DivScore: Zero-Shot Detection of LLM-Generated Text in Specialized Domains

Anonymous ACL submission

Abstract

Detecting LLM-generated text in specialized and high-stakes domains like medicine and law is crucial for combating misinformation and ensuring authenticity. However, current zero-shot detectors, while effective on general text, often fail when applied to specialized content due to domain shift. We provide a theoretical analysis showing this failure is fundamentally linked to the KL divergence between human, detector, and source text distributions. To address this, we propose DivScore, a zero-shot detection framework using normalized entropy-based scoring and domain knowledge distillation to robustly identify LLM-generated text in specialized domains. We also release a domain-specific benchmark for LLM-generated text detection in the medical and legal domains. Experiments on our benchmark show that DivScore consistently outperforms state-of-the-art detectors, with 14.4% higher AUROC and 64.0% higher recall (0.1% false positive rate threshold). In adversarial settings, DivScore demonstrates superior robustness than other baselines, achieving on average 22.8% advantage in AUROC and 29.5% in recall. Code and data are publicly available[1].

1   Introduction

As Large Language Models (LLMs) produce text of increasing fluency and realism, the need for reliable machine-generated text detection has become critical. This is especially true in high-stakes domains like healthcare and law, where undetected AI-written content could undermine trust and safety.

Initial approaches to detecting LLM-generated text employed supervised learning, such as GPT-2 detector (Solaiman et al., 2019) and GPT-3.5 detector (Guo et al., 2023). While effective on the data they were trained on, such supervised detectors show clear shortcomings when faced with new domains or more advanced language models. For
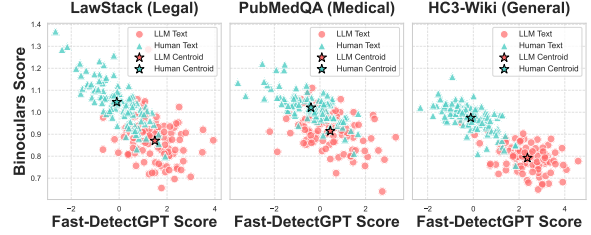

Figure 1: Fast-DetectGPT (Bao et al., 2023) and Binoculars (Hans et al., 2024) detector scores of LLM/Human texts in specialized (legal, medical) and general domain. In specialized domain, both detectors failed in creating scoring gap necessary for detecting LLM-generated text

instance, a detector trained on law articles may falter on healthcare text, and a classifier tuned to GPT-2 outputs may misclassify text from a newer LLM. Furthermore, every new LLM or domain requires the creation of a fresh labeled dataset and retraining of the detector, rendering the process resource-intensive and inefficient.

In response, researchers have developed detectors that require no training examples and instead exploit statistical irregularities of LLM-generated text. For instance, DetectGPT (Mitchell et al., 2023) uses an LLM's probability curvature to flag generated passages without any finetuning, while Binoculars (Hans et al., 2024) contrasts the outputs of two language models to achieve state-of-the-art detection accuracy with zero training data.

However, most existing zero-shot detection studies focus on general domain data. Specialized domains such as healthcare literature or legal documents present several additional challenges that make LLM-generated text detection more difficult. These domains typically have: (1) Sparse labeled data. Few or no annotated examples of LLM-generated text for training. (2) Complex language. Domain-specific jargon and intricate writing styles that differ markedly from common text. (3) Unfamiliar content. Topic areas and facts that

---

[1]https://anonymous.4open.science/r/DivScore-CF66

1

even advanced LLMs may not have seen frequently during pre-training.

As a result, detection methods that perform well on general text often struggle under such domain shift. As shown in Figure 1, we observe that Fast-DetectGPT (Bao et al., 2023) and Binoculars (Hans et al., 2024) method both tend to compress in specialized domains: both human and LLM-generated texts yield similar detector scores, eroding the gap needed for reliable discrimination. To address these gaps, we propose DivScore – a normalized entropy-based divergence metric for domain-specific, zero-shot detection of LLM-generated text in specialized domains. Our approach is specifically designed to be robust across different domains.

In essence, DivScore works by normalizing the domain-specific entropy of a candidate text against an estimated baseline, thereby measuring how surprising the text is relative to what a reference language model expects for that domain. Our method first highlights domain-specific alignment, and then applies entropy normalization to accentuate the differences between human-written and LLM-generated text, thereby enhancing the contrast necessary for detection. We implement DivScore using an unsupervised domain knowledge distillation framework, maintaining a zero-shot setting that requires no labeled data. In our evaluations, DivScore outperforms state-of-the-art (SOTA) zero-shot baselines under significant domain shifts, while remaining competitive in general domain settings. We make the following key contributions:

- We introduce **DivScore**. To the best of our knowledge, it is the first zero-shot detection method designed for identifying domain-specific LLM-generated text, inspired by domain alignment and entropy normalization.

- We release a high-quality benchmark for detecting LLM-generated text in specialized domains. It includes professionally legal and medical corpora, along with knowledge distillation datasets and an adaptation pipeline for detector evaluation and finetuning.

- We conduct extensive experiments and analysis. The results show that DivScore outperforms SOTA baselines by 14.4% in AUROC and 64.0% in recall (0.1% false positive rate), while also exhibiting strong robustness under adversarial conditions.

2 Related Work

Machine-generated text detection is typically formulated as a binary classification problem (Su et al., 2023; Bao et al., 2023; Mitchell et al., 2023; Solaiman et al., 2019; Hans et al., 2024). Supervised detectors, such as RoBERTa-based models (Solaiman et al., 2019), perform well on in-domain data, but require additional manual annotation, with limited scalability and flexibility (Bakhtin et al., 2019; Su et al., 2023; Mitchell et al., 2023).

Another group of studies aims to construct zero-shot detectors based on the discrepancy in LLM-measured statistical signatures between human and machine-generated text. Text statistics-based detectors such as LLM text entropy (Lavergne et al., 2008), log-probability (Solaiman et al., 2019) and per-token rank (Gehrmann et al., 2019) have proven to be effective in general LLM text detection. Mitchell et al. (2023) proposed DetectGPT to detect LLM-generated passages by the curvature of LLM text log-probability function under controlled text perturbations. Fast-DetectGPT (Bao et al., 2023) substitutes the perturbation method in DetectGPT with a more efficient LLM sampling step. Hans et al. (2024) further discovered that contrasting the LLM text perplexity of two closely related language models can effectively distinguish LLM-generated content.

However, the zero-shot property comes at the cost of limited adaptability in specialized domains. Current zero-shot detectors, including Binoculars (Hans et al., 2024), mainly focus on general domain content, such as news, essays, and social media posts (Narayan et al., 2018; Guo et al., 2023; Fan et al., 2018; Wu et al., 2025a; Dugan et al., 2024). Relying on general domain LLMs for statistical signature scoring, current methods implicitly assume sufficient domain familiarity from the detector. In specialized domains with rare knowledge (e.g., medical or legal), both human and LLM-generated texts could be unfamiliar to these detectors, collapsing the statistical gap and degrading detection performance. Our DivScore is designed for avoiding such problems.

3 Methodology

In this section, we first formalize why zero-shot detectors falter when the text distribution shifts to a specialised domain, setting the theoretical stage for **DivScore**. Guided by this analysis, we derive our core detector and then introduce a domain-
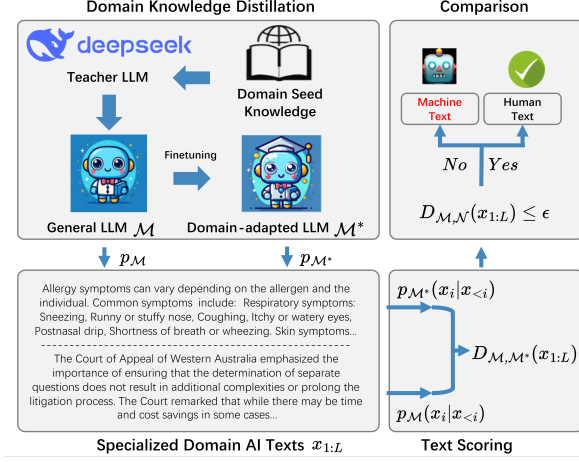
2

Figure 2: The overall framework of DivScore

3.1 Theoretical Analysis: Why Do Zero-Shot Detectors Fail under Domain Shift?

knowledge distillation step that narrows the gap between the detector and specialised language, markedly strengthening robustness and accuracy. The framework of DivScore is shown in Figure 2.

Most zero-shot detectors, such as entropy-based scoring (Lavergne et al., 2008) or perplexity-based Binoculars (Hans et al., 2024), assume that LLM-generated text exhibits statistical signatures (e.g., lower entropy, distinctive perplexity) that are distinguishable from human-written text.

However, when applied to corpus from specialized domains (see Figure 1), these signals often collapse: both human- and machine-generated texts may appear equally "surprising" or "familiar" to a general domain LLM. To formalize this phenomenon, we firstly analyse the theoretical performance of a general zero-shot detector that scores a text X based on its likelihood $p_Q(X)$ with a detector LLM. Specifically, we present the following theorem, which characterizes the relationship between detector performance and the two specific Kullback–Leibler (KL) divergence.

Theorem 1. *Given a zero-shot LLM-generated text detector D with detection score mapped from the detector LLM's text likelihood $p_Q(X)$ by any strictly monotonic function $f : \mathbb{R} \to \mathbb{R}$:*

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

where Q represents the text distribution calculated from D. The area under the receiver operating characteristic curve (AUROC) of D, denoted as

AUROC($D(X)$), *satisfies:*

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

where $\propto^+$ denotes positive monotonic association, and δ_{KL}, effective KL divergence, is given as:

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

with Q' indicating text distribution from the real source. P denotes text distribution from human. $D_{KL}(\cdot||\cdot)$ denotes KL divergence.

Theorem 1 shows that the performance of a zero-shot detector depends on the extent to which the human text distribution P diverges more from the real source distribution Q' than from the detector distribution Q. In specialized domains (e.g., medical or legal) with distinctive linguistic characteristics, the source distribution Q' may more closely resemble P than the general-domain LLM distribution Q. Consequently, a reduction in δ_{KL} diminishes the detector's ability to distinguish machine-generated text within such domains. See detailed theorem proof in Appendix A.

3.2 DivScore: Normalized Entropy-based Scoring

Building on the theoretical analysis, we design a detector that leverages the entropy of a domain-adapted LLM $\mathcal{M}^*$ (see subsection 3.3) to quantify how predictable or familiar a sentence $x_{1:L}$ is under its distribution. Given a $x_{1:L} = (x_1, ..., x_L)$, the mean token entropy is defined as:

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

Ideally, if $\mathcal{M}^*$ is well aligned with the distribution of LLM-generated text in the designated domain, LLM-generated text should yield lower entropy than human-written text. However, in practice, this assumption may not always hold due to subtle domain mismatches or the presence of highly predictable (i.e., low-entropy) human texts.

To address this, we introduce a normalization scheme based on the cross-entropy between a general LLM $\mathcal{M}$ and the domain-adapted LLM $\mathcal{M}^*$. The cross-entropy is defined as:

$$CE_{\mathcal{M},\mathcal{M}^*}(x_{1:L}) = -\frac{1}{L} \sum_{i=1}^{L} p_{\mathcal{M}}(x_i \mid x_{<i}) \\ \cdot \log p_{\mathcal{M}^*}(x_i \mid x_{<i}) \quad (5)$$

The above cross-entropy quantifies the degree of disagreement between the two models' probability estimates for text. For LLM-generated text in a specialized domain, the domain-adapted $\mathcal{M}^*$ is expected to diverge significantly from the general LLM $\mathcal{M}$, resulting in a high cross-entropy score. In contrast, for human-written text that is unfamiliar to both models, the disagreement is typically lower, leading to a lower cross-entropy value. This intuition is further supported by a detailed statistical analysis in Section 4.4. The final DivScore is computed as:

$$D_{\mathcal{M},\mathcal{M}^*}(x_{1:L}) = \frac{H_{\mathcal{M}^*}(x_{1:L})}{CE_{\mathcal{M},\mathcal{M}^*}(x_{1:L})} \quad (6)$$

A higher $D_{\mathcal{M},\mathcal{M}^*}(x_{1:L})$ driven by high entropy (unfamiliarity) and low cross-entropy (low model disagreement) - indicates that $x_{1:L}$ is more likely to be human-written. Conversely, a lower score suggests that the text is machine-generated. This DivScore normalization enhances the separation between human and LLM-generated text, especially in specialized domains where raw entropy-based signals alone are insufficient.

3.3 Enhancing the Detector via Unsupervised Domain Knowledge Distillation

DivScore employs a paired LLM $(\mathcal{M}, \mathcal{M}^*)$, consisting of a general-purpose and domain-adapted model, for scoring. In the basic DivScore setup, $\mathcal{M}^*$ can be any LLM adapted for the target domain. Further, motivated by Theorem 1 and aiming to enhance detection performance in highly specialized domains, we introduce a tailored improvement: constructing $\mathcal{M}^*$ by distilling domain-specific knowledge into the general model $\mathcal{M}$. The goal of this enhancement is to reduce the KL divergence $D_{\text{KL}}(Q \parallel Q')$ by aligning the detector's distribution with the domain-specific LLM-generated text. The process is as follows:

1. **Teacher Model Selection:** Select a large language model (LLM) with extensive domain expertise and a substantial parameter count to act as the teacher model (e.g., DeepSeek-R1).

2. **Domain Seed Knowledge Construction:** Collect a set of questions representative of the specialized domain. Using seed knowledge, prompt the teacher LLM to generate detailed and in-depth questions based on the initial input.

3. **Distillation Knowledge Generation:** Use the teacher LLM to generate high-quality answers and narrative explanation to the seed questions, forming a domain-specific distillation knowledge base consisting of outputs from the teacher LLM.

4. **Student Learning:** Finetune the general LLM $\mathcal{M}$ of DivScore with the generated domain-specific question-answer pairs, yielding the enhanced model $\mathcal{M}^*$ adapted to this specialized domain.

This distillation process for building $\mathcal{M}^*$ transfers domain knowledge from a teacher LLM to the general LLM $\mathcal{M}$ of DivScore, narrows the gap between the detector and the LLM text distributions, and thus increases δ_{KL} in Theorem 1, improving detection performance. The entire approach operates in a zero-shot manner, requiring no human-labeled text for training and inference, which makes it highly practical for real-world deployment in specialized domains.

We release all the above resources as a benchmark, which consists of (1) domain-specific knowledge distillation datasets and adaptation pipelines for medical and legal domains, and (2) 4,000 high-quality professional legal and medical documents paired with LLM-generated counterparts for detector evaluation.

4 Experiments

In this section, we evaluate DivScore's effectiveness in detecting machine-generated text within legal and medical domains. We further assess its robustness to adversarial attacks that attempt to bypass detection. Additionally, we analyse how domain knowledge distillation influences text scoring, and present ablation studies examining the effects of detector LLM selection, knowledge distillation, and scoring components on overall performance.

4.1 Implementation Details

This study investigates zero-shot, black-box detection settings that realistically reflect real-world deployments. In the zero-shot scenario, a detector must identify LLM-generated text without any task-specific labeled examples. The black-box constraint further withholds information about the source LLM and its logits, so detectors must rely on surrogate models for scoring. Together, these

4

Text Source	GPT-4o				O3-mini				DeepSeek-V3				DeepSeek-R1			
Datasets	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(↓9.63)	58.33(↓19.86)	65.10(↓29.98)	67.98(↓26.71)	98.70(↑14.13)	58.07(↑0.67)	73.35(↓9.02)	69.11(↓30.89)	82.39(↓17.33)	23.20(↓75.07)	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(↓0.48)	99.66(↑15.09)	77.29(↑19.89)	85.55(↑3.18)	91.10(↓8.9)	92.14(↑7.58)	52.31(↓45.96)	67.82(↓32.05)	87.35(↓11.75)	99.84(↓9.67)	84.22(↑29.67)	86.36(↓1.53)
Entropy	68.52(↓29.82)	98.29(↑2.87)	33.30(↓5.17)	91.85(↓3.23)	84.73(↓9.96)	96.95(↑12.38)	38.23(↓19.17)	83.53(↑1.16)	94.40(↓5.6)	99.87(↑0.15)	87.14(↓11.13)	98.22(↑1.65)	68.21(↓20.89)	99.39(↓9.22)	58.92(↑4.37)	94.70(↑6.81)
Rank	71.97(↓26.37)	87.54(↓7.88)	29.97(↓8.5)	85.73(↓9.35)	98.66(↑3.97)	96.36(↑11.79)	58.52(↑1.12)	92.58(↑1.01)	86.86(↓13.14)	94.56(↓5.16)	68.35(↓29.92)	93.33(↓6.54)	92.34(↑3.24)	92.49(↑7.32)	65.19(↑10.64)	87.16(↓0.73)
Log-Rank	83.37(↓14.97)	98.75(↓3.33)	32.40(↓6.07)	91.19(↓3.89)	95.07(↓0.38)	98.89(↑14.32)	47.05(↓10.35)	88.34(↑5.97)	99.57(↓0.43)	99.96(↓0.24)	34.09(↑14.8)	99.03(↓0.84)	87.92(↓1.18)	99.31(↓9.14)	67.70(↑13.15)	95.02(↑7.13)
Log-Likelihood	81.29(↓14.97)	98.30(↓2.88)	32.32(↓6.15)	90.15(↓4.93)	94.08(↓0.61)	98.20(↓13.63)	47.32(↓10.08)	86.77(↑4.4)	99.37(↓0.63)	99.94(↓0.22)	93.79(↑4.48)	98.70(↑1.17)	88.18(↓0.92)	99.14(↓9.97)	70.36(↑15.81)	94.73(↑6.84)
DetectGPT	53.71(↓44.63)	42.39(↓53.03)	31.05(↑42.76)	57.51(↓37.57)	63.15(↓31.54)	54.10(↓30.47)	32.55(↓24.85)	86.05(↑3.68)	78.68(↓21.32)	73.45(↓26.27)	47.36(↓50.91)	89.30(↓10.57)	87.74(↓1.36)	80.95(↓9.22)	43.36(↑11.19)	97.07(↑9.18)
Fast-DetectGPT	96.23(↑2.11)	87.79(↓7.63)	46.32(↑7.85)	69.19(↓25.9)	90.38(↓4.31)	77.55(↓7.02)	71.91(↑14.51)	58.81(↓23.56)	99.89(↓0.11)	99.65(↓0.07)	99.22(↓0.05)	98.59(↓1.28)	89.63(↓0.53)	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)	99.86(↑1.52)	99.07(↑3.65)	**99.17**(↑60.7)	99.27(↑4.19)	99.99(↑5.3)	99.87(↑15.3)	**99.45**(↑42.05)	99.76(↑17.39)	99.97(↓0.03)	99.87(↑0.15)	98.51(↑0.24)	99.93(↑0.06)	99.98(↑10.88)	99.98(↓9.81)	99.28(↑44.73)	99.99(↑12.1)
DivScore (law)	99.86(↑1.52)	99.61(↑4.19)	98.65(↑60.18)	99.54(↑4.46)	99.99(↑5.3)	99.81(↑15.24)	99.25(↑41.85)	99.87(↑17.5)	99.97(↓0.03)	99.81(↑0.09)	**98.61**(↑0.34)	99.94(↑0.07)	99.94(↑10.84)	99.93(↓9.76)	99.37(↑44.82)	99.99(↑12.1)
DivScore	**99.87**(↑1.53)	**99.90**(↑4.48)	98.88(↑60.41)	**99.84**(↑4.76)	**100.00**(↑5.31)	**99.96**(↑15.39)	99.20(↑41.8)	**99.87**(↑17.5)	99.99(↓0.01)	**99.99**(↑0.27)	98.54(↑0.27)	**99.99**(↑0.12)	**99.99**(↑10.89)	**100.00**(↓9.83)	**99.56**(↑45.01)	**100.00**(↑12.11)
TPR @ 0.1% FPR																
RoB-base	0.00(↓68.4)	35.40(↑30.3)	0.10(↓0.4)	0.00(↓27.7)	0.00(↓14.3)	62.60(↑62.6)	0.10(↓0.1)	0.10(↓0.3)	0.00(↓99.7)	36.90(↑1.8)	0.00(↓74.5)	0.20(↓84.8)	0.04(↓8.26)	59.20(↑59.2)	0.20(↑0.2)	0.10(↓0.2)
RoB-large	9.80(↓58.6)	26.70(↑21.6)	1.90(↑1.4)	0.70(↓27.0)	5.08(↓9.22)	78.80(↑78.8)	0.90(↑0.7)	0.30(↓0.1)	0.94(↓98.76)	23.20(↓11.9)	0.10(↓74.4)	0.30(↓84.7)	0.62(↓7.68)	82.30(↑82.3)	2.80(↑2.8)	1.30(↑1.0)
Entropy	0.00(↓68.4)	28.80(↑23.7)	0.10(↓0.4)	0.00(↓27.7)	0.60(↓13.7)	10.20(↑10.2)	0.00(↓0.2)	0.00(↓0.4)	5.90(↓93.8)	89.20(↑54.1)	4.40(↓70.1)	0.00(↓85.0)	0.00(↓8.3)	37.10(↑37.1)	0.10(↑0.1)	0.00(↓0.3)
Rank	13.50(↓54.9)	9.80(↑4.7)	0.00(↓0.5)	5.10(↓22.6)	34.30(↑20.0)	14.20(↑14.2)	0.00(↓0.2)	5.30(↑4.9)	33.30(↓66.4)	28.90(↓6.2)	2.80(↓71.7)	19.50(↓65.5)	14.40(↑6.1)	3.20(↓3.2)	0.00(↓10)	0.40(↓0.1)
Log-Rank	0.90(↓67.5)	59.90(↑54.8)	0.10(↓0.4)	0.00(↓27.7)	4.30(↓10.0)	48.70(↑48.7)	0.00(↓0.2)	0.00(↓0.4)	66.30(↓33.4)	98.00(↑62.9)	4.40(↓70.1)	0.00(↓85.0)	1.10(↑1.2)	63.20(↓63.2)	0.10(↑0.1)	0.00(↓0.3)
Log-Likelihood	0.20(↓68.2)	42.00(↑36.9)	0.10(↓0.4)	0.00(↓27.7)	1.50(↓12.8)	23.40(↑23.4)	0.00(↓0.2)	0.00(↓0.4)	40.20(↓59.5)	96.90(↑61.8)	4.40(↓70.1)	0.00(↓85.0)	0.30(↓8.0)	48.60(↑48.6)	0.10(↑0.1)	0.00(↓0.3)
DetectGPT	0.00(↓68.4)	0.00(↓5.1)	0.00(↓0.5)	0.00(↓27.7)	0.00(↓14.3)	0.00(↑0)	0.00(↓0.2)	0.20(↓0.2)	0.00(↓99.7)	0.00(↓35.1)	0.00(↓74.5)	0.10(↓84.9)	0.00(↓8.3)	0.00(↑0)	0.00(↑0)	2.70(↑2.4)
Fast-DetectGPT	42.20(↓26.2)	7.30(↑2.2)	0.90(↑0.4)	1.70(↓26.0)	12.50(↓1.8)	1.50(↑1.5)	0.80(↑0.6)	0.00(↓0.4)	92.70(↓7.0)	82.20(↑47.1)	73.80(↓0.7)	60.90(↓24.1)	7.70(↓0.6)	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**(↑25.7)	71.80(↑66.7)	**81.00**(↑80.5)	72.00(↑44.3)	99.20(↑84.9)	95.90(↑95.9)	**71.10**(↑70.9)	83.70(↑83.3)	98.30(↑1.4)	91.10(↑56.0)	77.00(↑2.5)	94.70(↑9.7)	97.60(↑89.3)	98.70(↑98.7)	88.30(↑88.3)	99.20(↑98.9)
DivScore (law)	91.20(↑22.8)	51.40(↑46.3)	66.60(↑66.1)	80.20(↑52.5)	97.80(↑83.5)	83.90(↑83.9)	48.10(↑47.9)	93.50(↑93.1)	97.70(↓2.0)	76.20(↑41.1)	78.00(↑3.5)	97.30(↑12.3)	94.20(↑85.9)	92.60(↑92.6)	89.70(↑89.7)	99.70(↑99.4)
DivScore	90.40(↑22.0)	94.50(↑89.4)	69.60(↑69.1)	88.00(↑60.3)	99.60(↑85.3)	99.50(↑99.5)	42.80(↑42.6)	96.50(↑96.1)	98.80(↓0.9)	99.10(↑64.0)	77.30(↑2.8)	98.10(↑13.1)	88.80(↑80.5)	99.90(↑99.9)	89.90(↑89.9)	99.90(↑99.6)

Table 1: Main Experiment - Detection AUROC score and TPR (at 0.1% FPR) of baseline methods and DivScore on medical (MIMIC, PubMedQA) and legal datasets (OALC, LawStack). Methods are benchmarked against the Binoculars detector, with arrows indicating performance differences (↑/↓) relative to Binoculars. **Bold** indicates the best result, and underline denotes the second-best.

constraints mirror practical situations in which neither in-domain reference data nor the identity of the generating LLM is available.

DivScore Detector DivScore employs Mistral-7B-v0.2 (Jiang et al., 2023) as the general LLM for text scoring and domain adaptation. The domain-adapted LLM is developed through a process of domain knowledge distillation, by finetuning on the Mistral-7B-v0.2. A combined knowledge distillation in both the medical and legal domains is conducted to optimize detector's capability within these specialized fields. In the main experiment, we additionally evaluate two variants of DivScore that undergo partial knowledge distillation in either the medical or legal domain alone, referred to as DivScore (med) and DivScore (law).

Domain Knowledge Distillation

We select DeepSeek-R1 (DeepSeek-AI, 2025) as the teacher LLM due to its strong medical and legal expertise (Tordjman et al., 2025; Yu et al., 2025) and its permissive MIT license, which allows full knowledge distillation. Medical seed knowledge is sourced from 22,000 GPT-4o-generated, verifiable medical questions by Chen et al. (2024). Legal seed knowledge is drawn from: (1) EQUALS (Chen et al., 2023), with 6,909 legal questions from professional law consultations; (2) Open Legal QA (Butler, 2023), with 2,124 GPT-4-synthesized questions based on Australian legal cases; and (3) Pile-of-Law (Henderson et al., 2022), with 15,393 legal questions from Reddit forums. Distillation knowledge is then generated by DeepSeek-R1 in the form of question answering. The resulting outputs are used to finetune the detector's general LLM with Low Rank Adaptation (LoRA). Details for seed knowledge construction and LoRA fine-tuning are provided in Appendix B.3.

Datasets We select four representative datasets in medical and legal domain to evaluate DivScore and baselines: (1) MIMIC-IV-Note (MIMIC) (Johnson et al., 2023), deidentified hospital notes for patients, (2) PubMedQA (Jin et al., 2019), questions and answers for biomedical research papers collected from PubMed, (3) Open Australia Legal Corpus (OALC) (Butler, 2023), a collection of Australian legislative and judicial documents, (4) LawStack (Moslem, 2025), legal questions and answers collect from Stack Exchange. Within each dataset, we select 1,000 pairs of texts, each consisting of a human-written text and its LLM-generated counterparts. The source LLMs for LLM-generated text are chosen from two model families: GPT-4o (OpenAI, 2024), O3-mini (OpenAI, 2025), DeepSeek-V3 (DeepSeek-AI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025), representing current SOTA LLMs available to the public. Detailed settings are provided in Appendix B.1.

Baseline Detectors We benchmark DivScore against a comprehensive set of zero-shot and supervised detection methods. For zero-shot detectors, we include Log-Likelihood (Solaiman et al., 2019), Entropy (Lavergne et al., 2008),

Datasets	MIMIC			PubMedQA			OALC			LawStack		
Attacks	w/o	Paraphrase	Substitute	w/o	Paraphrase	Substitute	w/o	Paraphrase	Substitute	w/o	Paraphrase	Substitute
AUROC												
RoB-base	66.79(↓31.55)	47.08(↓40.06)	55.01(↓21.28)	85.79(↓9.63)	68.59(↓17.15)	66.86(↓8.16)	58.33(↑19.86)	41.92(↑4.25)	40.87(↑23.97)	65.10(↓29.98)	51.62(↓35.65)	49.83(↓14.62)
RoB-large	93.05(↓5.29)	65.76(↓21.38)	60.94(↓15.35)	91.69(↓3.73)	78.72(↓7.02)	60.31(↓14.71)	<u>81.23</u>(↑42.76)	<u>59.50</u>(↑21.83)	<u>45.04</u>(↑28.14)	72.36(↓22.72)	48.53(↓38.74)	35.68(↓28.77)
Entropy	68.52(↓29.82)	9.84(↓77.30)	23.88(↓52.41)	98.29(↑2.87)	82.56(↓3.18)	<u>75.96</u>(↑0.94)	33.30(↓5.17)	10.28(↓27.39)	9.73(↓7.17)	91.85(↓3.23)	61.45(↓25.82)	61.77(↓2.68)
Rank	71.97(↓26.37)	57.54(↓29.6)	11.82(↓64.47)	87.54(↑7.88)	82.75(↓2.99)	52.59(↓22.43)	29.97(↓8.50)	27.23(↓10.44)	9.65(↓7.25)	85.73(↓9.35)	72.22(↓15.05)	43.18(↓21.27)
Log-Rank	83.37(↓14.97)	18.42(↓68.72)	25.53(↓50.76)	98.75(↓3.33)	<u>90.89</u>(↑5.15)	74.79(↓0.23)	32.40(↓6.07)	12.68(↓24.99)	8.71(↓8.19)	91.19(↓3.89)	67.26(↓20.01)	50.54(↓13.91)
Log-Likelihood	81.29(↓17.05)	14.78(↓72.36)	24.79(↓51.50)	98.30(↑2.88)	88.05(↑2.31)	69.89(↓5.13)	32.32(↓6.15)	11.37(↓26.30)	8.43(↓8.47)	90.15(↓4.93)	65.63(↓21.64)	48.53(↓15.92)
DetectGPT	53.71(↓44.63)	32.92(↓54.22)	34.30(↓41.99)	42.39(↓53.03)	38.82(↓46.92)	22.19(↓52.83)	31.05(↓7.42)	11.73(↓25.94)	9.49(↓7.41)	57.51(↓37.57)	42.18(↓45.09)	26.05(↓38.40)
Fast-DetectGPT	96.24(↓2.10)	**87.82**(↑0.68)	<u>78.01</u>(↑1.72)	87.77(↓7.65)	79.99(↓5.75)	39.75(↓35.27)	46.32(↑7.85)	48.69(↑11.02)	19.30(↑2.40)	69.19(↓25.90)	70.46(↓16.81)	26.39(↓38.06)
Binoculars	<u>98.34</u>	<u>87.14</u>	76.29	95.42	85.74	75.02	38.47	37.67	16.90	<u>95.08</u>	<u>87.27</u>	<u>64.45</u>
DivScore	**99.88**(↑1.54)	78.74(↓8.40)	**97.03**(↑20.74)	**99.64**(↑4.22)	**93.85**(↑8.11)	**95.21**(↑20.19)	**98.88**(↑60.41)	**88.07**(↑50.40)	**94.63**(↑77.73)	**99.84**(↑4.76)	**91.29**(↑4.02)	**94.09**(↑29.64)
TPR @ 0.1% FPR												
RoB-base	0.00(↓68.40)	0.00(↓11.30)	0.00(↓8.10)	<u>35.40</u>(↑30.30)	<u>14.20</u>(↑13.80)	<u>14.50</u>(↑14.20)	0.10(↓0.40)	0.00(↓0.30)	0.00(↑0.00)	0.00(↓27.70)	0.00(↓4.60)	0.00(↓0.60)
RoB-large	9.80(↓58.60)	0.70(↓10.60)	0.00(↓8.10)	26.70(↑21.60)	7.80(↑7.40)	2.60(↑2.30)	<u>1.90</u>(↑1.40)	<u>0.50</u>(↑0.20)	<u>0.40</u>(↑0.40)	0.70(↓27.00)	0.20(↓4.40)	0.00(↓0.60)
Entropy	0.00(↓68.40)	0.00(↓11.30)	0.00(↓8.10)	28.80(↑23.70)	0.40(↑0.00)	0.10(↓0.20)	0.10(↓0.40)	0.00(↓0.30)	0.00(↑0.00)	0.00(↓27.70)	0.00(↓4.60)	0.00(↓0.60)
Rank	13.50(↓54.90)	0.90(↓10.40)	0.00(↓8.10)	9.80(↑4.70)	1.50(↑1.10)	0.00(↓0.30)	0.00(↓0.50)	0.00(↓0.30)	0.00(↑0.00)	5.10(↓22.60)	0.40(↓4.20)	0.00(↓0.60)
Log-Rank	0.90(↓67.50)	0.00(↓11.30)	0.00(↓8.10)	59.90(↑54.80)	5.60(↑5.20)	0.80(↑0.50)	0.10(↓0.40)	0.00(↓0.30)	0.00(↑0.00)	0.00(↓27.70)	0.00(↓4.60)	0.00(↓0.60)
Log-Likelihood	0.20(↓68.20)	0.00(↓11.30)	0.00(↓8.10)	42.00(↑36.90)	1.10(↑0.70)	0.00(↓0.30)	0.10(↓0.40)	0.00(↓0.30)	0.00(↑0.00)	0.00(↓27.70)	0.00(↓4.60)	0.00(↓0.60)
DetectGPT	0.00(↓68.40)	0.00(↓11.30)	0.00(↓8.10)	0.00(↓5.10)	0.00(↓0.40)	0.00(↓0.30)	0.00(↓0.50)	0.00(↓0.30)	0.00(↑0.00)	0.00(↓27.70)	0.00(↓4.60)	0.00(↓0.60)
Fast-DetectGPT	41.80(↓26.60)	5.90(↓5.40)	4.60(↓3.50)	6.90(↑1.80)	2.10(↑1.70)	0.00(↓0.30)	0.90(↑0.40)	0.00(↓0.30)	0.00(↑0.00)	1.70(↓26.00)	1.50(↓3.10)	0.00(↓0.60)
Binoculars	<u>68.40</u>	<u>11.30</u>	<u>8.10</u>	5.10	0.40	0.30	0.50	0.30	0.00	<u>27.70</u>	<u>4.60</u>	<u>0.60</u>
DivScore	**96.40**(↑28.00)	**12.80**(↑1.50)	**43.40**(↑35.30)	**85.80**(↑80.70)	**28.80**(↑28.40)	**23.00**(↑22.70)	**69.60**(↑69.10)	**3.40**(↑3.10)	**11.20**(↑11.20)	**88.00**(↑60.30)	**6.10**(↑1.50)	**12.30**(↑11.70)

Table 2: Detector Robustness - Detection AUROC score and TPR (0.1% FPR) for baselines and DivScore on GPT-4o-generated datasets with and without Adversarial Attacks (Paraphrase, Word Substitution). "w/o" indicates datasets without attacks. Methods are benchmarked against the Binoculars (SOTA), with arrows indicating performance differences (↑/↓) relative to Binoculars. **Bold** indicates the best result, and <u>underline</u> denotes the second-best.

Rank, Log-Rank (Gehrmann et al., 2019), Detect-GPT (Mitchell et al., 2023), Fast-DetectGPT (Bao et al., 2023), and Binoculars (Hans et al., 2024). For supervised detectors, we evaluate two RoBERTa-based classifiers (RoB-base and RoB-large) released by OpenAI (Solaiman et al., 2019). In the zero-shot experiment setting, supervised detectors are assessed without further training on the current domain detection tasks. Detailed baseline definitions are provided in Appendix B.2.

Evaluation Metrics Previous studies (Su et al., 2023; Yu et al., 2024; Wu et al., 2025b) mainly use AUROC to assess detector performance. However, AUROC, as an aggregate metric, may not reflect performance in low false positive regimes, which is crucial in specialized domains where misclassifying human-written texts as LLM-generated can be highly harmful. Therefore, in addition to AUROC, we report the True Positive Rate (TPR) at a False Positive Rate (FPR) of 0.1%, following the evaluation protocol of Hans et al. (2024). This metric reflects the TPR when, out of 1,000 human-written texts, at most one false positive is permitted.

4.2 Main Results

Table 1 shows the detection performance of baselines and DivScore on four medical and legal datasets generated by GPT-4o, O3-mini, DeepSeek-V3 and DeepSeek-R1, respectively. As shown, DivScore and its variants with partial knowledge distillation—DivScore (med) and DivScore

(law)—demonstrate substantial performance gains over the Binoculars benchmark. For DivScore (med), it yields on average a 14.3% gain in AUROC and a 62.1% gain in TPR. For DivScore (law), it achieves a similar performance with 14.3% averaged gain in AUROC and 57.4% in TPR. The full DivScore, enhanced by distillation from both medical and legal domains, further pushes the average improvement to 14.4% in AUROC and 64.0% in TPR. These results underscore DivScore's ability to detect specialized LLM-generated content with few false positives. Additionally, improvements of DivScore over its domain-specific variants indicate its generalizability across different specialized domains, showing its potential in cross-domain specialized LLM-generated text detection. Appendix C.1 and Figure 6 provide a more detailed analysis of detection performance and the true vs. false positive trade-off for all detectors.

4.3 Detector Robustness

Recent studies show that paraphrasing (Sadasivan et al., 2023; Krishna et al., 2023) and word-level perturbations (Peng et al., 2023; Liang et al., 2023) can effectively evade current AI-text detectors by altering lexical and syntactic features. To assess DivScore's robustness in real-world application, we employ paraphrase (Krishna et al., 2023) and word-level substitution Peng et al. (2023) attacks on the GPT-4o texts. As shown in Table 2, both attacks substantially reduce the performance of zero-

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

Table 3: Statistics (mean ± standard deviation) of DivScore and its components: cross-entropy, entropy for each dataset. Statistics are computed across four text source LLMs, with and without domain knowledge distillation. The "Avg." row shows dataset-averaged values.
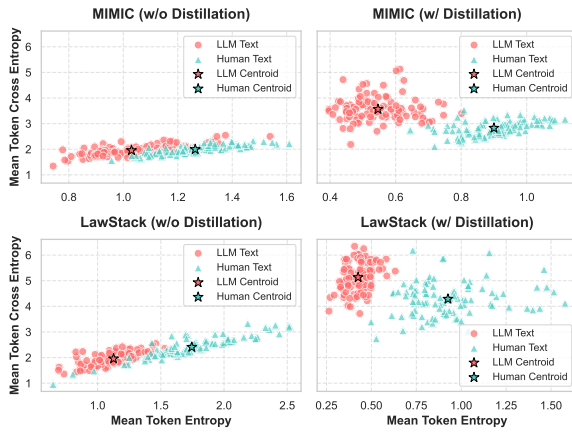

Figure 3: Text statistical signatures in DivScore for MIMIC and LawStack datasets with and without domain knowledge distillation. The data consists randomly selected 100 human-written and 100 GPT-4o-generated texts. Knowledge distillation greatly enlarges the cluster gap thereby enhances DivScore's performance.

shot and supervised baselines. In contrast, DivScore remains robust, achieving on average 22.8% higher in AUROC and 29.5% higher in TPR than Binoculars. This robustness is likely due to the proposed knowledge distillation, which aligns the detector with LLM knowledge in specialized domains. Since adversarial attacks must preserve semantic and domain-specific content for usability, DivScore can still effectively recognize LLM-generated text, maintaining high detection accuracy even under strong perturbations. Detailed settings are provided in Appendix B.4.

4.4 Knowledge Distillation Analysis

In this section, we examine how domain knowledge distillation influences the statistical signatures used by DivScore. We analyse the statistics of DivScore and its components - mean token entropy and mean

token cross entropy with and without distillation. Mistral-7B-v0.2-Instruct, finetuned only on general instructions, serves as $\mathcal{M}^*$ for the "without distillation" baseline. As shown in Figure 3, knowledge distillation enhances the separability between LLM-generated and human-written texts, which is essential for accurate detection. Table 3 additionally reports the mean and standard deviation of these signatures for both text types, giving the following findings.

(1) Cross-Entropy & Entropy. Knowledge distillation reduces the standard deviation of entropy (by 56.3% for LLM texts and 27.6% for human texts), making both clusters more compact. It also increases the centroid (mean) distance for cross-entropy by 36.2%, enlarging the gap between LLM and human text clusters. These results, as shown in Table 3, confirm that knowledge distillation improves the separability of LLM- and human-generated texts. Cross-entropy statistics further support the hypothesis presented in Section 3.2, namely that LLM-generated texts yield higher cross-entropy due to the disagreement between the domain-adapted LLM and the general-domain LLM.

(2) Detector Score. The proposed detector score, being the ratio of the above statistical signatures, shares a collective impact from knowledge distillation. As shown in Table 3, distillation leads to an overall 33.3% increase in inter-group centroid (mean) distance, 66.7% decrease in LLM score std and 40% decrease in human score std. These results demonstrate that domain knowledge distillation enhances the distinguishability of LLM-generated texts by increasing the gap between LLM and human texts (centroid distance) and reducing intra-group variation (standard deviation).

4.5 Ablation Studies

Ablation on Domain Knowledge Distillation. In Section 3.3, we analysed the motivation for incorporating LLM domain knowledge distillation to improve DivScore's performance. We further empirically assess its impact on detection. Specifically, we compare DivScore with: (1) three variants where the domain adaptation for general LLM $\mathcal{M}$ is constructed by finetuning on human-written texts in medical, legal domain, and their combination, and (2) the baseline DivScore without domain adaptation, as referred in Section 4.4.

Figure 4 presents the AUROC averaged across all datasets and text source LLMs. Incorporating
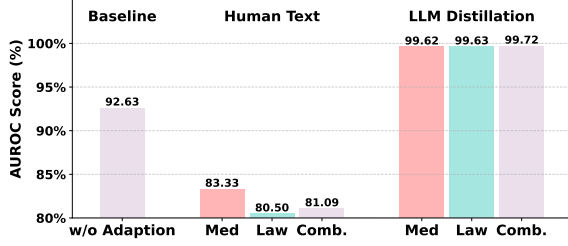
7

Figure 4: Comparison of detector AUROC scores across three settings: baseline (w/o domain adaptation), human domain text finetuning, and LLM domain knowledge distillation. For the later two settings, three variants are included: adaptation to medical, legal domain and their combination. AUROC scores are averaged across all evaluation datasets.
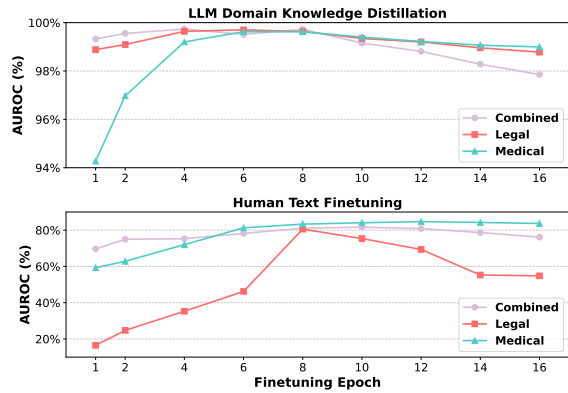

Figure 5: Detector performance (AUROC %) of LLM distillation (proposed) and human text finetuning across finetuning epochs. AUROC scores are averaged across all evaluation datasets.

LLM domain knowledge distillation significantly enhances detection performance, with respective AUROC improvements of 18.6% and 7.1% compared to DivScore variants adapted from human text corpus and the baseline. These findings suggest that LLM distillation is highly effective for constructing domain-adapted LLMs for DivScore, outperforming baseline and human text finetuning. We further investigated the impact of the number of finetuning epochs on the performance of each variant. As illustrated in Figure 5, all variants reach their peak or near-peak AUROC at epoch 8, after which performance declines. This observation supports our choice of finetuning epochs and ensures fair comparisons in the ablation studies. Intuitively, insufficient training leads to inadequate domain adaptation, while excessive finetuning may lead to forgetting of LLM linguistic characteristics in general domain, reducing detection effectiveness.

Ablation on Detector Components. We further

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 4: Ablation results (AUROC %) for DivScore and its scoring variants across four detector LLM architectures, showing that the proposed design consistently outperforms entropy or cross-entropy alone. AUROC scores are averaged across all text source LLMs.

validated DivScore's design for two key detector components: the choice of LLM architecture and the effectiveness of the scoring mechanism. We compared four open-source LLMs: Falcon-7B (Almazrouei et al., 2023), Qwen2.5-8B (Qwen, 2025), Llama3-8B (Grattafiori et al., 2024), and the proposed LLM, Mistral-7B-v0.2. We also evaluated each scoring component by comparing the overall detector score with its components: token entropy and token cross-entropy. The detector performance, averaged across all datasets and source LLMs (Table 4), yields three main findings: (1) DivScore achieves robust detection (Avg. AUROC > 95) across all LLM architectures, with Mistral-7B performing best; (2) the full detector scoring mechanism outperforms its components, confirming the necessity of the DivScore design. See detailed ablation settings in Appendix B.5.

5 Conclusions

In this work, we introduce DivScore, a theoretically grounded, zero-shot detector for identifying LLM-generated text in specialized and high-stakes domains such as medicine and law. Our analysis reveals that existing general zero-shot detectors suffer in these settings due to a distributional mismatch. To address this, DivScore leverages a normalized entropy-based scoring mechanism, augmented by domain knowledge distillation, to robustly detect LLM-generated specialized content. Empirical results on medical and legal domains show that DivScore consistently surpasses SOTA baselines, while remaining robust to adversarial attacks.

8

6 Limitations

We identify the following limitations of our work: (1) Domain Adaptation Requirement. Our approach requires tailored adaptation to target specialized domains. While the proposed detector demonstrates generalizability beyond a single domain, effective adaptation remains essential for high detection performance. This adaptation process may pose practical challenges, particularly in low-resource settings. Future work could focus on improving the efficiency and scalability of domain adaptation. (2) Language Scope. Like most existing studies, our work focuses solely on detecting machine-generated text in English. The performance of DivScore in other languages, especially non-English and low-resource languages, remains unexamined. Expanding detection capabilities across languages—particularly in specialized and high-stakes contexts—is a critical direction for future research, including comprehensive benchmarking and adaptation strategies.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, 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 Søgaard. 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 Trhlík, 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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 François 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

10

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 GECScore. 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*.

A Proof of Theorem

We start the analysis with a standard entropy-based detector d_H, which utilizes an LLM (with generation distribution Q) to compute text entropy scores for classification. Given input medical text X, the detector's performance can be formalized as follows:

Theorem 1. *Given a zero-shot LLM-generated text detector D with detection score mapped from the detector LLM's text likelihood $p_Q(X)$ by any strictly monotonic function $f : \mathbb{R} \to \mathbb{R}$:*

$$D(X) = f(p_Q(X)), \tag{7}$$

where Q represents the text distribution calculated from D. The performance of D, denoted as $\mathrm{AUROC}(D(X))$, satisfies:

$$\mathrm{AUROC}(D(X)) \propto^+ \delta_{KL}, \tag{8}$$

where $\propto^+$ denotes positive monotonic association, and δ_{KL}, effective KL divergence, is given as:

$$\delta_{KL} = D_{KL}(P \parallel Q') - D_{KL}(Q \parallel Q'), \tag{9}$$

with Q' indicating text distribution from the real source. P denotes text distribution from human. $D_{KL}(\cdot \parallel \cdot)$ denotes Kullback–Leibler(KL) divergence.

Proof. We begin the proof by considering the LLM entropy detector $d_H(X)$ as a special case of $D(X)$. Detector d_H classifies sampled text x as machine-generated if its entropy under Q falls below a threshold τ:

$$d_H(x) = \begin{cases} 1 & \text{if } H_Q(x) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $H_Q(x)$ denotes the LLM entropy of text x sampled from X, based on LLM distribution Q, given by:

$$H_Q(x) = \sum_{x_i} p_Q(x_i|x_{<i}) \log p_Q(x_i|x_{<i}) \tag{11}$$

Distributional Assumptions:

Assume the text variable X's LLM entropy values follow normal distributions:

$$\begin{aligned} H_Q(X) &\sim \mathcal{N}(\mu_{Q'}, \sigma_{Q'}^2) && \text{if } X \sim Q' \\ H_Q(X) &\sim \mathcal{N}(\mu_P, \sigma_P^2) && \text{if } X \sim P \end{aligned} \tag{12}$$

with means derived from cross-entropy among detector distribution Q, human distribution P and LLM text generation distribution Q':

$$\begin{aligned} \mu_Q &= H(Q) + D_{\mathrm{KL}}(Q \parallel Q') \\ \mu_P &= H(P) + D_{\mathrm{KL}}(P \parallel Q') \end{aligned} \tag{13}$$

ROC Curve Construction:

The True Positive Rate (TPR) and False Positive Rate (FPR) are defined as:

$$\text{TPR} = \mathbb{P}_{X \sim Q'}(H_Q(X) < \tau)$$
$$\text{FPR} = \mathbb{P}_{X \sim P}(H_Q(X) < \tau) \tag{14}$$

given Equation 12,

$$\text{TPR}(\tau) = \Phi\left(\frac{\tau - \mu_Q}{\sigma_{Q'}}\right)$$
$$\text{FPR}(\tau) = \Phi\left(\frac{\tau - \mu_P}{\sigma_P}\right) \tag{15}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

AUROC Derivation:

The AUROC is the integral of TPR over FPR, denote FPR as p for simplicity:

$$\text{AUROC} = \int_0^1 \Phi\left(\Phi^{-1}(p)\frac{\sigma_P}{\sigma_{Q'}} + \frac{\delta_{KL}}{\sigma_{Q'}}\right) dp \tag{16}$$

where $\delta_{KL} = \mu_P - \mu_Q = D_{\text{KL}}(P \parallel Q') - D_{\text{KL}}(Q \parallel Q')$, given Equation 13.

Let $z = \Phi^{-1}(p)$, then $dp = \phi(z)dz$, where $\phi(z)$ is the Cumulative Density Function (CDF) of standard normal distribution. Equation 16 simplifies to:

$$\text{AUROC} = \int_{-\infty}^{\infty} \Phi\left(\frac{\sigma_P}{\sigma_{Q'}}z + \frac{\delta_{KL}}{\sigma_{Q'}}\right) \phi(z)dz$$
$$= \Phi\left(\frac{\delta_{KL}}{\sqrt{\sigma_P^2 + \sigma_{Q'}^2}}\right) \tag{17}$$

Since the CDF of normal distribution is monotonically increasing, then, for detector d_H, we have:

$$\text{AUROC}(d_H(X)) \propto^+ \delta_{KL} \tag{18}$$

We further extend such conclusion to the complete definition of $D(X)$. Let $s(x) = f(p_Q(x))$, where $f : \mathbb{R} \to \mathbb{R}$ is any strictly monotonic function. For any $x_1, x_2 \in \mathcal{X}$:

$$H_Q(x_1) > H_Q(x_2) \iff \begin{cases} s(x_1) > s(x_2) & (f \nearrow) \\ s(x_1) < s(x_2) & (f \searrow) \end{cases} \tag{19}$$

Equation 19 implies $s(\cdot)$ preserves/reverses the exact ranking induced by $H_Q(\cdot)$. For any threshold τ_s for detector s, define equivalent H_Q threshold:

$$\tau_H := H_Q(f^{-1}(\tau_s)) \tag{20}$$

Suppose $f \nearrow$, for any τ_s:

$$\text{TPR}(\tau_s) = \mathbb{P}_{X \sim Q'}(s(X) < \tau_s)$$
$$\text{FPR}(\tau_s) = \mathbb{P}_{X \sim P}(s(X) < \tau_s) \tag{21}$$

According to Equation 20:

$$\text{TPR}(\tau_s) = \mathbb{P}_{X \sim Q'}(H_Q(X) < \tau_H)$$
$$\text{FPR}(\tau_s) = \mathbb{P}_{X \sim P}(H_Q(X) < \tau_H) \tag{22}$$

Similarly, for $f \searrow$, for any τ_s:

$$\text{TPR}(\tau_s) = \mathbb{P}_{X \sim Q'}(H_Q(X) < \tau_H)$$
$$\text{FPR}(\tau_s) = \mathbb{P}_{X \sim P}(H_Q(X) < \tau_H) \tag{23}$$

The ROC curves are therefore identical for detector $s(X)$ and $H_Q(X)$ in either settings, giving:

$$\text{AUROC}(D(X)) = \text{AUROC}(d_H(X)) \tag{24}$$

Therefore, we have

$$\text{AUROC}(D(X)) \propto^+ \delta_{KL}. \tag{25}$$

$\square$

B Experimental Settings

This section describes the experimental configurations for our main evaluation, adversarial attacks, and ablation studies. Table 5 provides detailed specifications of the large language models used throughout our tests. All experiments were performed on a Linux server with 8 NVIDIA H100-80G GPUs.

B.1 Text Detection Datasets

In this study, we focus on detecting LLM-generated text within written English. All detection texts are originally composed in English. The datasets are organized in two formats: (1) Documents: The MIMIC (Johnson et al., 2023) and OALC (Butler, 2025) datasets consist of independent, human-written medical and legal documents; (2) QA pairs: The PubMedQA (Jin et al., 2019) and Law-Stack (Moslem, 2025) datasets contain paired, human-written questions and answers. To compile our test set, we generate 1,000 samples for each dataset and each source model (GPT-4o, O3-mini, DeepSeek-V3, DeepSeek-R1) by calling API service.

For the document datasets, we generate corresponding LLM-generated texts by using the first one-third of each document's tokens as a prompt and instructing the text source LLMs to continue

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 5: Details of LLMs used in the experiments, including machine-generated text source models, adversarial attack models and detector models

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

Table 6: Prompt Template for MIMIC Dataset

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

Table 8: Prompt Template for PubMedQA and Law-Stack Dataset

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

Table 7: Prompt Template for OALC Dataset

the writing. To ensure a fair comparison, the human-written texts are also truncated, retaining only the remaining two-thirds of the original document. For the QA pair datasets, the human text consists of the human-authored answers, while the LLM-generated text is produced by prompting the questions to the text source LLMs. Specifically, we initiate the text generation process by sending the following prompt to the LLM API service, as shown in Table 6, Table 7 and Table 8.

B.2 Baseline Detectors

In this section, we present the benchmark detectors and their configurations used in the main experiments. For zero-shot detectors, classification scores are obtained by extracting the logits from the detector's surrogate language model (and its variants). Classical zero-shot methods—including Log-Likelihood, Rank, Log-Rank, and Entropy—employ Mistral-7B-v0.2 to ensure consistency with the DivScore detector LLM setup. Recent zero-shot detectors, such as DetectGPT, Fast-DetectGPT, and Binoculars, are implemented using the original LLM configurations as specified in their respective publications. For supervised detectors, we utilize the pre-trained models released by the original authors and apply them directly for out-of-distribution (OOD) text detection in our main experiments. Detailed descriptions of the baseline methods are provided below:

Log-Likelihood This zero-shot method uses a language model to compute the log-probability of

each token in a given text (Solaiman et al., 2019). Specifically, the averaged log-probability across all tokens is used as the detection score. A higher score corresponds to an increased probability of the text being generated by LLM.

Rank This zero-shot method utilizes the absolute ranking of tokens as determined by a language model (Gehrmann et al., 2019). By computing the mean rank score, a higher average value indicates an increased probability that the text is machine-generated.

Log-Rank An enhanced version of Rank method. This zero-shot method applies logarithmic transition to each token's absolute rank (Gehrmann et al., 2019). A higher mean log-rank value indicates a higher likelihood that the text is machine-generated.

Entropy This zero-shot approach leverages a language model to calculate the mean entropy score across all tokens in a given text (Lavergne et al., 2008). Since human-written content tends to exhibit greater unpredictability from the perspective of an LLM, a lower mean entropy score suggests a higher likelihood of the text being machine-generated.

DetectGPT A zero-shot method using probabilistic curvature for detection (Mitchell et al., 2023). The approach involves perturbing text samples and analysing the resulting changes in a pretrained language model's log-likelihood. The underlying theory suggests that LLM-generated texts tend to occupy local optima in the LLM's log probability space, causing perturbations to mostly decrease their probability scores. In contrast, human-written texts does not follow such property. In our experiments, DetectGPT utilizes its default optimal setting in the literature: GPT-Neo-2.7B (Black et al., 2021) as the surrogate model and T5-3B (Raffel et al., 2020) as mask-filling model.

Fast-DetectGPT Bao et al. (2023) proposed a new zero-shot approach to address issue of high computational cost in DetectGPT. It employs an LLM sampling method to replace the perturbation step of DetectGPT. Leveraging conditional probability curvature as metric, Fast-DetectGPT

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

Table 9: Prompt Template for Medical Seed Knowledge

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

Table 10: Prompt Template for Legal Seed Knowledge

identifies the word-choice pattern differences in between LLM and human text. We chose the optimal settings reported in the literature, using GPT-Neo-2.7B as the scoring model and GPT-J-6B (Wang and Komatsuzaki, 2021) as the reference model.

Binoculars Hans et al. (2024) proposes a zero-shot detection method that leverages a pair of LLMs to compute the ratio of a text's perplexity to its cross-perplexity. This perplexity ratio measures how one model reacts to another's token-level predictions, exploiting systematic differences in LLM and human-generated text. A lower score indicates higher likelihood of machine-generated origin, as LLM-produced text tends to be less surprising "relative to the baseline perplexity of an LLM acting on the same string" (Hans et al., 2024). In the main experiments, we chose the optimal detector settings reported in the literature, using Flacon-7B (Almazrouei et al., 2023), Flacon-7B-Instruct (Almazrouei et al., 2023) as observer and performer model.

OpenAI RoBERTa Detectors A set of popular and competitive supervised detectors with RoBERTa architecture. Solaiman et al. (2019) introduced a labelled dataset comprising GPT-2-generated outputs and WebText samples for

14

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

Table 11: Prompt Template for EQUALS legal question translation

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

Table 12: Domain Adaptation: Distillation Knowledge Base Summary. (M: Million Tokens)

supervised training on LLM-generated text detection. Leveraging this dataset, they finetuned two classifiers based on RoBERTa-base and RoBERTa-large model, achieving over 95% accuracy on GPT-2 text detection. In the main experiments, we adopt the original model weights released by OpenAI and evaluate their OOD detection performance on LLM-generated medical and legal texts without additional training.

B.3 LLM Knowledge Distillation

Domain Seed Knowledge To facilitate detector domain adoption, seed knowledge serves as a catalyst, prompting the teacher LLM to generate more elaborate and detailed domain knowledge based on its initial information. The prompt of seed knowledge is formulated in QA format and sent to DeepSeek-R1 through its official API service.

In this work, we focus on knowledge distillation using written English text. All seed knowledge datasets, except for EQUALS, are originally composed in English, and all prompted responses from DeepSeek-R1 are generated in English. For the EQUALS (Chen et al., 2023) dataset, whose original legal questions are in Chinese, we utilize DeepSeek-V3 to translate the questions into English prior to prompting the teacher LLM. Table 11 shows the prompt for translation.

Distillation Knowledge base We construct the distillation knowledge base using the response content generated by the teacher LLM (DeepSeek-R1). Since the reasoning content produced by R1 reflects the model's internal self-reasoning rather than professional expert output, it is not utilized for detector domain adaptation. A summary of the medical and legal distillation knowledge bases is provided in Table 12.

Student Learning We enforce the student The LoRA hyperparameters were configured as LoRA

rank: 64 and alpha scaling factor: 128. Supervised finetuning selectively updates the self-attention projection matrices (query, key, value) and feed-forward network components (gate, up projections) within the Mistral architecture. The finetuning is conducted with a bfloat16 parameter precision with number of training epoch set as 8, learning rate set as $2e^{-4}$, batch size set as 12 and maximum sequence length set as 2048.

B.4 Detector Robustness

To assess DivScore's robustness in real-world scenarios, we apply two types of adversarial attacks to the dataset. For the paraphrase attack, we utilize the DIPPER paraphraser introduced by Krishna et al. (2023), an 11B-parameter T5-based LLM finetuned for paragraph-level paraphrasing with controllable perturbation strength, determined by lexical diversity and content reordering. DIPPER receives the candidate text along with these two parameters as input and generates paraphrased outputs at the specified perturbation level. In our experiments, both the lexical diversity and content reordering factors are set to 20%, representing a moderate attack intensity.

For the word substitution attack, we follow the protocol of Peng et al. (2023), selecting the most frequent words in the text for semantically coherent replacement using a BERT-base model (Devlin et al., 2019). Specifically, the BERT model operates at its maximum sequence length (512 tokens), replacing the top 20% most frequent words within a 50-token context window. Common stop-words, which lack substantial semantic content, are excluded from substitution.

15

Knowledge Base	Seed Knowledge Composition	Tokens Size
Human (med)	iCliniq (Li et al., 2023), MedAlpaca (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 13: Domain Adaptation: Human Corpus Summary (M: Million Tokens)

B.5 Ablation Study

In the ablation study on domain knowledge distillation, we benchmark the proposed domain adaptation approach with human text finetuning. For human text finetuning, we select medical and legal corpus to conduct the same LoRA finetuning as knowledge distillation with detailed settings referred in Section B.3. As shown in Table 13, human corpus consists of text in medical and legal domain matching the distillation knowledge base in token size. The baseline detectors refers to the same setting in Section 4.4 where we employ Mistral-7B-v0.2-Instruct as $\mathcal{M}^*$ to represent the "without distillation" scenario where the detector failed in gaining specialized knowledge. In the ablation study on detector components, details of LLM architectures are listed in Table 4. All detectors have the same settings in domain knowledge distillation.

C Additional Results

This section records the additional results in the experiments.

C.1 Main Experiment

We present the main experiment's additional visualization results in this appendix section. Figure 6 visualizes the AUROC curves of all baseline detectors and the proposed detector in the main experiments. The x-axis (false positive rate) uses a $\log_{10}$ scale to better illustrate detection performance at low false positive rates. This visualization complements the main experiment's metric (TPR at 0.1% FPR) by providing a more comprehensive view of low-FPR performance.

As shown in Figure 6, the proposed detector enhanced with knowledge distillation in medical,

legal domain and their combination achieves the best overall AUROC score in all test cases. Notably, under strict FPR thresholds (below 10^{-3} to 10^{-1}), the performance gaps between our detector and baseline methods become particularly significant. This indicates DivScore's ability to detect LLM-generated profession texts with minimal false positive harm.

C.2 Detection Performance on General Domain Texts

The main focus of this study is to detect LLM-generated content in specialized domains. In this appendix section, we further evaluate the proposed detector's performance in general domain LLM-generated text detection which is more well studied by current research. We choose the HC3 dataset (Guo et al., 2023) for LLM-generated text detection. Specifically, we evaluate three sub-datasets: Reddit posts, Open-QA pairs and Wikipedia content in HC3, covering general LLM-generated text detection scenarios in social media, daily question answering and general encyclopedia content facing the public.

We employs the same settings in the main experiments, randomly selecting 1000 human texts from each datasets and 1000 LLM texts generated with GPT-4o and O3-mini. The prompt is in the QA format same with [xx]. We then benchmark the proposed detector with two state-of-the-art general domain detectors: Fast-DetectGPT (Bao et al., 2023) and Binoculars (Hans et al., 2024). As shown in Table 14, the proposed detector demonstrates strong performance across general domain datasets. Notably, DivScore (law) achieve substantial AUROC and TPR improvements over Binoculars on detecting O3-mini generated texts, with the averaged AUROC (%) gains reaching 19.79 and TPR (%) gains reaching 35.63. For GPT-4o paradigm, the proposed method, while showing slightly lower AUROC scores in Reddit and Wiki dataset, still achieves an overall higher recall at low false positive rates.

These results confirm that DivScore, though tailored for specialized domains, generalizes well beyond specialized domains. It remains highly competitive and often superior to SOTA methods in general domain LLM-generated text detection tasks.
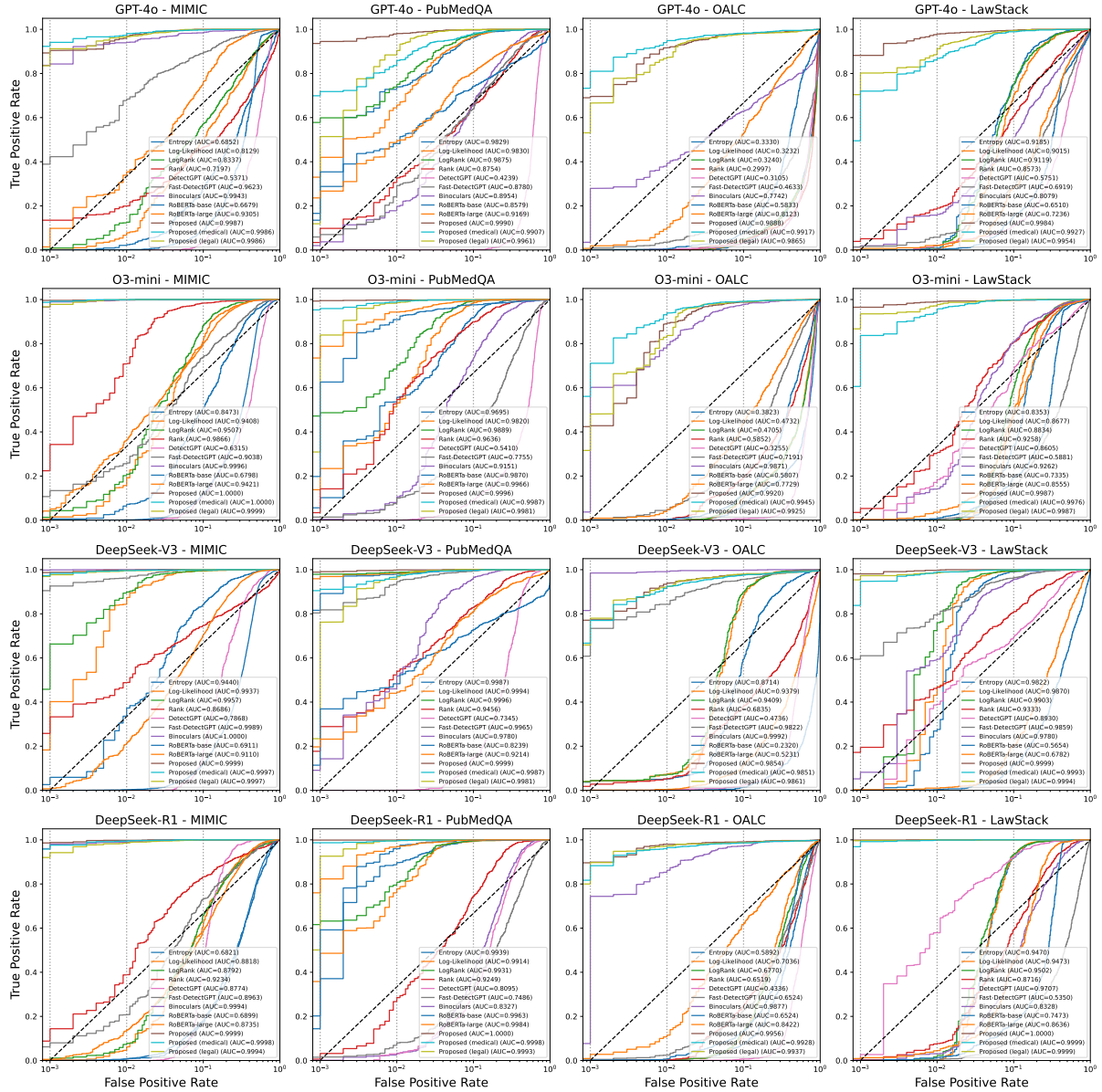
Figure 6: ROC Curves of Baseline Detectors and DivScore in the Main Experiment

C.3 Case Study

This section present four representative case studies, drawn from key experimental datasets — MIMIC, PubMedQA, OALC, and LawStack — with GPT-4o serving as the source LLM. These cases were chosen for their high level of detection difficulty. Notably, for all four cases, state-of-the-art detectors, including Binoculars (Hans et al., 2024) and Fast-DetectGPT (Bao et al., 2023), failed to identify the generated content.

We start from the MIMIC dataset, containing de-identified hospital discharge summaries. The hospital discharge summary typically includes information about the patient's admission, diagnosis, treatment, response to treatment, and discharge

plan. It is a crucial medical document that extends beyond hospital use, playing a pivotal role in commercial sectors such as healthcare insurance by ensuring precise patient care documentation and facilitating efficient claims processing. Table 15 demonstrates the selected MIMIC detection case with GPT-4o as the text source LLM.

In the MIMIC case, GPT-4o generated a continuation of the discharge report using the background information from the first third of the text as the prompt. Enhanced by LLM medical domain knowledge distilled from DeepSeek-R1, the proposed detector created a significant scoring gap between the LLM-generated and authentic human text ($0.103241 \rightarrow 0.275714$), indicating confident

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

Table 14: General Domain LLM-generated Text Detection Performance, AUROC (%) & TPR (at 0.1 % FPR) of selected methods on HC3-reddit, HC3-OpenQA, HC3-Wiki dataset with GPT-4o and O3-mini as text source LLMs. Methods are compared against the Binoculars benchmark, with arrows indicating performance differences (↑/↓) relative to Binoculars.

zero-shot classification. The main contribution to such gap is the Cross-Entropy component, where base and enhanced LLM show great disagreement on LLM-generated text (4.218750) while holding a similar probability estimation on human text (2.734375). This case is expected to be correctly classified with detector threshold set by below 0.1% FPR (i.e., fewer than 1 false-positive case in detection).

Table 16 presents a case from the OALC dataset, with formal legislative and judicial documents from Australia government and courts. Legal documents such as statutes and regulations are characterized by their precise language, formal structure, and strict logical consistency. In this OALC case, the LLM-generated continuation closely mirrors the formal tone and structure of the original court document, with similar structure ans linguistic characteristic, leading to detection failures in Binoculars and Fast-DetectGPT. The proposed detector, leveraging domain-specific legal knowledge distilled from DeepSeek R1, distinguishes the LLM-generated text from the authentic human-written document with a clear scoring gap (0.163043 → 0.084449). The decrease entropy, again plays the key role in the detection (0.820312 → 0.443359). With the detector threshold set at less than 0.1% FPR, both the human and LLM-generated texts are confidently and correctly classified, demonstrating the detector's robustness in handling highly specialized legal content.

Table 17 shows the case selected from Pub-MedQA. PubMedQA contains paired questions and answers extracted from biomedical research papers in the PubMed database. Texts with similar forms of biomedical research QAs play a crucial role in public health promotion, doctor training and evidence-based medical practice. In the Pub-MedQA case, enhanced by LLM medical domain knowledge distilled from DeepSeek-R1, the proposed detector created a notable scoring gap between the LLM-generated and authentic human text (0.272876 → 0.116497). Such gap is attributed to a significant smaller entropy estimation in LLM-generated text (0.535156), compared to human text (1.304688). The entropy gap could be the result of a successful knowledge distillation on LLM's medical knowledge. Enhanced with LLM medical knowledge, the scoring LLM is expected to show lower entropy in LLM-generated answers, as the model tends to produce more predictable and consistent patterns compared to the broader diversity found in human writing. This finding highlights the role of entropy component in the proposed detection framework.

Table 18 shows a case from the LawStack dataset, which contains legal question-answer pairs covering a wide range of jurisdictions and legal topics. Such QA pairs are commonly used in legal consulting, compliance training, and public legal education. In this case, the LLM-generated answer receives nearly one-third of detection score compared with the human-written answer, indicating a confident classification from DivScore. Such

18

gap, as shown in Table 18, is a collective result of both scoring components. With entropy and cross-entropy both shifting as expected, the DivScore detector could successfully classify both LLM and human texts at the most restrictive threshold (FPR below 0.1%).

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: OBSTETRICS/GYNECOLOGY Allergies: Patient recorded as having No Known Allergies to Drugs Attending: ___. Chief Complaint: 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 RUBEOLA 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 fingersticks 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 control. 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 control 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 hypoglycemia 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 hyperglycemia/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. Prognosis: 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 15: MIMIC Case Study: Comparison of human and LLM generated hospital discharge summary with detection scores

OALC: Open Australia Legal Corpus

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 16: OALC Case Study: Comparison of Human and LLM-generated Legal document with detection scores

---

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 17: PubMedQA Case Study: Comparison of Human and LLM-generated answers to biomedical research questions with detection scores

LawStack: Answer to Legal Questions

System Prompt:
Answer this question truthfully:

User Prompt:
I have read the question and answer here: Are parking tickets enforcable 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 18: LawStack Case Study: Comparison of Human and LLM-generated answers to legal issue with detection scores