无名稿件编号（将由编辑插入）

用于学术诚信的人类与机器文本文体计量分析

Hezam Albaqami · Muhammad AsifAyub · Nasir Ahmad · Yaseen Ahmad ·Mohammed M. Alqahtani ·Abdullah M. Algamdi · Almoaid A. Owaidah · Kashif Ahmad

收稿和接收日期应稍后插入

摘要 本研究提出一种基于自然语言处理（NLP）的框架，通过作者归属判定和文体变化检测来验证学生内容的真实性，从而应对学术诚信面临的关键挑战，包括抄袭、伪造以及教育内容的作者身份验证。尽管已有一些初步探索，但该主题的若干方面仍有待研究。在

Hezam Albaqami 计算机科学与人工智能系，计算机科学与工程学院，吉达大学，吉达 21493，沙特阿拉伯。电子邮箱：haalbaqamii@uj.edu.sa Muhammad Asif Ayub 工程与技术大学，白沙瓦，巴基斯坦。电子邮箱：asifayub836@gmail.com Nasir Ahmad 芒斯特理工大学，科克，爱尔兰。电子邮箱：nasir.ahmad@mtu.ie Yaseen Ahmad 工程与技术大学，白沙瓦，巴基斯坦。电子邮箱：ahmadyaseen1628@gmail.com Mohammed M. Alqahtani 计算机科学与人工智能系，计算机科学与工程学院，吉达大学，吉达 21493，沙特阿拉伯。电子邮箱：Mmalqahtani@uj.edu.sa Abdullah M. Algamdi 计算机科学与人工智能系，计算机科学与工程学院，吉达大学，吉达 21493，沙特阿拉伯。电子邮箱：amalgamdi6@uj.edu.sa Almoaid A. Owaidah 管理信息系统系，经济与行政学院，阿卜杜勒阿齐兹国王大学，吉达，吉达 21589，沙特阿拉伯。电子邮箱：aowaidah@kau.edu.sa Kashif Ahmad 芒斯特理工大学，科克，爱尔兰 电子邮箱：kashif.ahmad@mtu.ie

arXiv:2601.01225v1 [cs.CL] 3 Jan 2026

与现有解决方案不同，本文针对四项相关任务对该主题进行了全面分析，包括：（i）人类文本与机器文本的分类；（ii）区分单作者与多作者文档；（iii）检测多作者文档中的作者变更；以及（iv）识别协作生成文档的作者。针对这些任务提出的解决方案，使用 Gemini 通过两种不同提示词生成的两个数据集进行评估，其中包括常规指令集和严格指令集。在实验中，研究人员观察到，使用严格提示生成的数据集时，所提解决方案的性能有所下降，这表明通过精心设计的提示词检测机器生成文本具有一定复杂性。生成的数据集、代码及其他相关材料已在 GitHub 上公开，预计将为该领域未来的研究提供基线[1], 。

关键词 作者归属判定 · 作者检测 · 文体变化检测 ·机器文本检测 · 生成式人工智能 · 文本分类

1 引言

近年来，人们对生成式人工智能（人工智能）及其应用表现出极大的兴趣。因此，人们推出了多种有趣的人工智能内容/文本生成工具，例如 ChatGPT。这些工具能够根据需求，为各种用途准确生成语法正确的文本[12]。例如，用户提出问题后，系统便可即时生成文章、电子邮件、社交媒体帖子、论文，甚至使用不同编程语言编写的代码。这类自动文本生成工具在多个领域都很有用，尤其是在教育领域；与传统学习方式相比，它们能提供多样化的知识来源，且偏见较少。然而，如果使用不当，这些工具可能会严重损害学习成果和教育目标。例如，学生可能依赖这些工具完成作业，却没有充分参与其中，也没有理解问题的实质。学术界广泛讨论了在作业、考试、论文和项目报告中抄袭及不道德使用这些工具的问题，全球范围内也对学术诚信表达了重大关切 [25]。学生过度依赖此类工具，可能会严重影响其批判性思维、问题解决能力、沟通能力（即与教师和同学互动的能力）以及协作能力。同样，这些工具也引发了人们对新闻业 [17], 艺术[29], 和执法等不同应用领域中内容可靠性与真实性日益增长的担忧。例如，此类工具大幅增加了虚假信息和假新闻的生成与传播。这些假新闻和错误信息可能影响政策 [28], 、政府公信力，并引发混乱 [3], ，进而危及公共安全。

令人振奋的是，目前已有一些基于人工智能和自然语言处理（NLP）的技术，可以自动检测通过此类工具生成的内容（文本）；然而，只需稍微修改自动生成的文本（即为文本增添一些人类风格），就能轻易绕过这类算法／工具。有趣的是，也有一些基于人工智能的工具能够使机器生成文本更具人类风格，这让机器生成文本的检测更具挑战性。此

[1] https://github.com/MASIFAYUB/Stylometry-Analysis-of-Human-and-Machine#

外，学术界还面临着作者身份检测这一重要挑战，即核实作业、论文、项目报告和学位论文是否由学生本人撰写，而非由其他人（例如同学或外部专家）代写。这一挑战同样会损害学术诚信和学习成果，影响重大。鉴于基于人工智能的内容生成工具在学术界引发广泛关注，以及教育领域中不道德和欺诈行为所带来的后果，研究界亟须迅速应对这些挑战。

文献中已有一些解决这些问题的初步尝试。然而，现有解决方案大多依赖训练机器学习算法来区分机器生成文本与人类文本，而只需将机器生成文本润色得更像人类所写，就能轻易绕过这些方法。此外，该领域也缺少大规模基准数据集。为克服这些局限，我们提出了一种基于文体计量的解决方案，并构建了一个大规模基准数据集。该解决方案有助于应对学术界在确保学术诚信方面面临的多项挑战。例如，它还可以验证一份文档是否由同一位作者创作，或是否有其他作者（人类或机器）参与创作。在这项工作中，我们处理四项任务，包括：(i) 区分机器生成文本与人类生成文本；(ii) 对单作者文档和多作者文档进行分类；(iii) 找出多作者文档中作者发生变化的部分（段落）；以及 (iv) 识别多作者文档中某段文本的作者。这四项任务在教育环境中至关重要。例如，区分机器生成文本与人类生成文本有助于防止在考核中使用基于人工智能的解决方案。第二项任务可预测教育内容创作过程中是否存在潜在协作（无论合法与否）。最后两项任务则有助于识别由其他作者创作的部分。风格分析主要通过运用内在特征，识别单作者文档或合作创作文档中写作风格的不同模式。文体变化检测（SCD）是风格分析的一项关键应用，可用于内在抄袭检测（IPD） [23]。IPD 不同于外部抄袭检测，后者会将文档与现有来源或参考资料进行比对。相比之下，IPD 旨在仅通过检查输入文档来检测抄袭，判断其中是否有部分内容并非出自同一作者之手，而该作者可能是人类，也可能是人工智能工具。IPD 利用作者在整篇文档中的写作风格一致性来检测抄袭，无需依赖任何其他参考语料库，因此是应对这些挑战的更有效解决方案。风格变化分析并非新课题，但以往研究主要聚焦于人类作者。我们认为，这种细致的分析将为该领域的未来研究提供一个基线。

本研究的主要贡献可总结如下：

– 本文提出了一种基于风格变化分析的新颖自然语言处理框架，旨在应对生成式人工智能时代学术界面临的一些关键挑战。这些任务包括区分单作者文档与多作者文档、对机器生成文本和人类生成文本进行分类，以及识别多作者文档中发生作者变更的段落。

– 本文还分析了精心设计的提示词对基于先进自然语言处理模型的抄袭检测和文档认证工具性能的影响– 本文生成了两组不同的机器生成文本，并将其嵌入人工编写的数据集中，为该领域的后续研究提供了宝贵的资源。数据集和源代码均已公开。– 本文强调了该主题面临的关键挑战，并指出了一些潜在的解决方案。

本文其余部分安排如下：第2节概述相关文献。第3节介绍所提出的方法论。第4节讨论实验设置、所开展的实验以及实验结果。第5节总结全文。

2 相关工作

随着生成式人工智能时代对文档认证和学术诚信的全球关注日益增长，近年来，人们对机器文本生成与分析表现出浓厚兴趣 [15,10]。文献中报告了若干探讨该领域不同方面的有趣研究。机器文本分析的一个关键方面是人类文本与机器文本分类，大多数相关文献都聚焦于自动区分机器生成文本与人类文本 [12]。这项任务通常被视为二元分类问题，并采用各种技术和策略加以处理[32]。例如，Bahad等人 [4] 微调了预训练模型 RoBERTa [22]„ 以对涵盖多个主题的人工智能生成文本和人类生成文本进行分类。Ji等人 [16]则将这项任务视为三元分类问题，新增第三类“未决”文本样本，这些样本可能属于任一类别（即人类或人工智能）。引入第三类旨在凸显模型在文本分类方面的困难与局限，并强调需要向用户提供清晰且有意义的解释。人工智能生成文本由四种最先进的模型生成，包括 GPT-4。0 [1],谷歌的 Gemini Pro [31],LLaMA 3.370B [9], 和 Qwen272B [5]。

这项研究的另一个重要但较少被探索的方面，是所谓的文体计量分析 [24], ，它是文本处理的关键应用之一。Zamir等人 [34], 在关于文档认证和来源追溯的研究中，暗示了文体计量分析的应用，尤其是文体变化检测（SCD）在这一任务中的应用。除了区分机器生成文本和人类文本之外，SCD分析还可以应对学术界在作者身份检测方面面临的其他相关挑战，即验证作业、论文、项目报告和学位论文是否由学生本人撰写，而不是由其他人类（同学或其他专家）代为完成。现有文献已经报告了若干关于SCD和分析的有趣研究，探索了这一概念的不同方面和应用。风格变化分析的关键应用之一是内在抄袭检测（IPD） [23]。与外部抄袭不同，外部抄袭是将文档与现有来源/参考文献进行比对，而IPD旨在仅通过检查输入文档来发现抄袭。随后，它判断输入文档的某些部分是否并非出自同一作者，而该作者可能是人类或基于人工智能的

工具。IPD利用作者在整篇文档中的写作风格一致性进行抄袭检测，而不依赖任何其他参考语料库[26]。SCD也可用于文档认证 [34]。例如，Zamir等人 [33] 使用SCD识别多作者文档以及发生作者变更的文本片段。其他一些研究也使用文体计量分析，对单人撰写的文档和协作生成的文档进行分类 [36,30]。SCD还曾成为基准竞赛PAN的一部分 [18,35], ，并引起了学术界的广泛关注。现有文献已经报告了一些关于文体计量分析的有趣研究；然而，文体计量总体而言，尤其是SCD分析，仍有多个方面有待探索。现有文献大多仅基于人类撰写的文档。近期很少有研究使用风格变化分析，在一般主题上区分人类撰写的文档与机器撰写的文档。例如，Kumar等人 [20] 采用文体计量分析检测Twitter帖子中的机器生成文本，依赖措辞、标点和语言多样性等传统文体计量特征。Opara等人 [27] 也依赖传统文体计量特征，对人类生成文本和机器生成文本进行分类。最近，若干研究探索了基于Transformer的模型用于文体计量分析的可行性，为词语、句子或整篇文档提供上下文嵌入。这些模型已被证明在文体计量分析的不同应用中非常有效，例如作者归属判定 [11],、作者身份检测 [2], 和SCD [33]。

现有文献主要使用机器生成文本探讨第一项任务（即人类文本与机器生成文本的分类）。另外三项任务仅在人类作者的背景下进行过研究。此外，现有文献缺乏包含人类文本与机器生成文本的基准数据集，而该数据集可以通过为各种相关任务提供标注，实现对 SCD 的详细分析。因此，本文通过在两种不同提示词下生成相应的机器文本数据集，分析 SCD 分析中的四项不同任务。

3 方法论

图 1 展示了所提出方法论的框图，整体可分为三个部分：(i) 数据集创建，(ii) 数据预处理，以及 (iii) 文本分类与分析。首先，我们通过智能提示生成机器生成文本，构建了数据集。数据集创建完成后，我们应用了多种预处理技术来处理数据中的各种问题，为使用最先进的自然语言处理算法进行文本分类与分析做好准备。最后，我们采用多种自然语言处理算法，分析并区分单作者和多作者文档中的机器生成文本与人类生成文本。在接下来的小节中，我们将详细介绍各个步骤。



图 1 所提出的方法论。

3.1 数据集创建

在方法论的这一阶段，我们使用 Gemini 模型（gemini-1.5-flash-latest），通过两种不同的提示词生成了大量机器生成文本段落。随后，我们将新生成的人工智能生成文本段落随机插入现有的人类生成数据集中，用于训练和评估针对四项任务的模型。图 2 展示了数据集生成流程图。如图所示，文本是使用多个免费的 API 密钥生成的。

数据生成过程根据现有数据集的结构进行设计，该数据集包含人类作者撰写的文本 [36]。在人类文本数据集中，文本样本按单作者文件和多作者文件进行组织。此外，数据集包含四位作者的文本样本。在我们的研究中，我们将多作者文件视为机器生成文本的基础，并将人工智能/大语言模型（LLM）视为第五位作者来生成机器生成文本。因此，如图所示，对于每个样本，我们首先判断其属于单作者文本还是多作者文本；如果是多作者文件，则继续后续流程。这种方法有助于避免文本生成过程中的重复工作和劳动。需要指出的是，跳过单作者文件并不会影响机器文本与人类文本的分类任务，因为在该实验中，被视为第五位作者的机器生成文本会与人类文本合并。

在多作者文档中，作者人数和段落数量各不相同，有些较少，有些较多。考虑到这一点，并保持机器生成文本与人类文本之间的合理比例，我们根据每份文档的段落数量，添加一个段落或多个段落。因此，对于每份多作者文档，我们检查段落数量（文档中段落的变更列表）是否小于或等于 5；若是，则插入一个段落，若大于 5，则插入几个机器生成段落。成功随机生成并插入机器生成文本后，我们会相应更新作者变更列表，该列表在某些任务中充当真实标注。

对于机器生成文本，我们生成了两种不同的提示：一种是普通提示，另一种是更严格的提示。在普通提示中，我们针对人类文本基准数据集中涵盖的更广泛主题。我们通过建模技术提取主题，得到一份从人类文本中提取的主题列表。在主题建模中，我们使用了三种不同的主题建模算法，即 (i) BERTopic[13], 潜在狄利克雷分配（LDA） [7], 和非负矩阵分解（NMF） [21]。这些算法生成了一份从人类文本中提取的主题列表。我们选取了每种算法提取的前十个主题，然后将其输入 Gemini 模型，通过添加主题的领域和子领域来生成扩展主题。随后，我们将扩展主题作为提示词的一部分输入 Gemini 模型，用于生成机器生成文本。表 1 列出了用于文本生成提示词的主题。在 **严格**提示中，我们没有

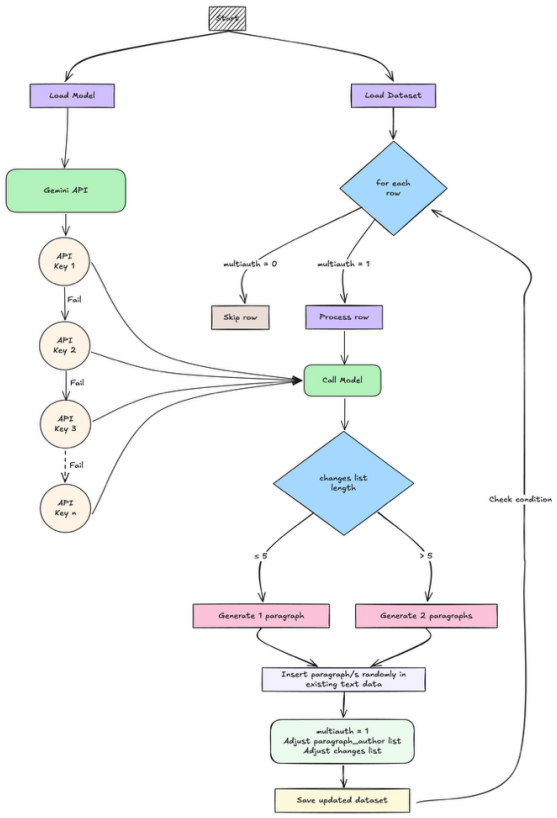

图 2 数据生成过程流程图。两个数据集使用相同的流程图，仅指令集有所不同。

文件系统与数据管理	编程与软件开发
网络与网络安全	云计算与虚拟化技术
人工智能、机器学习与数据科学	硬件与系统性能
软件安装与故障排除	企业信息技术与 DevOps
科技行业趋势	用户体验与人机交互

表 1 提示词中用于文本生成的主题列表。

使用预先定义的主题；相反，模型提供了人工样本，并要求生成相同语境下的文本。两种提示中所使用指令的比较见第 2 节。

3.2 数据预处理

在将数据输入模型之前，我们对其进行了分析和清洗。数据预处理期间执行了多项任务，包括数据清洗、重复项检测与删除，以及数据平衡。首先，通过分析文本和作者信息，检查数据集中的空值和重复条目。随后，使用类别权重进行数据平衡。需要指出的是，在此

普通提示中使用的指令	严格提示中使用的指令
– 生成一个段落，聚焦于以下任一领域中的主题： 文件系统与 数据管理、编程与软件开发 开发、网络与网络安全、云计算与虚拟化 技术、人工智能、机器学习与数据科学、硬件 与系统性能、软件安装与 故障排除、企业信息技术与 DevOps、科技行业 趋势、用户体验与 HC – 请勿添加任何介绍性文字或解释， 前言，或诸如“这是一个新的 段落。 – 确保生成的五个段落各具特色。 – 每个段落应为 40 到 80 个词。 – 涵盖 特征， 原则， 挑战， 采用 趋势及优缺点。 – 使用信息丰富、专业且注重事实准确性的语气。 正确性。 – 避免冗余或含糊的推测。	– 根据单引号中所引文本的上下文， 生成一个新段落，其中若干词语将 与其他所提供段落中的平均词数相 等 – 仅输出该段落；不要包含任何输 介绍性文本、解释、前言或元信息—— 例如“这是一个新段落。”。 – 避免 额外 标签， 标注， or 机器生成的 生成标记。请勿总结 或直接使用所提供段落中的观点。 – 确保该段落与主题相关，且 提出新的视角或见解。 – Do not 包括 介绍性 or 说明性 文本——仅直接生成段落。 – 专注于提供有意义的内容，以补充整体议程或主题。 充整体议程或主题。 – 注意生成段落的长度。 图表，确保其与.....的平均词数相符 其他段落。

表2 提示词中用于生成机器生成文本的指令集与模型的比较。

阶段，我们探索了多种数据平衡技术，包括过采样、欠采样、SMOTE [8]，SMOTEENN [6]，ADASYN [14]，NearMiss、Tomek 和 NearMiss-Tomek。不过，最终采用的方法是通过为少数类分配更高的权重，使其得到更多关注。这些权重被纳入交叉熵损失函数，因此少数类错误会受到更大的惩罚。

3.3 文本分类

对于文本分类，我们使用了四种最先进的模型：BERT（基础模型）、ALBERT、DistilBERT 和 RoBERTa。选择这些模型是基于它们在类似任务中已得到验证的性能 [19]。这些模型的较低层和中间层能够提取较为传统的文体计量特征，例如功能词、标点和句子节奏等。BERT基础模型由 12 个 Transformer 层和注意力头组成，词元长度为 512。DistilBERT 是基础模型的一个变体，它体积更小、速度更快，同时不会显著影响性能。RoBERTa 是基础模型的另一种变体，采用了不同的训练策略，包括使用更大规模的训练集和更长的训练时间，以及移除下一句预测。我们使用不同的丢弃率和注意力丢弃率，在新数据集上对所有模型进行微调，以获得更好的结果。有关超参数的更多详情，请参见第4节。

超参数	值	超参数	值
批次大小	32	梯度累积	2
学习率	0.00001	权重衰减	0.01
训练轮数	5	日志记录	50

表3 实验中使用的超参数汇总。

4 实验与结果

4.1 实验设置

为实现本研究的多方面目标，我们开展了多项实验，探讨该应用的不同方面。这些实验包括：

– **机器生成文本与人类生成文本分类**：这是一项二元分类任务，模型需要预测文档是由人类作者还是机器生成的。– **单作者与多作者文档分类**：这也是一项二元分类任务，模型需要预测文档是由单一作者创作，还是由多位作者共同创作。单作者文档包括人类和机器生成的文本。对于多作者文档，机器生成的文本会随机嵌入由多位人类作者创作的文档中。– **作者变更检测**：这也是一项二元分类任务，旨在识别作者发生变更的连续段落。在多作者文档中，每对连续段落都标记为 0 或 1，分别表示“无变更”和“作者变更”。– **作者识别**：在这项实验中，我们需要识别每篇文档的作者，文档可能由单一作者创作，也可能由多人协作完成。因此，这项任务被视为一项多标签分类任务，共有五位不同的作者，其中包括作为作者之一的机器/人工智能。

在实验过程中，我们采用了各种技术和策略。例如，考虑到计算资源有限，我们使用了梯度累积技术，从而能够使用更大的批次大小。实验中使用的批次大小为 32；然而，累积 2 个步骤的梯度实际上使其达到了 64。学习率初始化为 0.00001，并通过带有 200 个预热步数的余弦调度器逐渐衰减。同样，我们在训练过程中使用了权重衰减，以防止过拟合并提高模型的泛化能力。此外，我们每 500 个步骤进行一次验证，并根据加权 F1 分数保存最佳模型检查点。我们还启用了混合精度（fp16）训练，以加快计算速度并降低内存使用量。我们采用日志记录，每 50 个步骤记录一次指标，并仅保留最近的五个检查点，以监控进度。此外，我们将 70%、15% 和 15% 的样本分别用于训练、验证和测试。

模型	普通提示				严格提示			
	准确率	精确率	召回率	F1 分数	准确率	精确率	召回率	F1 分数
distilbert-base-uncased	0.9998	0.9998	0.9998	0.9998	0.9477	0.9463	0.9477	0.9463
Albert-base-v2	0.9995	0.9995	0.9995	0.9995	0.9456	0.9440	0.9456	0.9441
Roberta-base	0.9997	0.9997	0.9997	0.9997	0.9487	0.9473	0.9487	0.9474
Bert-base-uncased	0.9998	0.9998	0.9998	0.9998	0.9483	0.9469	0.9483	0.9467

表4 两个数据集上机器文本与人类文本分类的实验结果。

4.2 实验结果

4.2.1 机器文本与人类文本分类

表4展示了第一项任务的实验结果，涵盖使用普通提示和严格提示生成的文本，并从准确率、精确率、召回率和 F1 分数等方面进行了评估。总体而言，两个数据集上的结果都非常好，展现了基于 Transformer 的自然语言处理模型区分人类文本和机器生成文本的能力。正如预期的那样，与严格提示相比，在包含通过普通提示生成的人工智能生成文本的数据集上，结果显著更好。这表明，当人工智能文本通过精心设计的提示生成时，将其与人类文本区分开来具有更大的复杂性。关于各个模型的性能，在两个数据集上，各模型的结果（所有指标）均未观察到显著差异。此外，精确率和召回率处于相同范围内，表明模型在正确预测正样本和负样本之间实现了平衡的权衡。我们还分析了这些模型在各个类别上的表现（即对人工智能文本和人类文本进行分类）。我们观察到，在严格提示的情况下，各模型在各个类别上的表现差异比普通提示更大；与人工智能生成文本相比，人类文本的得分更高。

为进一步探究模型在两个数据集上的性能差异，我们使用 Sentence-BERT 嵌入，从类间分离度和类内凝聚度两个方面进行比较语义分析。对于任何分类任务而言，这种基于类间分离度和类内凝聚度的分析都至关重要，尤其是在机器生成数据集的情境下，因为它能揭示各数据集在多大程度上保留了语义组织。使用普通提示词生成的数据集，其嵌入中类别 0（人类文本）与类别 1（机器生成文本）之间的平均余弦相似度非常低（0.047）。在嵌入空间中，这两个类别形成了完全分离的簇。使用严格提示词生成的数据集也呈现出类似趋势，其类间相似度如图3所示。然而，类内凝聚度这一同样会影响数据集分类性能的关键因素，则呈现出不同的趋势。就类内凝聚度而言，我们注意到，在使用严格提示词生成的数据集中，类别 1 的样本彼此具有中等相似度，平均类内凝聚度为 0.14；而类别 0 的样本则相当多样，平均类内凝聚度为 0.03。需要指出的是，类别 0 是人类生成文本，在两个数据集中保持不变。因此，我们仅针对数据集 2 中的机器生成类别计算类内凝聚度；数据集 2 使用普通提示词生成。数据集 2 的类别 1 平均类内凝聚度相对

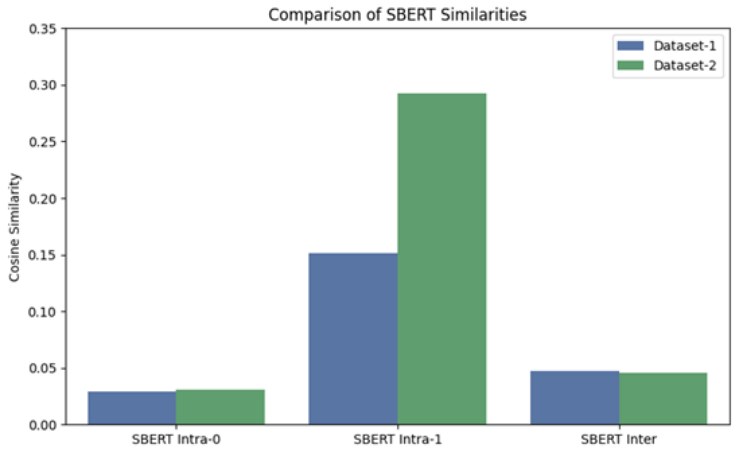

图 3 两个数据集的 SBERT 相似度比较。数据集 1 是使用普通提示生成的数据集，而数据集 2 则包含使用严格提示生成的机器生成文本。此外，类别 0 和类别 1 分别代表人类文本和机器生成文本。

类别/作者	数据集 1（严格提示）			数据集 2（普通提示）		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
类别 0（人类文本）	0.9567	0.9795	0.9780	0.9995	1.00	0.9998
类别 1（机器生成文本）	0.8779	0.7691	0.8199	1.00	0.9971	0.9985

表5 其中一个模型（Albert）的分类得分，展示了两个数据集各个类别的得分，以说明数据集 2 中类别 1 的高类内凝聚度所产生的影响。

较高（0.29），表明机器生成样本更加紧凑、彼此相似。数据集 1 中类别 1 的类内凝聚度较低，主要是因为提示词中的指令要求生成的文本必须严格相似，并且要与每份文档中人类生成文本处于相同语境。

机器生成数据的高类内凝聚度使模型对机器生成文本的分类得分高于使用严格提示生成的文本。两个数据集中各个类别的分类得分也证明了这一点。作为示例，我们在表5中给出了Albert模型的分类结果。

4.2.2 单作者与多作者文档的分类

表6展示了单作者与多作者文档的分类实验结果。可以看出，与使用严格提示词生成的数据集上此前的任务相比，这些模型的表现非常出色。模型得分较高，反映出该任务的复杂度较低；模型能够更准确地区分单作者与多作者文档。如前所述，在本实验中，单作者文档既包含人类生成的文本，也包含机器生成文本；而多作者文档则包含多位人类作者撰写的文本，以及随机嵌入文档不同位置的机器生成文本。这种性能提升

模型	普通提示				严格提示			
	准确率	精确率	召回率	F1 分数	准确率	精确率	召回率	F1 分数
Distilbert-base-uncased	0.9866	0.9869	0.9866	0.9866	0.9845	0.9844	0.9845	0.9844
Albert-base-v2	0.9841	0.9850	0.9841	0.9842	0.9853	0.9854	0.9853	0.9853
Roberta-base	0.9924	0.9925	0.9924	0.9925	0.9824	0.9825	0.9824	0.9822
Bert-base-uncased	0.9857	0.9862	0.9857	0.9858	0.9774	0.9773	0.9774	0.9772

表6 使用常规提示和严格提示进行人工智能生成文本的单作者与多作者文档分类任务的实验结果。

模型	普通提示				严格提示			
	准确率	精确率	召回率	F1 分数	准确率	精确率	召回率	F1 分数
Albert-base-v2	0.674	0.6982	0.674	0.6806	0.6818	0.6964	0.6818	0.6867
Bert-base-uncased	0.5878	0.6795	0.5878	0.5928	0.64	0.6908	0.64	0.6484
Distilbert-base-uncased	0.6699	0.7023	0.6699	0.6774	0.6973	0.7147	0.6973	0.7026
Roberta-base	0.6667	0.6913	0.6667	0.6735	0.6948	0.7149	0.6948	0.7006

表 7 使用常规提示和严格提示生成的人工智能生成文本在多作者文档中的作者变更检测任务上的实验结果。

可能归因于与首个区分人类文本和机器文本的任务相比，人工智能生成文本所占比例相对较低。

4.2.3 多作者文档中的作者变更检测

表 7 展示了作者变更检测任务的实验结果。该任务需要预测后续段落中的作者变更（即判断下一段是否由同一作者撰写）。该任务涉及五位作者，其中四位是人类作者。与之前的任务相比，模型的表现显著下降。使用常规提示和严格提示生成的数据集所取得的最高 F1 分数分别为 0.680 和 0.70。表现较差主要是由于任务本身较为复杂，文献 [33] 也证实了这一点。为进一步探究实验结果，我们分析了各类别得分，并注意到模型在正类上的表现更好，即大多数情况下都能正确预测作者变更。然而，我们也观察到一些类别 0 上的误分类案例，这类样本表示作者没有发生变化。

4.2.4 多作者文档中的作者识别

表 8 展示了最终实验的实验结果，我们在该实验中预测文档的作者。由于一份文档可能由多位作者共同撰写，因此这是一个多标签分类问题。与其他任务相比，两个数据集上的 F1 分数都非常低。计算该任务的 F1 分数时，我们为每个类别使用了最优阈值，因为这些类别的样本数量和分布各不相同。为找到最优阈值，我们采用网格搜索方法，遍历 0.05 到 0.95 之间的不同阈值（步长为 0.01），并计算各阈值下的 F1 分数。类别 1 至 5 使用的最终阈值依次为 0。46，0。48，0。56， 0。51 和 0。9。

另一个有趣的观察是，各模型得出的精确率和召回率存在显著差异。每个模型的精确率都非常

模型	普通提示			严格提示		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
Albert-base-v2	0.3756	0.9106	0.5318	0.3869	0.7433	0.5089
Bert-base-uncased	0.3564	0.9314	0.5155	0.3835	0.7306	0.503
Distilbert-base-uncased	0.355	0.9357	0.5147	0.383	0.7094	0.4974
RoBERTa	0.3561	0.9369	0.516	0.3842	0.7344	0.5045

表 8 普通提示和严格提示生成的人工智能生成文本在多作者文档中的作者识别任务上的实验结果。

类别/作者	普通提示			严格提示		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
作者1	0.5396	0.999	0.70	0.6483	0.7676	0.7029
作者2	0.2674	0.9991	0.4212	0.3005	0.7014	0.4208
作者3	0.1542	0.7297	0.2546	0.2027	0.5624	0.2980
作者4	0.0964	0.0086	0.0158	0.0890	0.5404	0.1528
作者5	0.9990	0.9960	0.9978	0.8516	0.7983	0.8241

表9 Albert模型在两个数据集各类别上的作者识别任务实验结果。

低于召回率，这表明模型过度预测了正样本。在分析过程中，我们注意到模型在某些类别上的表现非常差。作为参考，表9列出了ALBERT-base模型各类别的结果。各项得分都很差，尤其是识别作者2、3和4时的精确率。导致这些类别表现较差的一个潜在原因是这些类别中的样本数量较少。

5 结论与未来工作

本文分析了生成式人工智能时代学术诚信面临的一些关键挑战，并针对抄袭和作者身份验证提出了解决方案。文章探讨了四项挑战，包括区分机器生成文本与人类生成文本、识别文档由单一作者还是多位作者撰写、识别作者发生切换的段落，以及识别多作者文档中每个段落的作者。实验使用了两组嵌入人类生成数据集中的机器生成文本，展示了提示词（用于生成机器文本的一组指令）对基于人工智能的抄袭检测和内容真实性验证工具的影响。实验中，我们观察到模型预测机器生成文本的性能显著下降，这表明精心设计的提示词能够绕过此类工具。基于人工智能的解决方案在区分单作者文档和多作者文档方面表现更佳，这为确保教育内容的真实性带来了很大希望，例如，确保作业由单一作者完成。同样，识别协作撰写的文档中的作者变更和作者身份，也有助于确保学术诚信。然而，这两项任务的实验结果较低，表明它们具有复杂性。

尽管已有一些初步尝试，该领域仍有若干方面有待探索。例如，在教育环境中，该领域缺少大规模基准数据集，无法涵盖项目报告、论文、学位论文及其他作业等各种类型教育内容的不同方面。对此，结合语言、语义、语境及元数据／特征的多模态解决方案可能非常有效，尤其有助于识别协作生成文档中的作者切换和作者身份。此外，通过探索对抗性框架来构建稳健的解决方案，有助于应对精心设计的提示词所生成的人工智能教育内容的检测挑战。

6 作者贡献声明

H. A、K. A. 和 M. A.A. 设计了实验，M.A.A 和 Y.A. 开发了软件，M. A. A. 和 H. A. 开展了实验，N.A.、M. M. A. 和 N.A. 分析了结果。H.A、Y.A、M. M. A. 和 K.A. 撰写了论文初稿。A. M. A.、A. A. O.、N. A、M. M. A. 和 K. A. 审阅了稿件。

7 数据可用性

代码及相关数据已公开发布于 GitHub https://github.com/MASIFAYUB/Stylometry-Analysis-of-Human-and-Machine#，可用于研究用途。

8 利益冲突

作者声明不存在利益冲突。

参考文献

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等：GPT-4技术报告。arXiv预印本 arXiv:2303.08774（2023） 2. Almutairi, A., Kang, B., Al Hashimy, N.：Bibert-av：采用预训练BERT和双向长短期记忆网络的孪生网络增强作者身份验证。载于：普适安全国际会议，第17–30页。Springer（2023） 3. Awwal, A.：考察假新闻和错误信息对新冠疫情危机传播造成的混乱影响。《教育》**2023**（2020） 4. Bahad, S., Bhaskar, Y., Krishnamurthy, P.：微调语言模型以检测人工智能生成文本与人类撰写文本。载于：第18届语义评估国际研讨会论文集（SemEval-2024），第918–921页（2024） 5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., 等：Qwen2.5-VL技术报告。arXiv预印本 arXiv:2502.13923（2025） 6. Batista, G.E., Prati, R.C., Monard, M.C.：多种机器学习训练数据平衡方法的行为研究。《ACM SIGKDD探索通讯》**6**（1），20–29（2004） 7. Chauhan, U., Shah, A.：使用潜在狄利克雷分配进行主题建模：综述。《ACM计算调查》（CSUR）**54**（7），1–35（2021）

8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.：SMOTE：合成少数类过采样技术。《人工智能研究杂志》**16**，321–357（2002） 9. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., 等：Llama 3模型家族。arXiv电子预印本，第arXiv–2407页（2024） 10. Đerić, E., Frank, D., Vuković, D.：探讨在高等教育中使用生成式人工智能工具的伦理影响。载于：《Informatics》，第12卷，第36页。MDPI（2025） 11. Fabien, M., Villatoro-Tello, E., Motlicek, P., Parida, S.：BERTAA：用于作者归属判定的BERT微调。载于：第17届自然语言处理国际会议（ICON）论文集，第127–137页（2020） 12. Fariello, S., Fenza, G., Forte, F., Gallo, M., Marotta, M.：区分人类与机器：人工智能生成文本检测的进展与挑战综述。《国际交互式多媒体与人工智能期刊》**9**（3）（2025） 13. Grootendorst, M.：BERTopic：采用基于类别的TF-IDF方法进行神经主题建模。arXiv预印本 arXiv:2203.05794（2022） 14. He, H., Bai, Y., Garcia, E.A., Li, S.：ADASYN：面向不平衡学习的自适应合成采样方法。载于：2008年IEEE国际神经网络联合会议（IEEE世界计算智能大会），第1322–1328页。IEEE（2008） 15. Iqbal, T., Qureshi, S.：综述：深度学习中的文本生成模型。《沙特国王大学计算机与信息科学杂志》**34**（6），2515–2528（2022） 16. Ji, J., Li, R., Li, S., Guo, J., Qiu, W., Huang, Z., Chen, C., Jiang, X., Lu, X.：检测机器生成文本：不仅仅是“人工智能与人类”的对比，而且可解释性问题十分复杂。arXiv预印本 arXiv:2406.18259（2024） 17. Jones, B., Luger, E., Jones, R.：生成式人工智能与新闻业：基于风险的快速综述。《生成式人工智能与新闻业：快速风险评估》（2023） 18. Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.：PAN-2018作者识别任务概述：跨领域作者归属判定与文体变化检测。载于：CLEF 2018评测实验室工作论文集。法国阿维尼翁，2018年9月10–14日/Cappellato, Linda |编|; 等，第1–25页（2018） 19. Konuma, R., Huaping, Z., Gao, C., Wang, J.：运用BERT微调与文体计量进行日语作者归属判定。载于：智能多语言信息处理：第一届国际会议，IMLIP 2024，中国北京，2024年11月16–17日，论文集，第2395卷，第293页。Springer Nature（2025） 20. Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., Liu, H.：Twitter时间线中人工智能生成文本的文体计量检测。arXiv预印本 arXiv:2303.03697（2023） 21. Lee, D., Seung, H.S.：非负矩阵分解算法。《神经信息处理系统进展》**13**（2000） 22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.：RoBERTa：一种稳健优化的BERT预训练方法。arXiv预印本 arXiv:1907.11692（2019） 23. Manzoor, M.F., Farooq, M.S., Abid, A.：以文体计量为基础的乌尔都语内在抄袭检测框架：运用机器学习、深度学习和大语言模型进行全面分析。《神经计算与应用》，第1–35页（2025） 24. Michailidis, P.D.：文体计量研究领域的科学计量学研究。载于：《Informatics》，第9卷，第60页。MDPI（2022） 25. Mulenga, R., Shilongo, H.：高等教育中的学术诚信：理解并应对抄袭问题。Acta Pedagogia Asiana **3**（1），30–43（2024） 26. Oberreuter, G., Velásquez, J.D.：应用于抄袭检测的文本挖掘：利用词语检测写作风格偏差。《专家系统与应用》**40**（9），3756–3763（2013） 27. Opara, C.：StyloAI：通过文体计量分析区分人工智能生成内容。载于：人工智能教育国际会议，第105–114页。Springer（2024） 28. Silva, E.C.d.M., Vaz, J.C.：错误信息和假新闻如何影响公共政策？国际文献综述。arXiv预印本 arXiv:2406.00951（2024） 29. Škiljić, A.：当艺术遇上技术，或反之：人工智能生成艺术作品与版权法交叉领域的关键挑战。《IIC知识产权与竞争法国际评论》**52**（10），1338–1369（2021） 30. Strøm, E.：通过解决二元分类问题实现多标签文体变化检测。载于：CLEF（工作论文），第2146–2157页（2021）

31. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K. 等：Gemini：高性能多模态模型系列。arXiv 预印本 arXiv:2312.11805（2023） 32. Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L.S., Wong, D.F.：大语言模型生成文本检测综述：必要性、方法与未来方向。计算语言学，第 1–66 页（2025） 33. Zamir, M.T., Ayub, M.A., Gul, A., Ahmad, N., Ahmad, K.：多作者文档的文体计量分析：作者身份与作者风格变化检测。arXiv 预印本 arXiv:2401.06752（2024） 34. Zamir, M.T., Ayub, M.A., Khan, J., Ikram, M.J., Ahmad, N., Ahmad, K.：通过作者身份分类实现文档溯源与真实性验证。载于：2023 年第 1 届智能城市高级创新国际会议（ICAISC），第 1–6 页。IEEE（2023） 35. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.：PAN 2020 风格变化检测任务概述。CLEF（工作笔记）**93**（2020） 36. Zangerle, E., Mayerl, M., Tschuggnall, M., Potthast, M., Stein, B.：PAN21 作者身份分析：风格变化检测（2021）