Modern Linguistics 现代语言学, 2026, 14(6), 530-537
Published Online June 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.146553

Hans汉斯

英语作文自动评价技术发展现状及教学应用启示

黄云龙，曾伊美

江西理工大学外国语学院，江西 赣州

收稿日期：2026年5月6日；录用日期：2026年6月2日；发布日期：2026年6月12日

摘　要

作文自动评价(Automated Essay Evaluation, AEE)研究已历经半个多世纪的发展。十年前，基于神经网络的深度学习技术引入作文自动评价研究，自动评价模型经历了一轮新的快速多元发展。目前，自动评价系统已从最初的浅层特征统计、评分模拟，发展到了能更好“理解”文章深层语义、分析多维度特征并提供有效反馈的较成熟阶段，应用潜力持续提升。本文以研究最为广泛的英语作文自动评价系统为例，通过分析现阶段主流研究中任务目标的多元化发展特征、优点与不足，为实际教育场景了解和利用现有成果、规避现存不足提供参考，以期更好地服务教育实践。

关键词

作文自动评价，发展趋势，应用启示，人机协同

Current Status of Automated Evaluation Technology for English Essay and Its Application Implications for Teaching

Yunlong Huang, Yimei Zeng

School of Foreign Languages, Jiangxi University of Science and Technology, Ganzhou Jiangxi

Received: May 6, 2026; accepted: June 2, 2026; published: June 12, 2026

Abstract

Automated Essay Evaluation (AEE) has been an active area of research for over half a century. Since 2016, the integration of deep learning techniques has further enhanced the performance of AEE

文章引用: 黄云龙, 曾伊美. 英语作文自动评价技术发展现状及教学应用启示[J]. 现代语言学, 2026, 14(6): 530-537.
DOI: 10.12677/ml.2026.146553

models, while the research objectives have become significantly diversified through increased emphasis on evaluation transparency, practical utility, and applicability. Contemporary AEE systems have developed capabilities in "understanding" deeper semantic content, analyzing multi-dimensional linguistic features, and generating constructive feedback. The present paper examines the diversification of task objectives in current AEE researches, along with their respective strengths and limitations. It aims to provide practical insights to better understand and leverage existing advancements, thereby facilitating more effective utilization of cutting-edge research findings in educational practice.

Keywords

Automated Essay Evaluation, Development Trends, Application Implications, Human-Computer Collaboration

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

作文自动评价(Automated Essay Evaluation, AEE)是自然语言处理技术在教育领域的重要应用。20 世纪 60 年代至今，作文自动评价研究经历了基于规则和浅层特征的早期探索阶段、基于统计机器学习和特征工程的中期发展阶段、以及引入深度学习技术后全面深化的现代发展阶段过程，其任务目标也一直在持续变化。上世纪 60 年代，作文自动评价从无到有，探索计算机自动评分的可行性和有效性是其主要目标。上世纪 90 年代至本世纪初，作文自动评价研究在传统统计机器学习和大数据资源的助力下，通过更复杂的特征工程和更先进的机器学习算法来提高模型性能，完成了一轮技术升级。2016 年起，深度学习技术的引入促使作文自动评价进入新一轮技术迭代升级。借助深度学习发展成果，作文自动评估也进一步向更高性能、更细粒度、更多维度评价，以及更适合应用场景的多元化目标方向发展。

2. 作文自动评价技术的发展演进

从评分一致性来看，当前基于深度学习技术模型的性能已达到、甚至可超越人类水平，因此，面向实用的问题得到更多关注，作文自动评价研究的任务目标也表现出了以下新特点：1) 从整体评分走向多特征细粒度评价，2) 从无反馈评价到有反馈评价，3) 从专属任务评价到跨任务评价，4) 从追求性能极致到平衡算力与性能的部署。

2.1. 整体评价 vs. 特征评价

整体评价聚焦对作文质量进行综合评分，是作文自动评价研究关注的基本目标。深度学习时代初期，验证深度学习技术在作文自动评价中的可行性是研究的主要焦点，因此，可靠的整体评分性能仍是此阶段的基本目标。近年来，基于预训练大语言模型和深度学习技术的模型为作文整体评分性能带来了显著提升：ASAP 数据集上的研究表明，多数模型报告的整体评分二次加权卡帕系数(Quadratic Weighted Kappa, QWK)达到人类评分专家水平[1]-[3]。但另一方面，深度学习模型也面临模型评分逻辑与人类专家的对齐问题。深度学习模型参数规模庞大、特征利用机制不够透明，因此评分模型具体在什么程度上依据了哪些特征完成作文评分仍不够清楚，模型可能面临“以不当的理据给出正确的评分”的风险。

特征评价的发展受到模型评分逻辑透明度和真实教育场景需求的共同推动。作文评价涉及多个维度的能力水平，系统评价的分数应该是不同维度能力评价的综合结果。传统特征分析技术主要利用自然语言处理工具、特定语言学规则和数量统计获取的特征，包括表达的正确性、词汇多样性和句法复杂度等。在深度学习模型中，对语言特征的分析和利用主要包括以下方式：1) 直接借用传统特征分析技术，2) 隐式神经网络特征联合学习，3) 基于神经网络的专门特征学习模块。首先，传统特征可作为重要补充特征与神经网络结合使用[2] [4] [5]。而基于神经网络的隐式特征学习主要通过共享表示学习和多任务特征联合学习等方式让模型自动学习对评分有用的特征，无需依赖手工设计特征。独立特征学习模块则通过专门网络或特定学习机制构建独立的特征学习模块，针对性地捕获分析特定类型的特征。当前相关研究现已经取得明显进步，逐渐接近人类水平[6]。但这类模型复杂度往往比较高、对计算资源有更高要求；且在文本宏观结构、修辞等特征的捕获分析能力上仍有不足。综合来看，随着基于深度学习技术的模型性能逐渐趋稳，研究目标也从单纯追求评分准确性逐步扩展到多维度、细粒度的特征评价，体现了技术逻辑的完善，也反映了对教育实践需求的考量。

2.2. 无反馈模型 vs. 反馈模型

无反馈评价仅仅给出综合分数或等级信息，不提供评语或指导性反馈。由于成本和效率的优势，无反馈模型受到早期研究的青睐。单纯评分任务对人工标注训练数据的要求更少，能更高效快捷地促进模型实现、验证、迭代升级和快速部署。不过现实中，学生仅知道分数而不知道如何改进是难以促进写作能力实质提高的。缺乏具体指导也成为无反馈评价系统的短板，不利于培养学生的修改和反思能力和持续写作的动机。

反馈评价模型将评价视为学习过程的重要组成部分，在评分的同时提供作文具体特征的质量信息、鼓励性评语和指导建议。早期的评价系统依靠人工规则提供有关拼写错误、语法错误和标点使用等问题和建议。这种反馈模式规则透明、可解释性强，对语言知识的学习具有帮助。中期阶段，反馈内容从简单的浅层错误发展为涵盖词汇多样性、句法复杂性、篇章连贯性、内容相关性、结构风格等更多维度的文本特征。引入深度学习技术后，系统对特征的利用由传统的显式转为隐式，这对系统提供特征反馈提出了新的挑战[7]。但另一方面，系统对深层语义特征的分析能力进一步增强，性能得到提升[8]。此外，生成式大语言模型(如 GPT、Llama)为模型提供了更强大的反馈生成能力[9]：这类系统不仅能指出不足，还能生成具体的修改建议或示例，大大提高了系统的实用性和指导作用[10] [11]。

2.3. 主题专属评价 vs. 跨主题评价

主题专属评价的目标是针对特定主题训练专门的评分模型。由于专注单一主题，这类模型能够深入学习和捕捉主题相关的特征，对相同或相似主题的作文评价时可达到很高的准确性。另外，由于无需处理不同主题间的差异和复杂关系，主题专属模型的结构往往较为简单，在训练和优化方面也更具优势。但另一方面，现实场景中的作文任务、主题和体裁丰富多样，面向专属任务的自动评价模型往往因泛化能力弱而受到限制，而对每个主题训练专门模型又会进一步增加模型开发的时间成本和经济开销。

跨主题评价模型的主要目标是训练能对不同主题作文进行高效评价的模型。其基本理念是模型训练不依赖所限定主题的训练数据，而是通过对反映文章质量的通用特征学习来完成跨主题作文评价。跨主题评价减少了对数据量的要求，也具有良好的灵活性和更广泛的适用性，可更好地满足写作题目和任务不断变化的实际场景需求。因此，跨主题评价已经成为近年来的重要方向[5] [12]-[14]。不足方面，跨主题模型更强调关注不同作文的通用质量特征，牺牲了对特定主题专属特征的分析，模型可能因此面临与特定主题的评价标准吻合度有限、性能不如专属模型好的问题。

2.4. 追求性能极致 vs. 综合平衡部署

作文自动评价研究一直关注增进对作文的特征表示学习以持续完善模型性能。随着模型的不断升级，模型的复杂度也与日俱增。基于特征工程的研究致力于尽可能多地发掘新的特征。部分系统的特征数量已经超过 2000 个[15]，模型在提高性能的同时也增加了计算负担。另一方面，深度学习评价系统的参数量规模庞大，训练耗时耗力、应用响应延迟，模型开发和部署受到限制。而近年来兴起的多特征评价、多任务学习等性能完善方法又进一步加剧了模型的复杂度。此外，实际教育场景中还可能面临多用户高并发提交等挑战。综合来看，现代评价模型在评分准确性上的确取得了可喜的突破，但同时也带来了显著更高的模型复杂度、计算资源消耗巨大以及实际部署难度和低响应速度等挑战。

综合平衡部署是利用作文自动评价研究成果助力教育场景的基本实践目标。教育应用场景中，模型部署除了需要考虑综合性能外，还需要考虑诸如可行性、稳定性、数据安全、使用便捷度和体验等多方面因素。基于深度学习技术的模型计算负担日益加重，为应对挑战，学界也积极探索不同轻量化、高效率的设计方案，以实现在保持模型性能的同时降低计算复杂度和资源消耗，例如，使用模型蒸馏[16]、修剪[17]、轻量化[1] [11]、参数高效化设计[18]、消费级 GPU 资源本地化部署方法[19]以及高并发条件下的资源调度[20]等优化策略或方案，保障系统在极大降低训练或部署成本的同时不过多损失模型的性能。可以看出，为了平衡实际场景条件制约和高性能模型的计算资源要求，近年来的研究已经不再单纯追求模型的性能极致，而是逐步走向可接受情况下牺牲部分模型性能来换取实用性的平衡方案。

3. 教学应用启示

3.1. 技术现状

进入深度学习时代后，作文自动评价研究的任务目标呈现出多元演进的特点，综合性能也得到明显提升，为实际应用提供了基础条件。性能上看，当前作文自动评价系统已经初步具备应用条件，但也存在一些潜在不足：

1. 模型难以兼顾细粒度语言特征评价，模型存在“以错误的理由得到正确的评分”风险

现有研究多直接依赖基于预训练语言模型的端到端架构实现评分。这类“黑盒”评分模式通过数千亿参数的非线性层将文本直接映射为分值，在超高维的分布式表征空间中，显式的语言学特征被解构融合并高度抽象化。这导致了模型虽然仍可通过海量高度抽象特征不断提升宏观指标(如 QWK)，但却难以再有效与人工专家作文评价所依赖的多层次、细粒度显式语言特征建立明确关联，影响模型决策透明度。另一方面，基于自注意力机制的大语言模型虽善于捕捉表层词汇依赖，但对超越共现的复杂语义理解仍存在不足。而数据驱动的模型训练以最小化损失函数为目标，倾向于选择最易优化的“快捷路径”(如局部情感倾向、文本长度或特定标点分布)，而不对所利用特征的语言学合理性加以区分。上述机制共同作用，削弱了传统评分中对显式语言学特征逻辑的绝对依赖，而将基于语言学特征的评价逻辑与基于其他表层相关性特征评分逻辑置于同等可能的优化路径空间。这导致模型可能将核心评分逻辑建立在简单的表层相关性、而非对内容的深度理解基础上，增加了陷入“以错误的理由得到正确的评分”陷阱的风险。此外，上述机制也极易导致模型评分受到文本长度、特定触发词或针对性的对抗文本的干扰，带来了模型决策的稳定性风险。目前，深度学习模型的决策透明度和稳定性问题已受到一些前沿对抗性研究[7]和可解释研究[21] [22]的关注，有望未来得到进一步完善。

2. 深度学习评价模型对文章整体逻辑、论证深度、修辞手段及内容创意的分析能力薄弱

此类困境的深层根源在于特征本身的高度抽象性和现有模型范式的能力局限。首先，文章的整体逻辑、论证深度、修辞手段和内容创意等本身抽象度更高，依赖深层次的复杂语义逻辑关系网络。然而，

Transformer 架构本质上是通过注意力机制来捕捉序列依赖，随着文本长度增加，宏观的论证结构、修辞等抽象信息在多层空间的传导过程中会更容易被局部语义噪声稀释，导致隐藏的长程语义逻辑依赖关系识别失效。而另一方面，基于概率最大化的模型范式客观上对修辞与创意存在压制。大语言模型的核心优化目标都是最大化下一个词的条件概率，这使得大模型天然倾向于拟合最普遍、最安全的常规表达。而修辞手段(如隐喻、反讽)和内容创意本质上属于对常规语言模式的“有益偏离”，因此，模型往往会倾向于将这些高阶创意表达误判为语义逻辑混乱或低概率噪声。最后，这些高度抽象的特征缺乏规模化、标准高的标注训练数据，依赖纯数据驱动的神经网络极难学习其内在规律，从而限制了模型的泛化能力上限。因此，实际使用中仍需要使用者批判地看待抽象特征的评价结果。

3.2. 教学环境部署潜在障碍与风险

将作文自动评价技术研究推向真实教学环境，不仅是一个技术迁移过程，更是一场涉及教学流程重组、教育生态适应与伦理信任重构的复杂系统工程。在实际部署过程中，可能面临以下现实障碍，需采取针对性的策略予以解决。

1. 教师工作量的变相增加与“技术排斥”心理

引入作文自动评价技术的初衷是减轻教师批改负担，但实际落地中首先需要教师花费时间学习系统操作、理解多维度评价指标的教育测量学含义；其次，由于作为自动评价系统目前在篇章逻辑与创意上的局限性，教师往往需要对系统生成的错误评语或误判分值进行二次校对与修正。这种“人机磨合”在成熟稳定之前可能变相增加教师的认知负荷与时间成本，导致一线教师产生技术排斥心理。

2. 学生群体的接受度、工具依赖与心理防御

学生对作文自动评价技术的接受度可能呈现两极分化的可能：一方面，部分学生对机械的“机器评分”可能产生心理排斥，导致评价的心理效度降低；另一方面，部分学生表现出过度工具依赖，为了“刷高分”而发展出迎合模型偏好的做法(如盲目堆砌高级词汇、拉长篇幅)，从而造成“为迎合算法而写作”的异化现象，偏离思辨能力的培养轨道。

3. 城乡教育信息化水平与数据资源的区域鸿沟

作文自动评价系统的部署高度依赖校园硬件算力、网络带宽以及师生的数字素养。在城市学校，高速网络和智能终端可支撑学生进行高频次的生机交互写作；而在中西部农村或偏远地区，不仅硬件设备受限、日常教学带宽不足，且缺乏专业的教育信息化维护人员。此外，现有的主流模型更多依托城市规范语料进行训练与微调，农村地区学生因方言习惯、文化背景差异导致的非规范表达，极易被模型误判，进而加剧教育实际应用的更深层鸿沟。

4. 人机评价冲突风险与仲裁问题

人机评价冲突是作文自动评价系统实际应用中需要面对的问题。当模型给出的评价与学生的自我预期或教师的人工判断发生显著冲突时(例如一篇极具文学创意的散文可能被模型因“结构不严密、词汇水平不足”而判定为低分)，如果缺乏合理透明的解释机制，极易引发信任危机。因此明确健全的仲裁机制有助于避免让系统和教师陷入信任危机。

3.3. 教学整合的深化路径

作文自动评价系统的技术进展为其教育嵌入提供了可能，但从教育教学的角度思考如何更好地用好当前的技术成果同样具有重要性。结合当前现状，我们认为技术与教学整合的不断深化是实现作文自动评价技术有效服务教育教学的基本路径。

1. 系统嵌入写作学习过程：从“环节辅助”到“流程重构”

从服务实际写作教学方面看，仅将系统视作“批改替代工具”并不能充分释放其教育潜能。作文自动评价系统技术与教学的深度整合，应首先推动作文自动评价从“辅助性工具”向“结构性要素”的转变，充分完善对写作教学过程的重构、人机互动机制的建立与学生写作综合素养的培育。传统写作教学通常遵循“布置–写作–批改–讲评”的线性流程。我们认为，作文自动评价系统的引入不应仅替代“批改”环节，而应推动流程向“迭代化、交互化、数据化”方向发展。充分发挥系统在草稿评价、错误修改、参考建议等重复性、基础性工作方面的优势，设计系统化嵌入写作过程方案，并将教师从繁重的基础性工作中解放出来，让教师聚焦任务设计、反馈把关以及立意深化、逻辑优化和情感表达提升等高层次的指导。具体来说：1) 在预写作阶段，系统可提供基于主题的词汇激活与思维链导引，帮助学生建立内容与思路框架；2) 在草稿与修订阶段，系统支持多轮即时反馈，学生可根据语言、结构等方面的建议进行反复修改，形成“写作–反馈–修改”的迭代循环。教师通过系统汇总的常见问题与进步轨迹，开展针对性课堂讲解；3) 在终稿与反思阶段，系统可生成写作能力发展图谱，突出个体在各项特征上的进步与待改进之处，引导学生进行结构化反思，并帮助教师进行学习成果的整体评估。这一流程重构可有助于提升写作的训练频次与密度，有利于将作文自动评价系统转化为支撑“过程写作教学法”的重要技术基础。

2. 构建基于学习科学的多模态评价生态系统

写作本质上是语言、思维与情感的多维整合活动，单一文本分析难以全面反映学生的写作素养。人机共生的综合生态系统建设应该持续考虑如何更好地基于学习科学构建人机互生的系统应用场景。基于当前的技术实际，具体实践可以通过以下方式给予进一步关注：1) 结合写作过程中的时间序列数据(如修改轨迹、停顿分布)、生理行为数据(如击键记录)等，更精细地刻画学生的写作策略与认知负荷，通过多模态数据融合全方位系统化了解写作学习过程。2) 纵向成长追踪。建立个人写作能力发展档案，可视化呈现学生在语言、结构、思维等维度上的长期进步轨迹，为个性化教学提供证据支持。

3. 结合实际需求建立教师专业发展支持系统

作文自动评价系统有效整合的实施最终取决于教师的意愿、认知能力与驾驭能力，因此，作文自动评价系统的教育应用应充分考虑一线教师实际，并构建面向教师的专业支持体系，具体可包括：1) 完善系统培训支持体系，帮助教师理解作文自动评价的技术逻辑、优势与盲区，掌握反馈筛选与整合策略。2) 教学设计案例库。通过实践积累提供不同学段、不同文体中作文自动评价系统与课堂教学结合的实践案例，降低教师探索成本，启发增效。3) 建立教师间自动评价系统使用经验、问题与教学创意的分享机制，形成实践共同体，推动校本化应用模式创新。4) 针对“人机磨合期”教师的负担问题，需要不断推进模型系统的易用性，例如，可尝试提供“一键导入–自动分析–差异高亮–生成报告”的极简化流程，将教师的操作学习成本降至最低。

4. 面向以学生为中心的评价体系设计

针对学生群体的接受度、工具依赖与心理防御问题，自动评价系统的应用应注重以学生为中心，关注从单一的“写作评价”目标到与“写作素养培育”的更高育人目标结合的发展导向。例如，作文自动评价系统可通过以下方式支持培养学生对写作过程的监控、评估与调节能力：1) 辅助设定个性化进阶目标。基于学生历史表现，系统可识别其最近发展区，并在每次任务中突出重点改进特征，帮助学生聚焦提升。2) 提供策略性反馈而非结论性评分。引导学生关注“如何修改”而非“得分多少”，例如：“尝试将这两个简单句合并为定语从句，以提升句间衔接”，引导学生从被动接受到主动思考。3) 结合写作日志提供反思性提示。系统可自动生成基于写作过程的反思问题(如：“你本次在段落衔接上有意识使用了哪些连接词？效果如何？”)，促进学生将写作经验转化为可迁移的自我调节策略。

5. 持续弥合城乡教育信息化水平和数字资源鸿沟

针对城乡教育信息化水平与地区经济文化背景差异，可以考虑开发“端云协同与轻量化部署架构”。具体来说，研发基于轻量化蒸馏模型(如 TinyBERT、MobileLLM)的离线版或低算力版客户端。核心特征提取在本地终端完成，复杂篇章分析在云端采用异步队列处理，可更好地适应低带宽的农村教学环境。另一方面，强化构建“多元文化包容性语料库”，在模型开发阶段，主动引入一定比例中西部、农村地区的学生作文语料，进行偏置校准，提升模型对地域性语言习惯和多元文化背景表达的泛化能力与包容度。

6. 建立人机协同的决策与冲突仲裁机制

作文自动评价系统透明度不足可能导致教师与学生对其输出结果的怀疑。因此，如何构建可信的人机协作机制是有效深度整合的基础。具体实践中，对于基于深度学习技术的自动评价系统可从以下方面加以完善：1) 可解释反馈设计。系统在提供作为评分依据的特征维度(如：“词汇多样性得分较低，因重复使用‘important’达 5 次”)，并可提供优秀片段作为参照。2) 教师介入修正机制。教师需要关注关键环节(如总分评定、高阶能力评价)评价结果，必要时介入修正机器评价结果，系统方面则可将教师的修正行为反馈至模型优化循环，形成“教学反哺技术”的可持续机制。3) 学生参与式校准。在适当阶段引导学生对比自身感知与系统评价，辨析评分依据的合理性，逐步培养学生对机器反馈的批判性使用能力。4) 针对潜在不一致评价冲突，建立“冲突仲裁与纠偏机制”，确立“教师拥有最终解释权”的最终决定地位，明确作文自动评价系统的“教学辅助”地位，其评分作为参考，但最终官方成绩与终结性评价以教师修正后的结果为准。此外，同时可考虑将教师修正后的文本作为“高价值对抗性样本”反哺至自动评价系统回路，实现系统的增量式自我纠偏。

4. 结语

进入深度学习时代后，作文自动评价在多维度、细粒度和面向教育实践的方面均有长足进步，具备了在低风险写作过程中应用的能力，可有效缓解写作学习中的困境，促进个性化学习。但由于客观限制，现有技术在抽象度高的特征评价上仍有不足，实践中可通过将教师纳入人在环路的设计，更好地保障模型充分发挥对学习者的价值。最后，通过技术研发和教学应用的双向靠拢和迭代升级，作文自动评价系统可以更贴合作文教学的现实需求。

致　谢

本文为江西省赣州市社科规划课题(2024-NDWX27-1018)、江西理工大学高层次人才科研启动课题(205200100646)、江西理工大学外国语学院高层次人才课题(WY2023-BSQDJJ010)支持成果，特此致谢。

参考文献

[1] Aljuaid, H., Alhothali, A., Alzamzami, O., Assalahi, H. and Aldosemani, T. (2025) ET-GNN: Ensemble Transformer-Based Graph Neural Networks for Holistic Automated Essay Scoring. *IEEE Access*, **13**, 58746-58758. https://doi.org/10.1109/access.2025.3556352

[2] Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., *et al*. (2024) Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, **12**, Article 3416. https://doi.org/10.3390/math12213416

[3] Tate, T.P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., *et al*. (2024) Can AI Provide Useful Holistic Essay Scoring? *Computers and Education*: *Artificial Intelligence*, **7**, Article ID: 100255. https://doi.org/10.1016/j.caeai.2024.100255

[4] Li, X., Yang, H., Hu, S., Geng, J., Lin, K. and Li, Y. (2022) Enhanced Hybrid Neural Network for Automated Essay Scoring. *Expert Systems*, **39**, e13068. https://doi.org/10.1111/exsy.13068

[5] Sun, J., Peng, W., Song, T., Liu, H., Zhu, S. and Song, J. (2024) Enhanced Cross-Prompt Trait Scoring via Syntactic Feature Fusion and Contrastive Learning. *The Journal of Supercomputing*, **80**, 5390-5407. https://doi.org/10.1007/s11227-023-05640-2

[6] Wang, J. and Liu, J. (2025) T-MES: Trait-Aware Mix-Of-Experts Representation Learning for Multi-Trait Essay Scoring. *Proceedings of the* 31*st International Conference on Computational Linguistics*, Abu Dhabi, 19-24 January 2025, 1224-1236.

[7] Wang, Y., Hu, R. and Zhao, Z. (2024) Beyond Agreement: Diagnosing the Rationale Alignment of Automated Essay Scoring Methods Based on Linguistically-Informed Counterfactuals. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, 12-16 November 2024, 8906-8925. https://doi.org/10.18653/v1/2024.findings-emnlp.520

[8] Liu, Y., Han, J., Sboev, A. and Makarov, I. (2024) GEEF: A Neural Network Model for Automatic Essay Feedback Generation by Integrating Writing Skills Assessment. *Expert Systems with Applications*, **245**, Article ID: 123043. https://doi.org/10.1016/j.eswa.2023.123043

[9] Han, J., Yoo, H., Myung, J.H., *et al*. (2023) LLM-as-a-Tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. https://doi.org/10.48550/arXiv.2310.05191

[10] Shi, H. and Aryadoust, V. (2024) A Systematic Review of AI-Based Automated Written Feedback Research. *ReCALL*, **36**, 187-209. https://doi.org/10.1017/s0958344023000265

[11] Song, Y., Zhu, Q., Wang, H. and Zheng, Q. (2024) Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Transactions on Learning Technologies*, **17**, 1880-1890. https://doi.org/10.1109/tlt.2024.3396873

[12] Chen, Y. and Li, X. (2024) PLAES: Prompt-Generalized and Level-Aware Learning Framework for Cross-Prompt Automated Essay Scoring. *Proceedings of the Language Resources and Evaluation Conference*, Torino, 20-25 May 2024, 12775-12786. https://doi.org/10.63317/4qnmmm7zj5aq

[13] Wang, J., Zhang, Q., Liu, J., Wang, X., Xu, M., Yang, L., *et al*. (2025) Making Meta-Learning Solve Cross-Prompt Automatic Essay Scoring. *Expert Systems with Applications*, **272**, Article ID: 126710. https://doi.org/10.1016/j.eswa.2025.126710

[14] Li, X. and Pan, W. (2025) KAES: Multi-Aspect Shared Knowledge Finding and Aligning for Cross-Prompt Automated Scoring of Essay Traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 24476-24484. https://doi.org/10.1609/aaai.v39i23.34626

[15] Tang, X., Chen, H., Lin, D. and Li, K. (2024) Incorporating Fine-Grained Linguistic Features and Explainable AI into Multi-Dimensional Automated Writing Assessment. *Applied Sciences*, **14**, Article 4182. https://doi.org/10.3390/app14104182

[16] Utami, N. and Ruskanda, F.Z. (2023) Automated Scoring of English Essays in CEFR Levels Using LSTM and Distil-BERT Embeddings. 2023 10*th International Conference on Advanced Informatics*: *Concept, Theory and Application* (*ICAICTA*), Lombok, 7-9 October 2023, 1-6. https://doi.org/10.1109/icaicta59291.2023.10390038

[17] 张冰雪, 邵小波, 熊振海, 等. 一种结合 BERT 修剪的作文自动评分模型[J]. 软件, 2020, 41(12): 89-94, 106.

[18] Sethi, A. and Singh, K. (2022) Natural Language Processing Based Automated Essay Scoring with Parameter-Efficient Transformer Approach. 2022 6*th International Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 29-31 March 2022, 749-756. https://doi.org/10.1109/iccmc53470.2022.9753760

[19] Ormerod, C. and Kwako, A. (2024) Automated Text Scoring in the Age of Generative AI for the GPU-Poor. *Chinese/English Journal of Educational Measurement and Evaluation*, **5**, Article 5. https://doi.org/10.59863/okuu1904

[20] 程相群. 面向高并发的作文自动评分系统的设计与实现[D]: [硕士学位论文]. 沈阳: 中国科学院大学(中国科学院沈阳计算技术研究所), 2022.

[21] Wang, Y. and Hu, R. (2021) A Prompt-Independent and Interpretable Automated Essay Scoring Method for Chinese Second Language Writing. In: Li, S., *et al*., Eds., *Chinese Computational Linguistics*, Springer, 450-470. https://doi.org/10.1007/978-3-030-84186-7_30

[22] Vanga, R.R., Sindhu, C., Bharath, M.S., Reddy, T.C. and Kanneganti, M. (2023) Autograder: A Feature-Based Quantitative Essay Grading System Using Bert. In: Tuba, M., Akashe, S. and Joshi, A., Eds., *ICT Infrastructure and Computing*, Springer, 71-81. https://doi.org/10.1007/978-981-99-4932-8_8