了

Methods	Text Source LLMs				
	MIMIC	PubMedQA	OALC	LawStack	Avg
Entropy (Mistral)	99.32	99.78	85.43	99.27	95.95
Cross-Entropy (Mistral)	12.69	83.17	21.91	52.45	42.56
DivScore (Mistral)	99.96	99.96	99.04	99.93	99.72
Entropy (Falcon)	94.69	99.49	89.34	98.12	95.41
Cross-Entropy (Falcon)	25.37	73.19	54.26	46.58	49.85
DivScore (Falcon)	99.70	99.13	95.87	99.13	98.46
Entropy (Qwen)	98.81	99.88	89.03	99.38	96.78
Cross-Entropy (Qwen)	20.86	94.97	19.73	61.36	49.23
DivScore (Qwen)	99.96	94.43	97.45	99.33	97.79
Entropy (Llama)	98.01	99.89	86.75	99.12	95.94
Cross-Entropy (Llama)	10.31	64.55	17.10	32.23	31.05
DivScore (Llama)	99.91	99.87	99.03	99.86	99.67

Table 3: 消融结果（AUROC %）展示了 DivScore 及其评分变体在四种检测器 LLM 架构上的表现，表明所提出的设计始终优于单独使用熵或交叉熵。AUROC 得分是对所有文本来源 LLM 平均得出的。

18.6% 和 7.1% 的 AUROC 提升。这些发现表明，LLM 蒸馏对于构建适应领域的 LLM 以用于 DivScore 非常有效，优于基线和人类文本微调。我们进一步研究了微调周期数对各变体性能的影响。如图 4 所示，所有变体在第 8 个周期达到峰值或接近峰值 AUROC，之后性能下降。这一观察支持了我们对微调周期的选择，并确保消融研究中的公平比较。直观上，不足的训练导致领域适应不足，而过多微调可能导致 LLM 在通用领域的语言特性遗忘，降低检测效果。

检测器组件的消融。我们进一步验证了 DivScore 在两个关键检测器组件上的设计：LLM 架构的选择及其评分机制的有效性。我们比较了四个开源的 LLM: Falcon-7B (Almazrouei et al., 2023)、Qwen2.5-8B (Qwen, 2025)、Llama3-8B (Grattafiori et al., 2024) 以及我们提出的 LLM, Mistral-7B-v0.2。我们还通过比较总体检测器得分及其组件（标记熵和标记交叉熵）来评估每个评分组件。在所有数据集和源 LLM 上（表 3）平均计算的检测器性能得出了三个主要发现：(1) DivScore 在所有 LLM 架构上实现了稳健的检测（平均 AUROC > 95），其中 Mistral-7B 表现最佳；(2) 完整的检测器评分机制优于其组件部分，证实了 DivScore 设计的必要性。详细的消融设定见附录 ??。

在这项工作中，我们引入了 DivScore，这是一种理论上有依据的零次检测器，用于识别医学和法律等专业及高风险领域的 LLM 生成文本。我们的分析表明，现有的通用零次检测器在这些环境中表现不佳，这是由于分布上的不匹配。为了解决这个问题，DivScore 利用了一种基于归一化熵的评分机制，并通过领域知识蒸馏进行增强，以稳健地检测 LLM 生成的专业内容。在医学和法律领域的实证结果显示，

DivScore 在面对对抗性攻击时始终优于现有的 SOTA 基线。

4 限制

我们识别出以下工作中的局限性：(1) 领域适应性需求。我们的方法需要针对特定领域进行定制的适应。虽然所提出的检测器在超越单一领域方面体现了泛化性，但有效的适应对于达到高检测性能仍然至关重要。这个适应过程可能带来实际挑战，尤其是在低资源环境中。未来的工作可以集中于提高领域适应的效率和可扩展性。(2) 语言范围。与大多数现有研究一样，我们的工作仅着重于检测英文中的机器生成文本。DivScore 在其他语言，尤其是非英语和低资源语言中的表现尚未经过检验。扩展跨语言的检测能力——特别是在专业及高风险环境中——是未来研究的关键方向，包括全面的基准测试和适应策略。

References

- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, M  rouane Debbah,   tienne Goffinet, Daniel Hessel, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. *The falcon series of open language models*. *Preprint*, arXiv:2311.16867.
- Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc’Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. *GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow*. Version 1.0.
- Umar Butler. 2023. Open australian legal qa. <https://huggingface.co/datasets/isaacus/open-australian-legal-qa>. Version 2.0.0, Hugging Face, DOI: 10.57967/hf/1479. Accessed: 2025-05-09.
- Umar Butler. 2025. Open australian legal corpus. <https://huggingface.co/datasets/isaacus/open-australian-legal-corpus>. Accessed: 2025-05-09.
- Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders S  gaard. 2023. *LeXFiles* and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.
- Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. *Equals: A real-world dataset for legal question answering via reading chinese laws*. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL ’23*, page 71–80, New York, NY, USA. Association for Computing Machinery.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. *Huatuogpt-o1, towards medical complex reasoning with llms*. *Preprint*, arXiv:2412.18925.
- DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. <https://arxiv.org/abs/2412.19437>. Accessed: 2025-05-09.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://arxiv.org/abs/2501.12948>. Accessed: 2025-05-09.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. *Preprint*, arXiv:1810.04805.
- Liam Dugan, Alyssa Hwang, Filip Trhl  k, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. *RAID: A shared benchmark for robust evaluation of machine-generated text detectors*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. *Hierarchical neural story generation*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. *GLTR: Statistical detection and visualization of generated text*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, pages 17519–17537. PMLR.
- Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. *Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset*. Preprint, arXiv:2207.00220.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. *Pubmedqa: A dataset for biomedical research question answering*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.
- Thomas Lavergne, Tanguy Urvoy, and Fran  ois Yvon. 2008. Detecting fake content with relative entropy scoring. In *Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship and Social Software Misuse - Volume 377, PAN’08*, page 27–31, Aachen, DEU. CEUR-WS.org.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).
- Gongbo Liang, Jesus Guerrero, and Izzat Alsmadi. 2023. Mutation-based adversarial attacks on neural text detectors. *arXiv preprint arXiv:2302.05794*.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.
- Yasmin Moslem. 2025. Law-StackExchange (Revision 6a14705). <https://huggingface.co/datasets/ymoslem/Law-StackExchange>. Accessed: 2025-05-09.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. *Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2024. GPT-4o System Card. <https://arxiv.org/abs/2410.21276>. Accessed: 2025-05-09.
- OpenAI. 2025. OpenAI o3 and o4-mini System Card. <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>. Accessed: 2025-05-09.
- Xinlin Peng, Ying Zhou, Ben He, Le Sun, and Yingfei Sun. 2023. Hidding the ghostwriters: An adversarial evaluation of ai-generated student essay detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Qwen. 2025. Qwen2.5 technical report. <https://arxiv.org/abs/2412.15115>. Accessed: 2025-05-09.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. [DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, et al. 2025. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature Medicine*, pages 1–1.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. <https://github.com/kingoflolz/mesh-transformer-jax>. Accessed: 2025-04-20.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025a. [A survey on llm-generated text detection: Necessity, methods, and future directions](#). *Computational Linguistics*, 51(1):275–338.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xuebo Liu, Lidia S. Chao, and Min Zhang. 2025b. [Who wrote this? the key to zero-shot LLM-generated text detection is GECSScore](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10275–10292, Abu Dhabi, UAE. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Yashaswat. 2024. Indian legal text - abstractive summarization dataset. <https://huggingface.co/datasets/Yashaswat/Indian-Legal-Text-ABS>. Accessed: 2025-05-17.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. 2024. [Text fluoroscopy: Detecting LLM-generated text through intrinsic features](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15838–15846, Miami, Florida, USA. Association for Computational Linguistics.

Yaoyao Yu, Leilei Gan, Yinghao Hu, Bin Wei, Kun Kuang, and Fei Wu. 2025. Evaluating test-time scaling llms for legal reasoning: Openai o1, deepseek-r1, and beyond. *arXiv preprint arXiv:2503.16040*.

我们从标准的基于熵的检测器 d_H 开始分析，该检测器利用 LLM（具有生成分布 Q ）计算文本熵分数用于分类。给定输入的医学文本 X ，检测器的性能可以形式化表示如下：

Theorem 1. 给定一个零样本 LLM 生成的文本检测器 D ，其检测分数由检测器 LLM 的文本可能性 $p_Q(X)$ 通过任何严格单调函数 $f: \mathbb{R} \rightarrow \mathbb{R}$ 映射而来：

$$D(X) = f(p_Q(X)), \quad (6)$$

，其中 Q 表示从 D 计算的文本分布。 D 的性能，记为 $\text{AUROC}(D(X))$ ，满足：

$$\text{AUROC}(D(X)) \propto^+ \delta_{KL}, \quad (7)$$

，其中 $\propto^+$ 表示正单调关联，而 δ_{KL} ，即有效的 KL 散度，给出为：

$$\delta_{KL} = D_{KL}(P \parallel Q') - D_{KL}(Q \parallel Q'), \quad (8)$$

，其中 Q' 表示来自真实来源的文本分布。 P 表示来自人类的文本分布。 $D_{KL}(\cdot \parallel \cdot)$ 表示 *Kullback-Leibler* (KL) 散度。

Proof. 我们从将 LLM 熵检测器 $d_H(X)$ 视为 $D(X)$ 的一个特例开始证明。检测器 d_H 将抽样的文本 x 分类为机器生成，如果其在 Q 下的熵低于阈值 τ ：

$$d_H(x) = \begin{cases} 1 & \text{if } H_Q(x) < \tau \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

其中 $H_Q(x)$ 表示文本 x 在 LLM 分布 Q 的基础上从 X 抽样的 LLM 熵，由以下公式给出：

$$H_Q(x) = \sum_{x_i} p_Q(x_i | x_{<i}) \log p_Q(x_i | x_{<i}) \quad (10)$$

分布假设：假设文本变量 X 的 LLM 熵值遵循正态分布：

$$\begin{aligned} H_Q(X) &\sim \mathcal{N}(\mu_{Q'}, \sigma_{Q'}^2) & \text{if } X \sim Q' \\ H_Q(X) &\sim \mathcal{N}(\mu_P, \sigma_P^2) & \text{if } X \sim P \end{aligned} \quad (11)$$

其均值从检测器分布 Q 、人类分布 P 和 LLM 文本生成分布 Q' 之间的交叉熵中得出：

$$\begin{aligned} \mu_Q &= H(Q) + D_{KL}(Q \parallel Q') \\ \mu_P &= H(P) + D_{KL}(P \parallel Q') \end{aligned} \quad (12)$$

ROC 曲线构建：真阳性率 (TPR) 和假阳性率 (FPR) 定义为：

$$\begin{aligned} \text{TPR} &= \mathbb{P}_{X \sim Q'}(H_Q(X) < \tau) \\ \text{FPR} &= \mathbb{P}_{X \sim P}(H_Q(X) < \tau) \end{aligned} \quad (13)$$

给定方程 11 ,

$$\begin{aligned} \text{TPR}(\tau) &= \Phi\left(\frac{\tau - \mu_Q}{\sigma_{Q'}}\right) \\ \text{FPR}(\tau) &= \Phi\left(\frac{\tau - \mu_P}{\sigma_P}\right) \end{aligned} \quad (14)$$

其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数 (CDF)。

AUROC 衍生: AUROC 是 TPR 在 FPR 上的积分, 为简单起见将 FPR 表示为 p :

$$\text{AUROC} = \int_0^1 \Phi\left(\Phi^{-1}(p)\frac{\sigma_P}{\sigma_{Q'}} + \frac{\delta_{KL}}{\sigma_{Q'}}\right) dp \quad (15)$$

其中 $\delta_{KL} = \mu_P - \mu_Q = D_{\text{KL}}(P \parallel Q') - D_{\text{KL}}(Q \parallel Q')$, 给定方程 12 。

设 $z = \Phi^{-1}(p)$, 则 $dp = \phi(z)dz$, 其中 $\phi(z)$ 是标准正态分布的累积分布函数 (CDF)。方程 15 简化为:

$$\begin{aligned} \text{AUROC} &= \int_{-\infty}^{\infty} \Phi\left(\frac{\sigma_P}{\sigma_{Q'}}z + \frac{\delta_{KL}}{\sigma_{Q'}}\right) \phi(z)dz \\ &= \Phi\left(\frac{\delta_{KL}}{\sqrt{\sigma_P^2 + \sigma_{Q'}^2}}\right) \end{aligned} \quad (16)$$

由于正态分布的 CDF 是单调递增的, 所以对于检测器 d_H , 我们有:

$$\text{AUROC}(d_H(X)) \propto^+ \delta_{KL} \quad (17)$$

我们进一步将此结论扩展到 $D(X)$ 的完整定义。设 $s(x) = f(p_Q(x))$, 其中 $f: \mathbb{R} \rightarrow \mathbb{R}$ 是任意严格单调函数。对于任意 $x_1, x_2 \in \mathcal{X}$:

$$H_Q(x_1) > H_Q(x_2) \iff \begin{cases} s(x_1) > s(x_2) & (f \nearrow) \\ s(x_1) < s(x_2) & (f \searrow) \end{cases} \quad (18)$$

方程 18 意味着 $s(\cdot)$ 保留/反转由 $H_Q(\cdot)$ 引起的确切排名。对于检测器 s 的任何阈值 τ_s , 定义等效 H_Q 阈值:

$$\tau_H := H_Q(f^{-1}(\tau_s)) \quad (19)$$

假设 $f \nearrow$, 对于任何 τ_s :

$$\begin{aligned} \text{TPR}(\tau_s) &= \mathbb{P}_{X \sim Q'}(s(X) < \tau_s) \\ \text{FPR}(\tau_s) &= \mathbb{P}_{X \sim P}(s(X) < \tau_s) \end{aligned} \quad (20)$$

根据方程 19:

$$\begin{aligned} \text{TPR}(\tau_s) &= \mathbb{P}_{X \sim Q'}(H_Q(X) < \tau_H) \\ \text{FPR}(\tau_s) &= \mathbb{P}_{X \sim P}(H_Q(X) < \tau_H) \end{aligned} \quad (21)$$

同样, 对于 $f \searrow$, 对于任何 τ_s : 因此在任一设置中, 探测器 $s(X)$ 和 $H_Q(X)$ 的 ROC 曲线都是相同的, 得到:

$$\text{AUROC}(D(X)) = \text{AUROC}(d_H(X)) \quad (22)$$

因此, 我们有

$$\text{AUROC}(D(X)) \propto^+ \delta_{KL}. \quad (23)$$

□

5 实验设置

本节描述了我们的主要评估、对抗性攻击和消融研究的实验配置。表格 4 提供了我们测试中使用的大型语言模型的详细规格。所有实验均在配置了 8 个 NVIDIA H100-80G GPU 的 Linux 服务器上运行。

在这项研究中, 我们专注于检测书写英文中由 LLM 生成的文本。所有检测文本最初都是用英语撰写的。数据集以两种格式组织: (1) 文档: MIMIC (Johnson et al., 2023) 和 OALC (Butler, 2025) 数据集由独立的、人类撰写的医学和法律文档组成; (2) 问答对: PubMedQA (Jin et al., 2019) 和 LawStack (Moslem, 2025) 数据集中包含配对的、人类撰写的问题和答案。为了编译我们的测试集, 我们通过调用 API 服务为每个数据集和每个源模型 (GPT-4o、O3-mini、DeepSeek-V3、DeepSeek-R1) 生成 1000 个样本。

对于文档数据集, 我们通过使用每个文档的前三分之一的令牌作为提示, 并指示文本源 LLM 继续书写来生成相应的 LLM 生成文本。为了确保公平比较, 人类书写的文本也被截短, 仅保留原始文档的后三分之二。对于问答对数据集, 人类文本由人类编写的答案组成, 而 LLM 生成的文本则是通过将问题提示给文本源 LLM 生成的。具体来说, 我们通过向 LLM API 服务发送以下提示来启动文本生成过程, 如表格 5、表格 6 和表格 7 所示。

在本节中, 我们介绍在主要实验中使用的基准检测器及其配置。对于零样本检测器, 通过从检测器的代理语言模型 (及其变体) 中提取 logits 来获得分类得分。经典的零样本方法——包括 Log-Likelihood、Rank、Log-Rank 和 Entropy——使用 Mistral-7B-v0.2 以确保与 DivScore 检测器 LLM 设置一致。最近的零样本检测器, 如 DetectGPT、Fast-DetectGPT 和 Binoculars, 使用各自出版物中指定的原始 LLM 配置实现。对于有监督的检测器, 我们使用原作者发布的预训练模型, 并在主要实验中直接应用于检测分布外 (OOD) 文本。以下是基准方法的详细描述:

Model	Model File/Service	Parameters	Usage
GPT-4o (OpenAI, 2024)	OpenAI/gpt-4o	NA	Text Source
O3-mini (OpenAI, 2025)	OpenAI/O3-mini	NA	Text Source
DeepSeek-R1 (DeepSeek-AI, 2025)	deepseek-ai/DeepSeek-R1	671B (MoE)	Text Source
DeepSeek-V3 (DeepSeek-AI, 2024)	deepseek-ai/DeepSeek-V3	685B (MoE)	Text Source
DIPPER (Krishna et al., 2023)	kalpeshk2011/dipper-paraphraser-xxl	11B	Paraphrase Attack
BERT-base (Devlin et al., 2019)	google-bert/bert-base-uncased	110M	Word Substitution Attack
Llama3 (Grattafiori et al., 2024)	meta-llama/Meta-Llama-3-8B	8B	Detector Model
Mistral-v0.2 (Jiang et al., 2023)	mistral-community/Mistral-7B-v0.2	7B	Detector Model
Qwen2.5 (Qwen, 2025)	Qwen/Qwen2.5-7B	7B	Detector Model
Falcon (Almazrouei et al., 2023)	tiiuae/falcon-7b	7B	Detector Model
Falcon-Instruct (Almazrouei et al., 2023)	tiiuae/falcon-7b-Instruct	7B	Detector Model
T5 (Raffel et al., 2020)	google-t5/t5-3b	3B	Detector Model
GPT-Neo (Black et al., 2021)	EleutherAI/gpt-neo-2.7B	2.7B	Detector Model
GPT-J (Wang and Komatsuzaki, 2021)	EleutherAI/gpt-j-6B	6B	Detector Model
RoBERTa-base (Solaiman et al., 2019)	openai-community/roberta-base-openai-detector	110M	Detector Model
RoBERTa-large (Solaiman et al., 2019)	openai-community/roberta-large-openai-detector	340M	Detector Model

Table 4: 实验中使用的 LLM 的详细信息，包括机器生成文本的源模型、对抗攻击模型和检测器模型

```
[
  { "role": "system", "content": "Continue this hospital discharge summary." },
  { "role": "user", "content": "<prefix: 1/3 document content> " }
]
```

Table 5: MIMIC 数据集的提示模板

```
[
  { "role": "system", "content": "Continue to write this legal text." },
  { "role": "user", "content": "<prefix: 1/3 document content> " }
]
```

Table 6: OALC 数据集的提示模板

Log-Likelihood 这种零样本方法使用语言模型来计算给定文本中每个标记的 log 概率。具体而言，使用所有标记的平均 log 概率作为检测分数。较高的分数对应于文本由 LLM 生成的概率增加。

Rank 这种零样本方法利用语言模型确定的标记的绝对排名。通过计算平均排名得分，较高的平均值表明文本是机器生成的概率增加。

Log-Rank 是 Rank 方法的增强版本。这种零样本方法对每个标记的绝对排名应用对数转换。较高的平均对数排名值表明文本是机器生成的可能性更高。

熵这种零样本方法利用语言模型来计算

```
[
  { "role": "system", "content": "Answer this question truthfully:" },
  { "role": "user", "content": "<prefix: question> " }
]
```

Table 7: PubMedQA 和 LawStack 数据集的提示模板

给定文本中所有标记的平均熵分数 (Lavergne et al., 2008)。由于从 LLM 的角度来看，人类撰写的内容往往表现出更大的不可预测性，因此较低的平均熵分数表明文本更有可能是机器生成的。

DetectGPT 使用概率速率进行检测的零样本方法 (Mitchell et al., 2023)。该方法涉及扰动文本样本并分析预训练语言模型中对数似然的结果变化。其理论基础表明，LLM 生成的文本往往在 LLM 的对数概率空间中占据局部最优，使得扰动通常会降低其概率得分。相反，人类撰写的文本不遵循这种特性。在我们的实验中，DetectGPT 使用文献中的默认最优设置：GPT-Neo-2.7B (Black et al., 2021) 作为代理模型，T5-3B (Raffel et al., 2020) 作为掩码填充模型。

Fast-DetectGPT Bao et al. (2023) 提出了一种新的零样本方法来解决 DetectGPT 中高计算成本的问题。它采用了一种 LLM 采样方法来替代 DetectGPT 的扰动步骤。利用条件概率速率

[
{ "role": "system","content": "You are a medical professional, reply the medical question with professional knowledge." },
{ "role": "user", "content": "<prefix: question>. Answer:" }
]

Table 8: 医学初始知识的提示模板

[
"role": "system", "content": "You are a legal document translator.",
"role": "user", "content": "Translate this text into English: <question prefix>. Translation output:"
]

Table 10: EQUALS 法律问题翻译的提示模板

[
{ "role": "system","content": "You are a legal professional, reply the legal consult with professional knowledge." },
{ "role": "user", "content": "<prefix: question>. Answer:" }
]

Table 9: 法律种子知识的提示模板

作为度量, Fast-DetectGPT 识别 LLM 和人类文本之间的用词模式差异。我们选择了文献中报道的最优设置, 使用 GPT-Neo-2.7B 作为评分模型, GPT-J-6B (Wang and Komatsuzaki, 2021) 作为参考模型。

Binoculars Hans et al. (2024) 提出了一种零样本检测方法, 利用一对 LLM 计算文本的困惑度与其交叉困惑度之间的比率。这个困惑度比率衡量了一个模型对另一个的标记级预测的反应, 利用 LLM 和人类生成文本的系统差异。得分越低表示机器生成的可能性越高, 因为 LLM 生成的文本倾向于“相对于其在相同字符串上操作时的基准困惑度”的惊奇程度较低 (Hans et al., 2024)。在主要实验中, 我们选择了文献中报道的最优检测器设置, 使用 Flacon-7B (Almazrouei et al., 2023)、Flacon-7B-Instruct (Almazrouei et al., 2023) 作为观察者和执行者模型。

OpenAI RoBERTa Detectors 一组流行且具有竞争力的使用 RoBERTa 架构的监督检测器。Solaiman et al. (2019) 引入了一个标记数据集, 包括 GPT-2 生成的输出和 WebText 样本, 用于对 LLM 生成文本检测的监督训练。利用此数据集, 他们微调了基于 RoBERTa-base 和 RoBERTa-large 模型的两个分类器, 在 GPT-2 文本检测上实现了超过 95 % 的准确率。在主要实验中, 我们采用了 OpenAI 发布的原始模型权重, 并在不进行额外训练的情况下评估其在 LLM 生成的医学和法律文本上的 OOD 检

测性能。

5.1 大模型知识蒸馏

域种子知识为促进检测器的领域采用, 种子知识起到催化剂的作用, 促使教师 LLM 根据其初始信息生成更精细和详细的领域知识。种子知识的提示以问答格式制定, 并通过其官方 API 服务发送给 DeepSeek-R1。

在这项工作中, 我们专注于使用英语文本进行知识蒸馏。除了 EQUALS 之外, 所有的初始知识数据集最初都是用英语编写的, 并且所有从 DeepSeek-R1 生成的提示响应都是英语。对于 EQUALS (Chen et al., 2023) 数据集, 其原始法律问题是中文的, 我们利用 DeepSeek-V3 将这些问题翻译成英语, 然后再提示教师 LLM。表 10 显示了翻译提示。

蒸馏知识库我们利用教师 LLM (DeepSeek-R1) 生成的响应内容构建蒸馏知识库。由于 R1 生成的推理内容反映了模型的内部自我推理, 而不是专业专家的输出, 因此它未被用于检测器领域的适应。医疗和法律蒸馏知识库的总结在表 11 中提供。

Knowledge Base	Seed Knowledge Composition	Tokens	Size
DivScore (med)	Verifiable Questions (Chen et al., 2024)	Medical	9.2M
DivScore (law)	EQUALS (Chen et al., 2023), Open Legal QA (Butler, 2023), Pile-of-Law (Henderson et al., 2022)	Legal	17.3M
DivScore	Combined from above		26.5M

Table 11: 领域适应: 萃取知识库总结。(M: 百万标记)

学生学习。我方强制学生 LoRA 的超参数配置为 LoRA 的秩: 64, alpha 缩放因子: 128。监督微调在 Mistral 架构内选择性更新自注意力投影矩阵 (查询、键、值) 和前馈网络组件 (门,

Knowledge Base	Seed Knowledge Composition	Tokens Size
Human (med)	iCliniq (Li et al., 2023), MedAl-paca (Han et al., 2023), MedRAG (Xiong et al., 2024)	9.5M
Human (law)	Indian-Legal-Text (Yashaswat, 2024), Legal-LAMA (Chalkidis et al., 2023)	19.6M
Combined	Combined from above	29.1M

Table 12: 领域适应：人类语料库摘要（M：百万标记）

上投影)。微调使用 bfloat16 参数精度进行，训练周期数设为 8，学习率设为 $2e^{-4}$ ，批量大小设为 12，最大序列长度设为 2048。

为了评估 DivScore 在现实场景中的鲁棒性，我们对数据集应用了两种类型的对抗攻击。对于转述攻击，我们使用由 Krishna et al. (2023) 引入的 DIPPER 转述工具，这是一种基于 11B 参数的 T5 模型，针对段落级的转述进行了微调，其扰动强度可以通过词汇多样性和内容重排来控制。DIPPER 接收候选文本以及这两个参数作为输入，并在指定的扰动水平上生成转述输出。在我们的实验中，词汇多样性和内容重排因素均设置为 20%，代表中等攻击强度。

对于单词替换攻击，我们遵循 Peng et al. (2023) 的协议，选择文本中最频繁出现的单词，并使用 BERT-base 模型 (Devlin et al., 2019) 进行语义一致的替换。具体而言，BERT 模型在其最大序列长度（512 个标记）下运行，在 50 个标记的上下文窗口中替换前 20% 个最常见的单词。常见的停用词，由于缺乏实质性的语义内容，不进行替换。

在领域知识蒸馏的消融研究中，我们将所提出的领域适配方法与人类文本微调进行基准测试。对于人类文本微调，我们选择医学和法律语料库进行相同的 LoRA 微调，详细设置参见第 5.1 节。如表 12 所示，人类语料库由医学和法律领域的文本组成，在标记大小上与蒸馏知识库匹配。基准检测器参考第 ?? 节中的相同设置，我们采用 Mistral-7B-v0.2-Instruct 作为 $\mathcal{M}^*$ 来代表“无蒸馏”情境，其中检测器未能获得专门知识。在检测器组件的消融研究中，LLM 架构的详细信息列在表 3 中。所有检测器在领域知识蒸馏中具有相同的设置。

6 附加结果

本节记录实验中的附加结果。

6.1 主要实验

我们在本附录部分展示了主要实验的附加可视化结果。图 5 绘制了主要实验中所有基线检测器和所提出的检测器的 AUROC 曲线。为了更好地说明在低误报率下的检测性能，横轴（误报率）使用 $\log_{10}$ 尺度。该可视化通过提供对低误报率性能的更全面视图，补充了主要实验的指标（TPR 在 0.1% FPR）。

如图 5 所示，结合知识蒸馏的检测器在医疗、法律领域及其组合中在所有测试用例中实现了最佳的 AUROC 总分。值得注意的是，在严格的 FPR 阈值（低于 10^{-3} 到 10^{-1} ）下，我们的检测器与基线方法之间的性能差距变得特别显著。这表明 DivScore 能够以最小的误报伤害检测由 LLM 生成的专业文本。

6.2 通用领域文本的检测性能

本研究的主要重点是检测特定领域内由大型语言模型生成的内容。在本附录部分中，我们进一步评估了所提出的检测器在通用领域大型语言模型生成文本检测中的性能，这一领域的研究相对较多。我们选择了 HC3 数据集 (Guo et al., 2023) 用于大型语言模型生成文本的检测。具体来说，我们评估了 HC3 中的三个子数据集：Reddit 帖子、开放问答对和维基百科内容，涵盖了社交媒体、日常问答和面向公众的普通百科全书内容的通用大型语言模型生成文本检测场景。

我们在主要实验中采用相同的设置，从每个数据集中随机选择 1000 个人类文本和 1000 个通过 GPT-4o 和 O3-mini 生成的 LLM 文本。提示采用表 7 中所示的 QA 格式。然后，我们将提出的检测器与两个最先进的通用领域检测器进行基准测试：Fast-DetectGPT (Bao et al., 2023) 和 Binoculars (Hans et al., 2024)。正如表 13 所示，提出的检测器在通用领域数据集上表现出强大的性能。值得注意的是，DivScore（法律）在检测由 O3-mini 生成的文本时，相较于 Binoculars 实现了显著的 AUROC 和 TPR 提升，平均 AUROC (%) 增益达到 19.79，TPR (%) 增益达到 35.63。对于 GPT-4o 范式，虽然在 Reddit 和 Wiki 数据集上的 AUROC 成绩略低，但在低误报率情况下仍显示出整体更高的召回率。

这些结果证实了尽管 DivScore 是为专业领域量身定制的，但在超出专业领域时同样具有良好的泛化能力。在一般领域的 LLM 生成文本检测任务中，它仍然具有很强的竞争力，且通常优于 SOTA 方法。

本节展示了四个具有代表性的案例研究，这些案例选自关键的实验数据集——MIMIC、

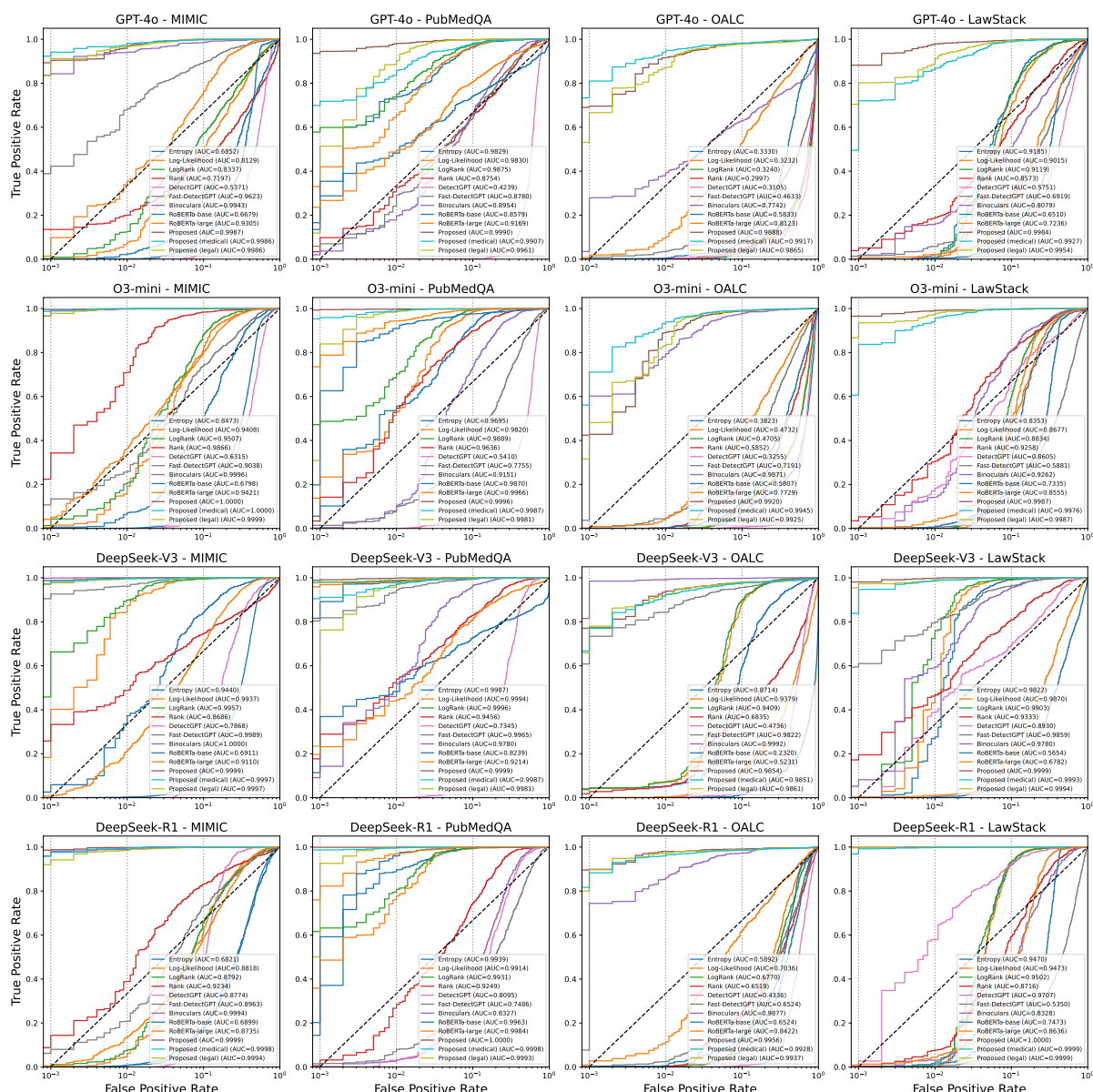


Figure 5: 基线检测器和 DivScore 在主要实验中的 ROC 曲线

PubMedQA、OALC 和 LawStack，源大型语言模型为 GPT-4o。这些案例之所以被选中，是因为它们具有很高的检测难度。值得注意的是，对于这四个案例，包括 Binoculars (Hans et al., 2024) 和 Fast-DetectGPT (Bao et al., 2023) 在内的最先进的检测器均未能识别出生成的内容。

我们从 MIMIC 数据集开始，该数据集包含去识别化的医院出院摘要。医院出院摘要通常包括有关患者入院、诊断、治疗、对治疗的反应以及出院计划的信息。这是一份重要的医疗文件，不仅限于医院使用，还在确保精确患者护理文档化和促进高效索赔处理的过程中，在如医疗保险等商业领域中发挥关键作用。表 14 展示了以 GPT-4o 作为文本来源的 LLM 所选择的 MIMIC 检测案例。

在 MIMIC 案例中，GPT-4o 使用文本前三分之一的背景信息作为提示生成了出院报告的续写。在 DeepSeek-R1 提炼的 LLM 医学领域知识的增强下，所提出的检测器在 LLM 生成的文本与真实人类文本 (0.103241 $\rightarrow$ 0.275714) 之间创建了显著的评价差距，表明有信心的零样本分类。造成这种差距的主要贡献是交叉熵组件，其中基础和增强的 LLM 在 LLM 生成的文本 (4.218750) 上显示出很大的分歧，同时在真人文本 (2.734375) 上保持相似的概率估计。预计在检测器阈值设置为低于 0.1 % FPR (即检测中少于 1 个误检病例) 的情况下，这个案例将被正确分类。

表 15 展示了来自 OALC 数据集的一个案例，其中包含澳大利亚政府和法院的正式立法和

Methods	GPT-4o			O3-mini		
	Reddit	OpenQA	Wiki	Reddit	OpenQA	Wiki
AUROC						
Fast-DetectGPT	<u>86.11</u> ($\downarrow 1.23$)	<u>76.26</u> ($\downarrow 4.83$)	<u>96.29</u> ($\downarrow 1.92$)	<u>79.41</u> ($\uparrow 2.83$)	<u>61.24</u> ($\uparrow 4.76$)	<u>89.31</u> ($\downarrow 4.17$)
Binoculars	87.34	81.09	98.21	76.58	56.48	93.48
DivScore (med)	79.22 ($\downarrow 8.12$)	97.10 ($\uparrow 16.01$)	94.84 ($\downarrow 3.37$)	<u>88.90</u> ($\uparrow 12.32$)	<u>97.27</u> ($\uparrow 40.79$)	<u>98.86</u> ($\uparrow 5.38$)
DivScore (law)	80.67 ($\downarrow 6.67$)	<u>96.17</u> ($\uparrow 15.08$)	<u>96.77</u> ($\downarrow 1.44$)	<u>89.17</u> ($\uparrow 12.59$)	<u>97.55</u> ($\uparrow 41.07$)	<u>99.37</u> ($\uparrow 5.89$)
DivScore	78.13 ($\downarrow 9.21$)	<u>95.96</u> ($\uparrow 14.87$)	90.72 ($\downarrow 7.49$)	<u>88.02</u> ($\uparrow 11.44$)	<u>97.19</u> ($\uparrow 40.71$)	<u>97.70</u> ($\uparrow 4.22$)
TPR @ 0.1 % FPR						
Fast-DetectGPT	<u>14.93</u> ($\uparrow 14.33$)	<u>1.50</u> ($\uparrow 1.2$)	<u>21.62</u> ($\uparrow 0$)	<u>6.61</u> ($\uparrow 6.31$)	<u>0.10</u> ($\uparrow 0$)	<u>4.39</u> ($\uparrow 1.06$)
Binoculars	0.60	0.30	21.62	0.30	0.10	3.33
DivScore (med)	6.31 ($\uparrow 5.71$)	17.65 ($\uparrow 17.35$)	<u>28.27</u> ($\uparrow 6.65$)	12.53 ($\uparrow 12.23$)	<u>11.52</u> ($\uparrow 11.42$)	<u>54.75</u> ($\uparrow 51.42$)
DivScore (law)	<u>12.73</u> ($\uparrow 12.13$)	5.22 ($\uparrow 4.92$)	57.84 ($\uparrow 36.22$)	<u>26.85</u> ($\uparrow 26.55$)	<u>1.90</u> ($\uparrow 1.8$)	<u>81.83</u> ($\uparrow 78.5$)
DivScore	4.41 ($\uparrow 3.81$)	<u>16.25</u> ($\uparrow 15.95$)	13.90 ($\downarrow 7.72$)	<u>16.43</u> ($\uparrow 16.13$)	<u>15.43</u> ($\uparrow 15.33$)	<u>43.94</u> ($\uparrow 40.61$)

Table 13: 通用域大语言模型生成文本检测性能，AUROC (%) & TPR (在 0.1 % FPR 时) 被选择的方法在 HC3-reddit、HC3-OpenQA、HC3-Wiki 数据集上使用 GPT-4o 和 O3-mini 作为文本来源的大语言模型。方法与 Binoculars 基准进行比较，箭头表示相对于 Binoculars 的性能差异 ($\uparrow$)。

司法文件。法律文件如法规和规章以其精确的语言、正式的结构和严格的逻辑一致性为特征。在这个 OALC 案例中，LLM 生成的续写紧密地模仿了原始法院文件的正式语气和结构，具有类似的结构和语言特征，导致望远镜和 Fast-DetectGPT 的检测失败。提出的检测器利用来自 DeepSeek R1 的领域特定法律知识，有效地区分 LLM 生成的文本和人类撰写的真实文件，并在评分上显示出明显差距 ($0.163043 \rightarrow 0.084449$)。熵的降低再次在检测中起到关键作用 ($0.820312 \rightarrow 0.443359$)。在检测器阈值设置为低于 0.1 % FPR 时，人类和 LLM 生成的文本都被自信且正确地分类，展示了检测器在处理高度专业化法律内容方面的鲁棒性。

表 16 展示了从 PubMedQA 中选取的案例。PubMedQA 包含从 PubMed 数据库中的生物医学研究论文中提取的成对问题和答案。具有类似形式的生物医学研究问答文本在公共健康促进、医生培训和循证医学实践中起到了关键作用。在 PubMedQA 案例中，通过从 DeepSeek-R1 中提取的 LLM 医学领域知识进行增强，所提出的检测器在 LLM 生成的文本和真实人类文本之间创造了显著的评分差距 ($0.272876 \rightarrow 0.116497$)。这种差距归因于 LLM 生成文本 (0.535156) 相比人类文本 (1.304688) 具有明显更小的熵估计。这种熵差异可能是成功提炼 LLM 医学知识的结果。通过 LLM 医学知识增强后，评分 LLM 预期会在 LLM 生成的答案中表现出较低的熵，因为模型倾向于产生比人类写作中更广泛的多样性更可预测和一致的模式。这一发现突出了所提出检测框架中熵组件的作用。

表格 17 展示了一个来自 LawStack 数据集的案例，该数据集包含覆盖广泛司法管辖区和法律主题的法律问答对。这类问答对通常用于法律咨询、合规培训和大众法律教育。在这个案例中，与人工撰写的答案相比，LLM 生成的答案检测评分接近三分之一，表明 DivScore 的分类较为自信。如表 17 所示，这种差距是评分组件共同结果。随着熵和交叉熵的预期变化，DivScore 检测器能够成功地在最严格的阈值 (FPR 低于 0.1 %) 下对 LLM 和人类文本进行分类。

MIMIC: Hospital Discharge Summary

System Prompt: Continue this hospital discharge summary.

User Prompt:

Name: ____ Unit No: ____ Admission Date: ____ Discharge Date: ____ Date of Birth: ____ Sex: F Service: OBSTET-
RICS/GYNECOLOGY Allergies: Patient recorded as having No Known Allergies to Drugs Attending: ____ Chief Com-
plaint: none Major Surgical or Invasive Procedure: none History of Present Illness: ____ yo G1P0 at 9 weeks gestation
with T1 DM, sent by Dr. ____ for diabetic control. She was on Lantus and Humalog prior to pregnancy and on NPH and
Humalog during pregnancy. Reports fingersticks have been average 250's. She feels well and is without complaints. Past
Medical History: PRENATAL COURSE (1)LMP ____ (2)No initial prenatal yet PAST OBSTETRIC HISTORY G1 PAST
GYNECOLOGIC HISTORY - no abnl paps or STIs PAST MEDICAL HISTORY - T1DM PAST SURGICAL HISTORY
- ureteral reimplantation at age ____ Social History: ____ Family History: non-contributory Physical Exam: VITALS: T
96.8, HR 96, BP 128/70, RR 26, O2 100 % RA GENERAL: NAD HEART: RRR LUNGS: CTAB ABDOMEN: soft, NT
EXTREMITIES: NT/NE PELVIC US: live IUP confirmed

Pertinent Results: ____ WBC-8.3 RBC-4.26 Hgb-13.0 Hct-37.0 MCV-87 Plt-314 ____ WBC-9.5 RBC-4.30 Hgb-13.2 Hct-
36.5 MCV-85 Plt-314 ____ Neuts-61.8 ____ Monos-4.8 Eos-2.0 Baso-0.3 ____

Human Text:

BLOOD HBsAg-NEGATIVE, HIV Ab-NEGATIVE ____ URINE pH-7 Hrs-24 Volu-1500 Creat-86 TotProt-<6 ____
URINE 24Creat-1290 RUBEOLO ANTIBODY, IgG (Final ____ POSITIVE BY EIA) RAPID PLASMA REAGIN TEST
(Final ____ NONREACTIVE) Rubella IgG/IgM Antibody (Final ____ POSITIVE by Latex Agglutination) EARLY OB
ULTRASOUND IMPRESSION: Single live intrauterine gestation. Size equals dates.

Brief Hospital Course: ____ G1 with poorly controlled T1DM admitted at 9wks for glycemic control. Ms ____ had finger-
sticks ranging from 151-236 on arrival to the emergency room. She had no complaints and electrolytes were normal. Her
hemoglobin A1C was 10.0 % . She was admitted to the antepartum floor and ____ consulted and continued to follow her
closely throughout this admission. Initially, it was unclear whether her recent elevated fingersticks were due to insulin
omission or increased requirements. Nutrition was consulted. Her regimen was titrated to achieve optimal glycemic con-
trol. She was continued on NPH in the am as well as an NPH sliding scale at night. She had a humalog sliding scale with
meals. Her regimen was increased and by hospital day # 4 her fingersticks were improved. She was discharged home and
will have close outpatient management with ____ and the Diabetes in Pregnancy clinic. She was instructed to call with
fingersticks greater than 200 for two hours. During this admission, Ms ____ had an early OB ultrasound which showed a
live IUP, size equal to dates. Prenatal labs were done and were unremarkable. She was seen by social services due to her
uncertain psychiatric history. Please see OMR note for details.

Medications on Admission: prenatal vitamins NPH 36 units qam and 20 units qhs Humalog per sliding scale with meals
Discharge Medications: 1. Insulin NPH Human Recomb Subcutaneous 2. Please follow the insulin doses and sliding
scale as detailed on the attached sheet. 3. Humalog Subcutaneous 4. B Complex-Vitamin C-Folic Acid 1 mg Capsule Sig:
One (1) Cap PO DAILY Discharge Disposition: Home Discharge Diagnosis: Type 1 Diabetic admitted for glycemic con-
trol Discharge Condition: stable Discharge Instructions: Please call your ob doctor please call if you experience vaginal
bleeding, leakage of fluid or any other concerns. Please continue current insulin regimen. Followup Instructions: ____

DivScore $D_{\mathcal{M}, \mathcal{M}^*}(x)$: 0.275714 , Entropy $H_{\mathcal{M}^*}(x)$: 0.753906 , Cross-Entropy $CE_{\mathcal{M}, \mathcal{M}^*}(x)$: 2.734375

LLM Text:

Patient discharged with the following instructions and management plan: Discharge Medications: 1. Insulin Therapy:
Continue current regimen of NPH and Humalog as managed during the hospital stay. Patient provided with updated
dosage chart and instructed to monitor blood glucose levels at home. 2. Prenatal Vitamins: One tablet daily. 3. Folic Acid:
0.4 mg daily.

Discharge Instructions: 1. Glycemic Control: Regular blood sugar monitoring at home. Aim for fasting blood glucose
levels of 70-95 mg/dL and postprandial levels <140 mg/dL. Contact healthcare provider if experiencing episodes of hy-
poglycemia or persistent hyperglycemia. 2. Dietary Recommendations: Continue diabetic-friendly diet. Referral to a
registered dietitian for nutritional counseling has been made. 3. Prenatal Care: Schedule an initial prenatal appointment
with Dr. ____ . Regular prenatal visits are important for monitoring maternal and fetal health. 4. Activity Restrictions:
No specific restrictions. Engage in normal physical activity as tolerated. 5. Warning Signs: Seek immediate medical
attention if experiencing symptoms such as abdominal pain, vaginal bleeding, decreased fetal movement, or signs of hy-
perglycemia/hypoglycemia.

Follow-Up: 1. Primary Care Provider: Appointment scheduled for follow-up on insulin management. 2. Obstetrician:
Appointment scheduled within 1 week post-discharge for prenatal evaluation and further obstetric care planning. Progno-
sis: The patient is stable on discharge with well-controlled Type 1 Diabetes Mellitus in pregnancy. Regular follow-up and
adherence to treatment plan will support a healthy pregnancy outcome. Conclusion: The patient was discharged in good
condition, with detailed instructions provided for managing her diabetes and prenatal care. She exhibited understanding
and agreement with the discharge plan. Please contact Dr. ____ or the OB/GYN clinic at ____ for any further questions or
concerns regarding this discharge plan.

DivScore $D_{\mathcal{M}, \mathcal{M}^*}(x)$: 0.103241 , Entropy $H_{\mathcal{M}^*}(x)$: 0.435547 , Cross-Entropy $CE_{\mathcal{M}, \mathcal{M}^*}(x)$: 4.218750

Detector Threshold @0.1 % FPR: 0.206166

Table 14: MIMIC 案例研究：比较人类和 LLM 生成的出院摘要与检测评分

System prompt: Continue write this legal text.

User prompt:

FEDERAL COURT OF AUSTRALIA Van der Velde v Ng [2008] FCA 480 TERRY VAN DER VELDE AND DAVID STIMPSON AND RIBY PTY LTD (ACN 081 153 219) IN LIQUIDATION AS TRUSTEE FOR THE JADE TRUST v KYM HON YOKE NG, THE OFFICIAL TRUSTEE AS TRUSTEE OF THE PROPERTY OF ROSS HASTINGS (A BANKRUPT) AND RAMS MORTGAGE CORPORATION LIMITED ABN 48 065 912 932 QUD 405 OF 2007 DOWSETT J 4 FEBRUARY 2008 BRISBANE IN THE FEDERAL COURT OF AUSTRALIA QUEENSLAND DISTRICT REGISTRY QUD 405 OF 2007 BETWEEN: TERRY VAN DER VELDE AND DAVID STIMPSON First Applicant RIBY PTY LTD (ACN 081 153 219) IN LIQUIDATION AS TRUSTEE FOR THE JADE TRUST Second Applicant AND: KYM HON YOKE NG First Respondent THE OFFICIAL TRUSTEE

Human Text:

AS TRUSTEE OF THE PROPERTY OF ROSS HASTINGS (A BANKRUPT) Second Respondent RAMS MORTGAGE CORPORATION LIMITED ABN 48 065 912 932 Third Respondent JUDGE: DOWSETT J DATE: 4 FEBRUARY 2008 PLACE: BRISBANE REASONS FOR JUDGMENT

1 In both Fraser v Deputy Commissioner of Taxation & Official Trustee (1996) 69 FCR 99 and Scott v Bagshaw (1999) 95 FCR 37, the relevant applicant was seeking to enforce a provable debt although, in the latter case, that may not have appeared from the statement of claim. In those circumstances, s 58(3) of the Bankruptcy Act 1966 (Cth) clearly applied.

2 That is not the present case, as I understand it. In those circumstances, I am of the view that leave is not necessary. However, should in the course of the trial it emerge otherwise, we can revisit the question. I certify that the preceding two (2) numbered paragraphs are a true copy of the Reasons for Judgment herein of the Honourable Justice Dowsett.

Associate: Dated: 14 April 2008 Solicitor for the First Applicant: Mills Oakley Counsel for the Second Applicant: The Second Applicant did not appear Counsel for the First Respondent: The First Respondent did not appear Solicitor for the Second Respondent: Rodgers, Barnes & Green Counsel for the Third Respondent: Mr D S Piggott Solicitor for the Third Respondent: Mallesons Stephen Jacques Date of Hearing: 4 February 2008 Date of Judgment: 4 February 2008 DivScore $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.163043 , Entropy $H_{\mathcal{M}^*}(x)$: 0.820312 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 5.03125

LLM Text:

AS TRUSTEE OF THE PROPERTY OF ROSS HASTINGS (A BANKRUPT) Second Respondent RAMS MORTGAGE CORPORATION LIMITED ABN 48 065 912 932 Third Respondent JUDGE: DOWSETT J DATE OF ORDER: 4 FEBRUARY 2008 WHERE MADE: BRISBANE REASONS FOR JUDGMENT

Introduction 1. This matter involves a complex set of legal proceedings concerning multiple parties and a series of claims arising from financial transactions, property interests, and trust management. The applicants, Terry Van der Velde and David Stimpson, along with Riby Pty Ltd acting as trustee for the Jade Trust, seek remedies against the respondents, who include Kym Hon Yoke Ng, the Official Trustee managing the bankruptcy estate of Ross Hastings, and RAMS Mortgage Corporation Limited.

Background 2. The facts of the case date back to transactions undertaken in 2005, where the applicants allege financial mismanagement and breaches of fiduciary duty by the first respondent, leading to significant financial detriments and potential loss of property interests administered under the Jade Trust.

3. The involvement of the Official Trustee arises from the concurrent bankruptcy proceedings associated with Ross Hastings, necessitating the management of his financial affairs, while the third respondent, RAMS Mortgage Corporation Limited, is implicated due to its interests in mortgage securities tied to the disputed property.

Legal Issues 4. The key legal issues revolve around the fiduciary duties allegedly breached by the first respondent, the implications of these breaches on the trust's financial stability, the proper scope of actions available by the liquidated Riby Pty Ltd against the respondents, and the determination of priority among the creditors, including RAMS Mortgage Corporation Limited.

5. Additional questions concern the jurisdiction of this Court in managing disputes related to bankruptcy and asset recovery, with particular attention to cross-claims presented by the second respondent regarding the equitable distribution of assets.

Proceedings and Orders 6. The Court will undertake a detailed examination of the evidence presented, including financial statements, transaction records, and testimonial accounts provided by all parties. This examination will allow for an informed judgement on breaches of duty and mismanagement claims.

7. Interim orders will be considered to safeguard the property interests of the applicants pending final resolution, including potential injunctions against foreclosure actions by mortgage holders.

Conclusion 8. The outcomes of this case will clarify the obligations of trustees and managers of financial trusts and provide guidance for creditor priorities within bankruptcy contexts, contributing to the broader understanding of fiduciary and financial stewardship in corporate and trust law. * * * I certify that the preceding text forms the official reasons for judgment handed down by the Honourable Justice Dowsett, on the 4th day of February, 2008. [Signature of Justice Dowsett] DivScore $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.084449 , Entropy $H_{\mathcal{M}^*}(x)$: 0.443359 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 5.25000

Detector Threshold @0.1 % FPR: 0.118119

Table 15: OALC 案例研究：人类和 LLM 生成的法律文件及检测评分的比较

PubMedQA: Answer to Biomedical Research Questions
System Prompt: Answer this question truthfully.
User Prompt: Is adjustment for reporting heterogeneity necessary in sleep disorders?
Human Text: Sleep disorders are common in the general adult population of Japan. Correction for reporting heterogeneity using anchoring vignettes is not a necessary tool for proper management of sleep and energy related problems among Japanese adults. Older age, gender differences in communicating sleep-related problems, the presence of multiple morbidities, and regular exercise should be the focus of policies and clinical practice to improve sleep and energy management in Japan. Detection Score $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.272876 , Entropy $H_{\mathcal{M}^*}(x)$: 1.304688 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 4.78125
LLM Text: Adjustment for reporting heterogeneity can be necessary in sleep disorders research to ensure accurate and reliable results. Reporting heterogeneity refers to variations in how information is documented or reported, which can stem from differences in assessment tools, study design, participant demographics, cultural differences, and subjective interpretations of symptoms. In sleep disorders, symptoms and their severity can be perceived differently by individuals due to personal, cultural, or situational factors. Without adjusting for these differences, there can be inconsistencies in data interpretation, potentially leading to skewed or misleading findings. Therefore, employing methods to adjust for reporting heterogeneity can enhance the validity and comparability of research outcomes across different studies or populations. DivScore $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.116497 , Entropy $H_{\mathcal{M}^*}(x)$: 0.535156 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 4.59375
Detector Threshold @0.1 % FPR: 0.130639

Table 16: PubMedQA 案例研究：人类与 LLM 生成的生物医学研究问题答案的比较及检测评分

LawStack: Answer to Legal Questions

System Prompt:

Answer this question truthfully:

User Prompt:

I have read the question and answer here: Are parking tickets enforceable But I would like to know that if a person got a parking ticket in Pennsylvania for a street sweeper violation and didn't know until they got a registered letter from the city. In that letter, they state that a person can plead not guilty (and send in a bond amount) or guilty. Which leads me to think that it's a criminal offense and not a civil one. If it's a criminal offense, wouldn't the burden of proof be on the city to prove that a particular person (and not just the owner) did the parking? Or is it really sufficient to serve the person who they think is the owner? Can they really issue a warrant for arrest if the owner might not have ever set foot in that state before? If personal service isn't made (as outlined in the above question) then is a parking ticket in PA really enforceable? What if the owner of the car doesn't live in Pennsylvania and instead lives in Texas? Can they be extradited to PA from TX or placed under arrest somehow?

Human Text:

The procedure depends on the local rules. For example, Philadelphia. In Philadelphia, parking offences are civil, not criminal - other governments may be different. Irrespective of if it is civil or criminal, the burden is on the city. All that is different is the burden: for civil it is the preponderance of the evidence (or balance of probabilities), for criminal it is beyond reasonable doubt. Under Philadelphia law, anyone driving with the owner's permission is an agent of the owner so serving them is serving the owner. There is an exemption that car rental companies can avail themselves of. The can take whatever enforcement action the law allows - this may include issuing an arrest warrant. Serving by post or affixing to the vehicle is legally service. If there is a valid warrant, they can be extradited from anywhere subject to relevant extradition law - most extradition laws require the crime to be "serious" - parking violations probably aren't. DivScore $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.203187 , Entropy $H_{\mathcal{M}^*}(x)$: 0.796875 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 3.921875

LLM Text:

In Pennsylvania, as in many other jurisdictions, parking violations, including those for street sweeper offenses, are generally considered civil infractions rather than criminal offenses. This means that the violation is typically against the vehicle and not directly against the individual who parked it. As such, the burden of proof is generally on the city to show that the vehicle was parked in violation of the law, rather than proving who specifically parked the car. The registered letter you received offering options to plead not guilty or guilty and pay a bond amount is a common procedural step in handling such civil infractions. While parking tickets are civil matters, failure to address them can lead to additional consequences, such as fines, vehicle registration holds, or even a bench warrant for failing to respond to the citation, which can create legal complications. Regarding jurisdiction, if the owner of the vehicle resides out of state, such as in Texas, Pennsylvania cannot extradite the individual solely for an unpaid parking ticket. Extradition typically applies to criminal offenses. However, unresolved parking tickets can lead to complications like increased fines or affecting the owner's ability to register their vehicle, depending on interstate compacts or agreements. If you believe that the ticket was issued in error or you have compelling reasons to challenge it, you may wish to contest the ticket through the legal process outlined in the letter. It can also be helpful to seek legal advice to understand the specific options and implications based on the circumstances of the case.

DivScore $D_{\mathcal{M},\mathcal{M}^*}(x)$: 0.078887 , Entropy $H_{\mathcal{M}^*}(x)$: 0.404297 , Cross-Entropy $CE_{\mathcal{M},\mathcal{M}^*}(x)$: 5.125000

Detector Threshold @0.1 % FPR: 0.099856

Table 17: LawStack 案例研究：比较人类和 LLM 生成的对法律问题的答案与检测评分