# 猫和老鼠——虚假文本生成能否超越检测系统？

Andrea McGlinchey[1] and Peter J Barclay[2]

[1] Lumerate, Canada
email: andrea.mcglinchey@lumerate.com
[2] School of Engineering, Computing, and the Built Environment
Edinburgh Napier University, Scotland
email: p.barclay@napier.ac.uk



**摘要** 大型语言模型（LLMs）可以在学术写作、产品评论和政治新闻等领域生成令人信服的“伪造文本”。已经研究了许多方法来检测人工生成的文本。虽然这可能预示着一场无休止的“军备竞赛”，但我们注意到，新出现的 LLMs 使用越来越多的参数、训练数据和能源，而相对简单的分类器在有限资源下展示了良好的检测准确率。为了探讨模型击败检测器的能力是否可能会达到平台期的问题，我们考察了统计分类器识别类似经典侦探小说风格的“伪造文本”的能力。经过 0.5 版本的提升，我们发现 Gemini 生成欺骗性文本的能力有所增强，而 GPT 则没有。这表明即使对于越来越大的模型，可靠地检测伪造文本仍然是可行的，尽管新的模型架构可能会提高它们的欺骗性。

**Keywords:** 大型语言模型，分类系统，假文本检测

## 1 背景

随着复杂文本生成模型的兴起，虚假新闻、社交媒体帖子及其他用于传播不实信息 [4] 的生成文本的数量有所增加。由于大型语言模型（LLMs）被恶意行为者滥用的风险较高，开发能够识别人工智能生成内容和非原创材料的检测器至关重要。

滥用大语言模型已被广泛研究，针对假新闻、学生提交的内容、用户贡献以及医学文本 [2] 的检测方法已经完成；然而相比之下，很少有研究关注创意写作。这一领域的日益重要性最近被作者们所强调，他们抗议 Meta 使用包含数百万本受版权保护书籍的数据库来训练其 AI 模型 [1]。

到目前为止，只有两项研究涉及了这一领域。“Ghostbuster” 研究 [7] 使用了三个数据准备域，其中一个为创意写作，这证明了检测难度稍高一

些。数据集是使用 GPT-3.5-turbo 创建的，并且在创意写作数据集中达到了 98.4%的 F1 分数。“AI Detective” 研究 [5] 仅专注于创意写作，同样使用了当时可用的最佳模型 GPT-3.5-turbo。该系统的 F1 分数范围从 94.2%到 100%，在较短片段上超过了 Ghostbuster 的表现。作者假设随着 LLMs 继续以更多的参数（权重）构建并用越来越多的数据进行训练，检测器的有效性可能会降低。

## 2 问题陈述

随着 LLM 工具的迅速发展，我们可能会发现识别人工生成文本变得越来越困难。这可能导致一种“猫鼠游戏”，其中生成器和检测器都在逐步改进，但都没有明显的优势。

然而，早期的工作如 [5] 和 [8] 已经表明传统的分类器可以达到很好的准确水平，我们注意到这类算法受到的资源限制相对较少。另一方面，大型语言模型则使用了大量的参数，拥有巨大的训练集，并且需要极高的能源需求 [3]。因此我们推测它们欺骗经过良好训练的传统检测器的能力可能会随时间趋于稳定。本文描述了我们早期尝试解决这一研究问题的方法。

> 后期版本的同一大型语言模型是否表现出生成欺骗性文本的能力增强，从而导致检测器的准确性下降？

## 3 方法论

创意小说可以通过大型语言模型从零开始生成，遵循提示生成，或重写人类编写的文章；虽然这两种方法都在引用的研究中有所涉及，在这里我们专注于重写方法以增加不同模型之间的可比性。这种方法对于检测改写的文本也很相关，这些改写的文本可能用于其他形式的抄袭。

数据集使用了六部阿加莎 · 克里斯蒂的公共领域小说创建。这些小说被分割成大约 100 个单词的段落，句点用作分隔符。然后随机化这些摘录并通过 OpenAI 和 Gemini 的 API 进行重写。要求重写的文本长度应与输入文本大致相同。提示已在 OpenAI 上进行了优化，并且为了保持一致性，在 Gemini 中也使用了相同的提示。对于 Open AI，测试了三个模型——最老可用的 GPT 3.5-turbo (GPT 3.5) 模型、针对速度和成本进行优化的下一个

版本 GPT 4o-mini，以及最新可用的 GPT 4.1 模型。在 Gemini 中测试了两个模型，分别是 1.5-flash 模型（Gemini 1.5）和 2.0-flash 模型（Gemini 2.0）。

正如早期工作 [5] 所强调的，很难让大语言模型生成所需长度的文本；人类文本的平均长度为 574，GPT 3.5 达到了 502 的平均值，GPT 4o-mini 达到了 603，最后 GPT 4.1 达到了 581 的平均值。Gemini 模型给出的文本片段要短得多，版本 1.5 的平均长度为 404，版本 2.0 的平均长度为 412。因此，在生成数据集后，对其进行修剪以平衡它们，使其具有大约相同的平均长度、中位数长度和标准差，以便公平地比较人类撰写和 AI 生成的文本。表 1显示了经过平衡的数据集指标。

最后，每个数据集都被随机化并分割为 80%用于训练和验证，保留 20%作为未见过的测试集。对于每次试验，80%的数据被用来创建一个平衡的、随机的人类编写和 AI 生成文本混合数据集，并且这个数据集被用来训练四个机器学习分类器：支持向量机；随机森林；朴素贝叶斯和一个多层感知器分类器。剩下的 20%则用于创建一个平衡的混合测试集，并用于校准每个分类器的准确性。

**表 1.** 比较文本块的统计信息

| 数据集 | 编号行数 | 平均长度 | 中位长度 | 标准差 |
|---|---|---|---|---|
| Original Human Text | 2713 | 574.31 | 578 | 65.80 |
| Balanced Human Text, GPT | 1735 | 579.70 | 580 | 43.33 |
| AI Text, GPT 3.5 | 2713 | 502.21 | 515 | 118.67 |
| Balanced AI Text, GPT 3.5 | 1735 | 571.41 | 563 | 67.57 |
| AI Text, GPT 4o-mini | 2713 | 602.86 | 605 | 78.58 |
| Balanced AI Text, GPT 4o-mini | 1735 | 578.21 | 583 | 40.53 |
| AI Text, GPT 4.1 | 2713 | 580.93 | 583 | 77.79 |
| Balanced AI Text, GPT 4.1 | 1735 | 580.70 | 582 | 70.95 |
| Balanced Human Text, Gemini | 1000 | 509.19 | 523.5 | 47.92 |
| AI Text, Gemini-1.5 | 2713 | 404.07 | 400 | 73.64 |
| Balanced AI Text, Gemini-1.5 | 1000 | 479.97 | 470 | 44.51 |
| AI Text, Gemini-2.0 | 2713 | 411.98 | 410 | 77.19 |
| Balanced AI Text, Gemini-2.0 | 1000 | 490.98 | 480 | 47.41 |

## 4 结果

生成的片段首先通过目测确认与人类撰写文本的相似性。表 2显示了一个示例的人类文本以及每个 LLM 是如何重写的。我们注意到 Gemini 生成的片段通常较短，且重写布局彼此相似。GPT 的重写似乎大多保持了与人类摘录相同的句子结构，并使用同义词来替换词汇。在检查中，GPT 模型的升级版本并没有明显产生不同类型重写；然而，可以看到 Gemini 1.5 和 2.0 之间的改进。

比较表 3中的结果，Gemini 2.0 一直证明是最具欺骗性的——其中的虚假结果大多是被误分类为人类生成的 AI 生成文本——从版本 1.5 到 2.0 的增幅超过 10%。(为了简洁起见，我们仅报告了准确率，这几乎与我们在平衡数据集上的 F1 分数相同)。

出乎意料的是，与 Gemini 相比，GPT 4.1 并没有产生比 3.5 版本更多的欺骗性文本。不那么意外的是，来自 GPT-4o-mini 的文本更容易被识别出来，MLP 分类器达到了近 98%的准确率。这个模型是为了速度和效率而设计的，因此可以预期它在能力上会弱于 GPT 3.5 和 4.1 版本。

为了更直观地比较模型，我们引入了一个新指标欺骗率，定义为 $\mathscr{D} = 1-\text{Recall}$；这是被误分类为人工编写的 AI 生成文本的比例。

每个模型相对于每个分类器的欺骗率显示在表 4中。SVM 分类器整体上表现良好（遭受较低的欺骗率），MLP 分类器对 GPT 模型的表现也很好。

## 5 讨论

我们必须谨慎地直接比较 GPT 和 Gemini 模型，因为它们具有不同的架构；虽然 GPT 和 Gemini 都使用了 Transformer 架构，但 Gemini 采用了专家混合方法，并且进行了优化以支持更长的上下文窗口。一个很好的概述可以在 [6] 中找到。我们对所有与 Gemini 相关的实验都使用了闪存版本，并注意到这些版本比蒸馏后的 GPT-4o-mini 模型更具欺骗性，后者在“规模”上可能大致相似。

总体而言，Gemini 模型在生成类人文本方面更为有效，至少在经典侦探小说领域是如此，尽管我们使用的提示词之前已被优化用于 GPT 模型。我们的结果表明，GPT 模型可能在其生成欺骗性文本的能力上达到了 plateau。

顺便一提，我们注意到 GPT 模型确实提高了生成指定长度文本的能力，而 Gemini 模型在这方面则能力较弱，不同版本之间几乎没有改进。

**表 2.** 不同模型重写同一人类摘录的比较

| 文本准备 | 文本摘录 |
|---|---|
| **原始人类文本** | His face, clean-shaven and exquisitely mobile, was stamped with an expression of power and force far beyond the ordinary. Magnetism seemed to radiate from him. Tuppence was undecided for the moment whether to put him down as an actor or a lawyer, but her doubts were soon solved as he gave her his name: Sir James Peel Edgerton. She looked at him with renewed interest. This, then, was the famous K.C. whose name was familiar all over England. She had heard it said that he might one day be Prime Minister. |
| **模型 GPT 3.5** | Tuppence observed his striking face, impeccably smooth and remarkably expressive, exuding an aura of dominance and intensity that surpassed the norm. Uncertain if he was a thespian or legal professional, her inquiry was swiftly answered when he introduced himself as Sir James Peel Edgerton. Instantly captivated, Tuppence recognized him as the renowned K.C. whose reputation spanned the entirety of England, with rumors suggesting he could potentially ascend to the position of Prime Minister in the future. |
| **模型 GPT 4o-mini** | His visage, impeccably groomed and remarkably expressive, bore an aura of authority and strength that surpassed the mundane. Charisma appeared to emanate from his presence. Tuppence found herself momentarily uncertain whether to classify him as an actor or a barrister, but her uncertainty quickly dissipated when he introduced himself: Sir James Peel Edgerton. Her interest piqued anew. This was indeed the renowned K.C., a name recognized throughout England. Rumors had circulated that he could potentially ascend to the position of Prime Minister. |
| **模型 GPT 4.1** | His face was smooth and expressive, marked by a commanding presence that set him apart from most men. There was a distinct energy about him, almost magnetic in its intensity. At first, Tuppence hesitated, uncertain if he was a member of the stage or the bar, but her uncertainty vanished when he introduced himself: Sir James Peel Edgerton. Instantly, she regarded him with heightened curiosity. This, after all, was the renowned King's Counsel known throughout England. Rumor even suggested he could someday become Prime Minister. |
| **模型 Gemini 1.5** | Undecided, Tuppence eyed Sir James Peel Edgerton. His exquisitely mobile, clean-shaven face radiated power and force, far beyond the ordinary. An actor? A lawyer? His name solved her doubts. Famous K.C., his name echoed throughout England. Prime Minister, some said. She looked at him with renewed interest; this man possessed magnetism. |
| **模型 Gemini 2.0** | The noted K.C. Sir James Peel Edgerton, whose famous name echoed throughout England, faced her. Tuppence hesitated. Was he lawyer or actor? Power and force radiated from him. His clean shaven face, exquisitely mobile, held magnetism beyond the ordinary. One day, she thought, he might even be Prime Minister. This explained the renewed interest she now had. |

**表 3.** LLM 文本生成 – 使用四种模型的准确性比较

| 用于的 LLM | 随机森林 | 支持向量机 | 多层感知机分类器 | 朴素贝叶斯 |
|---|---|---|---|---|
| Gemini 1.5 | 95.25% | 96.25% | 97.00% | 96.25% |
| Gemini 2.0 | 82.25% | 85.50% | 83.00% | 79.75% |
| GPT 3.5 | 89.48% | 92.94% | 94.09% | 94.81% |
| GPT 4o-mini | 92.36% | 96.83% | 97.84% | 97.26% |
| GPT 4.1 | 90.78% | 93.95% | 95.24% | 94.38% |

**表 4.** 使用欺骗率比较 LLM，$\mathscr{D} = 1-$ 召回率

| 使用的语言模型 (LLM) | 随机森林 | 支持向量机 | 多层感知器分类器 | 朴素贝叶斯 |
|---|---|---|---|---|
| Gemini 1.5 | 6.00% | 3.00% | 3.50% | 14.00% |
| Gemini 2.0 | 24.50% | 14.00% | 18.00% | 34.00% |
| GPT 3.5 | 17.29% | 9.22% | 9.22% | 8.93% |
| GPT 4o mini | 11.82% | 3.46% | 2.31% | 4.61% |
| GPT 4.1 Balanced | 14.12% | 6.63% | 6.05% | 9.22% |

尽管模型参数并未公开披露，GPT-4 据信拥有 1–2 万亿个参数（权重），相比之下 GPT-3.5 则有 1570 亿个参数。这意味着，在模型“规模”增加十倍的情况下，其生成文本的能力并没有显著提升。

Gemini 闪存模型是蒸馏模型，基于可能具有与 GPT 模型相当参数数量的 Gemini-pro 模型。假设在 Gemini 1.5 和 2.0 之间参数数量有显著增加，我们看到这确实导致了欺骗检测器的能力增强，准确率下降超过 10%。因此我们认为 Gemini 架构更适合这项任务，并且能够更好地从模型规模的增大中获益。

## 6 结论与进一步工作

我们的结果显示，对于 Gemini 模型而言，版本从 1.5 升级到 2.0 显著提高了生成经典侦探小说风格的可信文本的能力；然而，对于 GPT 模型来说，版本的增加并未显示出明显的改进。这表明 GPT 模型在生成欺骗性文本的能力方面可能已经达到平台期，并且模型架构的重要性可能会超过模型规模。

本研究是首次尝试探讨 LLMs 生成类人文本能力的未来趋势。进一步的研究受到模型参数不公开以及当前模型早期版本不可用的阻碍。另一条研究路线可以是对自托管模型如 NanoGPT 或 LLaMA 变体进行评估，在这些模型中，可以逐步增加参数数量，并将其与击败检测器的能力进行比较。Gemini 未来版本是否能继续提高其欺骗能力仍有待观察；然而，到目前为止的研究结果使我们对认为 LLMs 将越来越擅长生成欺骗性文本这一假设产生怀疑。

## 参考文献


1. Creamer, E. (2025), ‘Meta has stolen books’ : authors to protest in London against AI trained using ‘shadow library’ . https://www.theguardian.com/books/2025/apr/03/meta-has-stolen-books-authors-to-protest-in-london-against-ai-trained-using-shadow-library
2. Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access, 11, 70977-71002
3. Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. arXiv preprint arXiv:2505.09598.
4. Koplin, J. J. (2023). Dual-use implications of AI text generation. Ethics and Information Technology, 25(2), 32. https://doi.org/10.1007/s10676-023-09703-z
5. McGlinchey, A. C. & Barclay, P. J. (2024). Using Machine Learning to Distinguish Human-written from Machine-generated Creative Fiction. arXiv.Org. https://doi.org/10.48550/arXiv.2412.15253
6. Rahman, A., Mahir, S. H., Tashrif, M. T. A., Aishi, A. A., Karim, M. A., Kundu, D., ... & Eidmum, M. D. (2025). Comparative analysis based on deepseek, chatgpt, and google gemini: Features, techniques, performance, future prospects. arXiv preprint arXiv:2503.04783.
7. Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2023). Ghostbuster: Detecting Text Ghostwritten by Large Language Models. arXiv.Org. https://www.proquest.com/docview/2819139768
8. Wahde, M., Della Vedova, M. L., Virgolin, M., & Suvanto, M. (2024). An interpretable method for automated classification of spoken transcripts and written text. Evolutionary Intelligence, 17(1), 609-621.