Position: LLMs Can be Good Tutors in English Education

Jingheng Ye[1,2], **Shen Wang**[1], **Deqing Zou**[2], **Yibo Yan**[1,3],
Kun Wang[1], **Hai-Tao Zheng**[2,4*], **Ruitong Liu**[2], **Zenglin Xu**[5],
Irwin King[6], **Philip S. Yu**[7], **Qingsong Wen**[1*]
[1]Squirrel Ai Learning, [2]Tsinghua University,
[3]HKUST (Guangzhou), [4]Peng Cheng Laboratory, [5]Fudan University,
[6]The Chinese University of Hong Kong, [7]University of Illinois at Chicago
yejh22@mails.tsinghua.edu.cn

Abstract

While recent efforts have begun integrating large language models (LLMs) into English education, they often rely on traditional approaches to learning tasks without fully embracing educational methodologies, thus lacking adaptability to language learning. To address this gap, we argue that **LLMs have the potential to serve as effective tutors in English Education**. Specifically, LLMs can play three critical roles: (1) as *data enhancers*, improving the creation of learning materials or serving as student simulations; (2) as *task predictors*, serving as learner assessment or optimizing learning pathway; and (3) as *agents*, enabling personalized and inclusive education. We encourage interdisciplinary research to explore these roles, fostering innovation while addressing challenges and risks, ultimately advancing English Education through the thoughtful integration of LLMs.

1 Introduction

English Education has long been a cornerstone of global education and a critical component of K-12 curricula, equipping students with the linguistic and cultural competencies necessary for an interconnected world (Alhusaiyan, 2025; Katinskaia, 2025). However, traditional English teaching methods often fall short in addressing the diverse needs of learners (Hou, 2020). Challenges such as limited personalization, scalability constraints, and the lack of real-time feedback are particularly pronounced in large classroom settings (Ehrenberg et al., 2001). Addressing these shortcomings requires innovative approaches that not only enhance the quality of instruction but also adapt to the unique learning trajectories of students (Eaton, 2010).

Recently, LLMs have opened new possibilities for transforming English Education (Caines et al., 2023). LLMs exhibit remarkable natural language understanding and generation capabilities, making them promising candidates for roles traditionally filled by human tutors. Leveraging LLMs as AI tutors can overcome the inherent limitations of conventional teaching methods, offering scalable, interactive, and personalized learning experiences (Chen et al., 2024; Schmucker et al., 2024). Therefore, this position paper argues that **LLMs can be effective tutors in English education, complementing human expertise and addressing key limitations of traditional methods**.

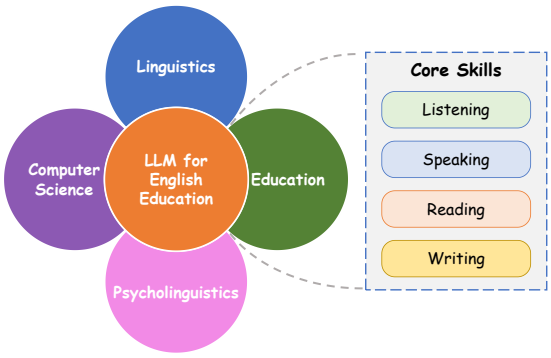

Figure 1: Involved disciplines of LLM for English Edu.

As shown in Figure 1, English Education intersects with multiple *disciplines*, each of which underscores the potential of LLMs to revolutionize this domain. From the perspective of (1) *computer science*, advancements in machine learning and NLP have enabled LLMs to process and generate human-like language at an unprecedented scale; (2) *linguistics* (Radford et al., 2009) contributes a deeper understanding of grammar, phonetics, and semantics, allowing LLMs to generate accurate and understandable language outputs; (3) *education* provides the foundation for designing effective pedagogical strategies, ensuring that LLMs can deliver personalized, engaging, and developmentally appropriate learning experiences; and finally, (4) *psycholinguistics* (Steinberg and Sciarini, 2013) bridges the gap between language acquisition and cognitive processes, enabling LLMs to optimize

*Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17516–17535
November 4-9, 2025

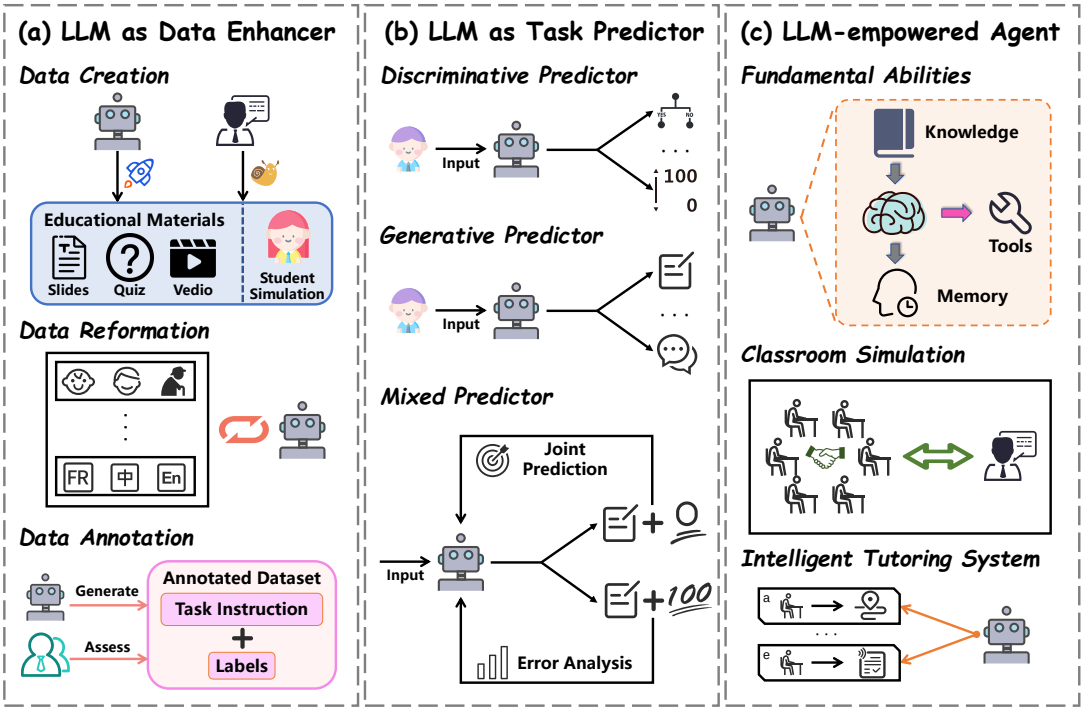

Figure 2: Overview of three roles of LLMs in English education. An overview of related literature is provided in Appendix A.

learner interactions by adapting to individual needs and fostering meaningful engagement. Together, these disciplines position LLMs as uniquely capable of addressing the multifaceted challenges of English education.

Moreover, English Education encompasses four *core skills*: listening, speaking, reading, and writing, each of which can be significantly enhanced by LLMs. For listening, LLMs can generate diverse audio materials (Ghosal et al., 2023) and facilitate interactive voice-based exercises, helping learners improve their ability to discern pronunciation, intonation, and contextual meaning. In speaking, LLMs can simulate realistic conversations (Siyan et al., 2024), provide pronunciation feedback, and scaffold learners' oral communication skills through iterative practice. For reading, LLMs can curate leveled texts, generate comprehension questions (Samuel et al., 2024), and engage learners in discussions that deepen their understanding of written content. Finally, in writing, LLMs can offer real-time grammar, syntax, and style feedback while assisting with idea generation and iterative revisions (Stahl et al., 2024a). By addressing these core skills holistically, LLMs have the potential to deliver a comprehensive and adaptive learning experience.

Despite these opportunities, the deployment of LLMs in English Education must be approached carefully, ensuring that their integration complements rather than replaces human tutors (Jeon and Lee, 2023). As illustrated in Figure 2, this paper explores three critical *roles of LLMs* in this context: their function as ***data enhancers*** (Section 4) to optimize learning materials, their capacity as ***task predictors*** (Section 5) to tailor educational solutions, and their potential as ***agents*** (Section 6) that deliver interactive and adaptive language instruction. By examining these roles, we aim to demonstrate how LLMs can address the limitations of traditional English teaching methods while advancing our understanding of intelligent tutoring systems. Additionally, we discuss potential challenges (Section 7) and future directions (Appendix B) for integrating LLMs into English Education, offering a technical guideline for researchers and educators to harness their transformative potential. We also describe the paradigm shift of leveraging AI for English Education, starting from the last century, as one of our contributions in Section 3.

2 Background

2.1 English Education

Traditional English Education methods often emphasize grammar rules, vocabulary memorization,

and repetitive practice, supplemented by limited opportunities for real-world application (Watzke, 2003). Such approaches are often constrained by the availability of skilled teachers, the diversity of learners' needs, and the lack of personalized feedback (Williams et al., 2004). Recently, many technologies for English Education have been proposed (Alhusaiyan, 2024), focusing on solving specific tasks instead of describing the whole picture of English tutoring. While intelligent language tutoring systems have the potential to create adaptive environments, attention to this field is relatively less compared to other subjects like science (Shao et al., 2025) and mathematics (Ahn et al., 2024). One key reason lies in the inherent complexity of language as an *ill-defined* domain (Schmidt and Strasser, 2022), posing a great challenge in establishing a valid automatic analysis of learner languages due to the vast variability and unpredictability of human language.

2.2 Large Language Models for Education

The potential of LLMs in education (Alhafni et al., 2024), particularly in English Education (Gao et al., 2024; Karataş et al., 2024; Cherednichenko et al., 2024), is immense. Benefiting from large-scale pretraining on extensive corpora, LLMs have demonstrated emergent abilities including (1) *in-context learning* (Dong et al., 2022), which allows the model to adapt to new tasks and provide contextually relevant responses based on a few examples provided during the interaction; (2) *instruction following* (Zeng et al., 2024), which enables the model to process and execute complex user instructions with high accuracy; and (3) *reasoning and planning* (Huang et al., 2024b), which allows the model to generate coherent, structured, and context-aware outputs, even for tasks that require multi-step thinking. However, these fundamental capabilities, while impressive, are insufficient to fully meet the unique demands of English Education. Teaching English requires more than generating grammatically correct sentences or providing accurate translations; it demands a nuanced understanding of pedagogy, learner psychology, and cultural context. Maurya et al. (2024) propose an evaluation taxonomy that identifies eight critical dimensions for assessing AI tutors. These dimensions can be broadly categorized into two groups. (1) *Problem-solving abilities* assess the technical capabilities of LLMs to perform tasks relevant to English Education. (2) *Pedagogical alignment abilities* evaluate how well the LLM aligns with effective teaching and learning principles. Pedagogical alignment includes the model's ability to adapt to the learner's proficiency level, provide scaffolded feedback, foster engagement, and maintain motivation. While LLMs can give direct answers, their ability to replicate these nuanced teaching strategies remains a challenge (Wang et al., 2024a).

3 Paradigm Shift

The development of AI models for English Education can be broadly categorized into four successive generations as shown in Figure 3: (1) *rule-based models*, (2) *statistical models*, (3) *neural models*, and (4) *large language models*.

Stage 1: Rule-based Models (1960s–1990s). Early solutions relied on handcrafted linguistic rules to process language in tightly constrained scenarios (Grosan et al., 2011; C Angelides and Garcia, 1993). Classical platforms like PLATO (Hart, 1981) and Systran (Toma, 1977) operated effectively for highly structured tasks (e.g., grammar drills) but struggled with complex, context-dependent interactions.

Stage 2: Statistical Models (1990s–2010s). With the increased availability of digitized corpora, methods such as the early version of Google Translate (Och, 2006) and Dragon NaturallySpeaking (Blair, 1997) pioneered statistical pattern mining. These approaches leveraged large datasets to infer linguistic rules and conduct specific tasks probabilistically, improving scalability yet still lacking deeper semantic understanding.

Stage 3: Neural Models (2010s–2020s). The advent of deep learning architectures (e.g., RNNs (Yu et al., 2019) and Transformers (Vaswani, 2017)) enabled more robust context modeling, sparking transformative applications like Grammarly (Fitria, 2021) and Duolingo (Vesselinov and Grego, 2012). These systems offered enhanced personalization and feedback, significantly augmenting learners' writing and reading comprehension.

Stage 4: Large Language Models (2020s–Present). Nowadays, various LLMs (e.g., ChatGPT (Achiam et al., 2023)) combine massive pretraining with instruction tuning, achieving impressive results in multi-turn dialogue, individualized scaffolding, and multimodal integration. Tools such as Khanmigo (Anand, 2023) demonstrate

LLMs' potential for real-time conversational practice, dynamic content creation, and inclusive educational support at scale.

Our position. We foresee next-generation LLMs with deeper alignment to pedagogical principles and stronger guardrails to mitigate misinformation and bias. Future models may integrate multimodal data (e.g., text, image, video, speech) to adapt to diverse learner profiles in real time. These improvements will reinforce the position that LLMs can evolve into more effective tutors for English Education.

4 LLMs as Data Enhancers

Education is a high-stake area where any hallucination could cause devastating harm to humans' cognition activities (Ho et al., 2024). One of the hallucination causes is from data (Huang et al., 2023). Therefore, high-quality and diverse data resources (Long et al., 2024) are critical to ensuring the reliability of incorporating LLMs into English Education. The 1) *creation*, 2) *reformation*, and 3) *annotation* of educational materials are crucial to delivering effective and engaging teaching. Traditional resource development methods often lack the scalability, adaptability, and personalization necessary to meet the diverse needs of learners (Feng et al., 2021; Shorten et al., 2021). In contrast, LLMs emerge as transformative tools capable of enhancing these processes (Wang et al., 2024c; Liu et al., 2024c). This section explores how LLMs serve as data enhancers in English Education.

4.1 Data Creation

Creating pedagogically sound and learner-specific data is a cornerstone of personalized learning. However, manually creating such resources is time-consuming and often fails to address the wide range of learner needs (Cochran et al., 2022). LLMs can revolutionize this process by generating tailored and diverse educational content or responses on demand (Zha et al., 2023; Cochran et al., 2023).

Educational Materials Generation. A primary use of LLMs in data creation is the *generation of educational questions* aligned with specific learning objectives. Due to their superior contextual understanding, classic rule-based approaches have largely been eclipsed by neural network-based techniques (Kurdi et al., 2020; Rathod et al., 2022; Mulla and Gharpure, 2023). LLMs can produce answer-aware (whose target answer is known) or answer-agnostic (whose answer is open) (Zhang et al., 2021), resulting in more nuanced exercises and assessments (Xiao et al., 2023).

Student Simulation. Simulating the learner's perspective is crucial for designing adaptive instructional materials. Traditional surveys and standardized tests often fail to capture the complexity of dynamic learner behaviors (Käser and Alexandron, 2024). In contrast, LLM-based approaches enable high-fidelity, context-aware *student simulations* (Liu et al., 2024d; Yue et al., 2024), generating synthetic learners who exhibit realistic mastery levels and evolving behaviors. For instance, *Generative Students* (Lu and Wang, 2024) create simulated learners with various competency levels, while *EduAgent* (Xu et al., 2024) integrates cognitive priors to model complex learning trajectories and behaviors better.

Discussion. While LLMs excel at generating educational content, current approaches mainly focus on question creation, leaving many areas of English Education underexplored. Essential tasks like generating culturally rich reading materials, context-dependent writing prompts, or dynamic comprehension exercises still lack diversity and depth. Additionally, the student simulations created by LLMs often fail to reflect long-term learning trajectories or the intricacies of individual learning progress.

4.2 Data Reformation

In addition to creating new content, LLMs can adapt *existing* materials to better align with current needs. This process, commonly referred to as data reformation, involves (1) changing data types or modalities, (2) paraphrasing materials to match learner proficiency, and (3) enriching raw data with auxiliary signals or contextual content.

Teaching Material Transformation. Transforming existing materials into different forms can yield more comprehensive and immersive learning experiences. For example, *Book2Dial*(Wang et al., 2024b) generates teacher-student dialogues grounded in textbooks, keeping the content both relevant and informative. Their approach includes multi-turn question generation and answering(Kim et al., 2022), dialogue inpainting (Dai et al., 2022), and role-playing. Likewise, *Slide2Lecture* (Zhang-Li et al., 2024) automatically converts lecture slides

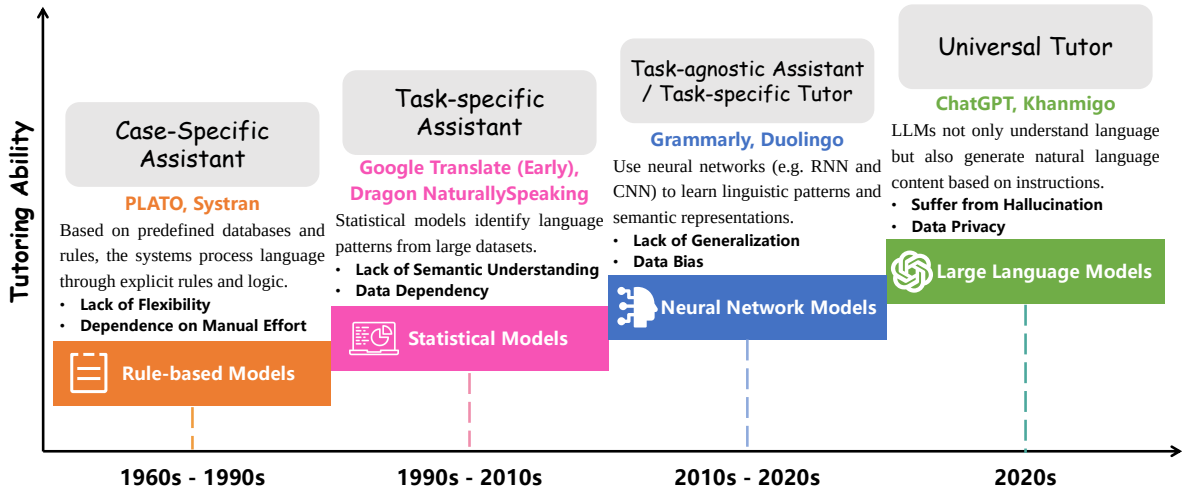

Figure 3: Roadmap of English Education.

into structured teaching agendas, enabling interactive follow-up and deeper learner engagement.

Simplification and Paraphrasing. Another vital application is simplifying or paraphrasing complex texts to specified readability levels (Huang et al., 2024a) without losing key concepts (Al-Thanyyan and Azmi, 2021). This is particularly beneficial in English Education settings, where language beginners often face advanced vocabulary and complex structures (Day et al., 2025). Recent advancements in controllable generation (Zhang et al., 2023) leverage model fine-tuning on curated datasets (Zeng et al., 2023) or decoding-time interventions (Liang et al., 2024), thereby allowing educators to specify text complexity, style, or tone.

Cultural Context Adaptation. Beyond linguistic correctness, cultural nuance is another crucial factor in English Education (Byram, 1989, 2008). LLMs can facilitate this process by recontextualizing existing materials to reflect the cultural and social norms of different areas (Liu et al., 2024a; Adilazuarda et al., 2024; Kharchenko et al., 2024). For instance, a short story originally set in an English-speaking environment may be adapted for Japanese students by adjusting the characters' names, idiomatic expressions, or social customs, while preserving core instructional goals. This cultural adaptation not only enhances learner engagement but also strengthens cross-cultural competencies.

Discussion. While LLM-based data reformation can significantly enhance English Education, several gaps warrant attention. Most current studies prioritize textual forms or single-modal approaches, which may overlook valuable *multimodal* resources such as interactive video and audio-based content (Ghosal et al., 2023). Furthermore, cultural adaptation, although promising, remains underexplored in practical classroom scenarios, particularly for underrepresented persons and culturally sensitive topics. AlKhamissi et al. (2024) demonstrate how cultural misalignment can increase bias. However, robust empirical *evaluations* are still limited across diverse learners and linguistic backgrounds.

4.3 Data Annotation

While *Data Creation* focuses on generating learner-specific data, it often prioritizes diversity and adaptability over precision. The approach is particularly useful for tasks with large label spaces (Ding et al., 2024). In contrast, *Data Annotation* emphasizes producing high-quality, meticulously labeled data that is essential for tasks requiring accuracy and consistency. Unlike data creation, annotated data often undergoes rigorous validation to ensure its accuracy and relevancy (Artemova et al., 2024).

Annotation Generation. LLMs can be central to generating a variety of annotations, including categorical labels, rationales, pedagogical feedback, and linguistic features such as discourse relations. Recent prompt engineering and fine-tuning techniques have further expanded LLMs' annotation capabilities. For instance, Ye et al. (2024) leverage GPT-4 to annotate structured explanations for Chinese grammatical error correction, while Samuel et al. (2024) examine GPT-4 as a surrogate for human annotators in low-resource reading comprehension tasks. Likewise, Siyan et al. (2024) deploy GPT-4-Turbo for audio transcript annotations. However, inconsistencies across LLMs (Törnberg,

2024) remain a serious challenge, posing risks to educational reliability.

Annotation Assessment. Although LLM-based annotation is efficient, it also raises critical issues of bias, calibration, and validity, particularly in low-resource language contexts (Bhat and Varma, 2023; Jadhav et al., 2024). Automated or semi-automated evaluation strategies have emerged to address these quality concerns. For example, LLMs-as-Judges (Li et al., 2024a,b; Gu et al., 2024) reduce human overhead by automating evaluation, an approach increasingly explored in education-focused applications (Chiang et al., 2024; Zhou et al., 2024). However, purely automated frameworks can still propagate errors or bias.

Discussion. Although LLMs provide efficient data annotation, the inconsistency across different models remains a critical concern, affecting the quality and reliability of annotated educational materials. These discrepancies hinder the creation of universally reliable educational content, especially in diverse linguistic and cultural contexts. Additionally, automated annotations often lack the nuance needed for pedagogical applications, making it essential to involve human oversight in critical cases to mitigate errors or biases.

Our position. We acknowledge the current limitations in LLM-based data creation, reformation, and annotation for English Education. However, we believe that with continued interdisciplinary collaboration, these challenges can be addressed. *Future advancements* should focus on enhancing the accuracy and diversity of generated content, improving multi-modal and culturally sensitive learning materials, and integrating more robust systems for human-LLM collaboration (Li et al., 2023; Wang et al., 2024e) in data annotation. This will ensure that LLMs can fully realize their potential as effective tutors in English Education.

5 LLMs as Task Predictors

Task-Based Language Learning (TBLL) (Nunan, 1989; Willis, 2021) as a methodological approach is one of the effective English Education methods. LLMs have demonstrated remarkable capabilities in understanding and generating human language, making them well-suited for addressing numerous tasks in English Education. These tasks can be broadly categorized into three types based on their nature and the role of LLMs: 1) *Discriminative*, 2) *Generative*, and 3) *Mixed* of the above two roles.

5.1 Discriminative Task Predictors

Discriminative tasks in English Education primarily involve classifying learner inputs or grading their future performance. Below are some applications that are still calling for improvements:

Automated Assessment. The task aims to automatically grade students' assignments, including essay scoring (Seßler et al., 2024; Li and Liu, 2024; Syamkumar et al., 2024), short answer grading (Schneider et al., 2023; Henkel et al., 2024), and spoken language evaluation (Gao et al., 2023; Fu et al., 2024). LLMs can process learners' submissions to judge grammar, lexical diversity, coherence, and even spoken fluency, providing instant feedback. This scalability is particularly appealing for large classes, where human evaluators are often overwhelmed and unable to provide timely, personalized critique (Mizumoto and Eguchi, 2023).

Knowledge Tracing. Given sequences of learning interactions in online learning systems, Knowledge Tracing identifies and tracks students' evolving mastery of target skills (Shen et al., 2024b; Xu et al., 2023). LLM-based methods of Knowledge Tracing have been explored in cold-start scenarios (Zhan et al., 2024; Jung et al., 2024), offering strong generalization by inferring latent learner states from limited data. These approaches can support adaptive learning pathways, giving personalized recommendations based on predicted performance and knowledge gaps.

Discussion. Despite their promise in automating and personalizing these discriminative tasks, LLMs still grapple with notable limitations that hinder their utility as robust tutoring tools. First, *misalignment of assessment with expert instructors* poses risks: machine-generated scores may deviate from established rubrics or neglect qualitative nuances, leading to potential discrepancies in grading quality (Kundu and Barbosa, 2024). Second, the *lack of empathy* compounds this issue, as assessments devoid of human judgment risk discouraging learners or overlooking subtle motivational factors (Sharma et al., 2024). Knowledge tracing approaches, while promising in cold-start scenarios, struggle with capturing the complexity of long-term learning trajectories and deeper cognitive processes (Cho et al., 2024). These concerns point to the need for more

transparent and human-centered methods in utilizing LLMs for assessment.

5.2 Generative Task Predictors

Generative tasks involve producing new content or responses. LLMs are known to be adept at these tasks due to their generation capabilities.

Grammatical Error Correction and Explanation. In English writing, errors often reveal learners' gaps in grammar and vocabulary (Hyland and Hyland, 2006). LLMs can detect and correct these errors (Bryant et al., 2023; Ye et al., 2023), offering concise explanations (Ye et al., 2024) that reinforce language rules. By streamlining error detection and corrections, learners deepen their linguistic understanding.

Feedback Generation. Quizzes and exercises remain vital in English Education for practice and targeted remediation (Rashov, 2024). LLMs enhance this process by delivering prompt, personalized feedback that pinpoints strengths and addresses weaknesses (Borges et al., 2024). This scalability enables learners to self-regulate and refine their skills without relying solely on human graders (Stamper et al., 2024).

Socratic Dialogue. Moving beyond straightforward Q&A, Socratic questioning promotes critical thinking and self-reflection (Paul and Elder, 2007). *SocraticLM* (Liu et al., 2024b), for example, aligns an LLM with open-ended, inquiry-based teaching principles, guiding learners through iterative exploration rather than prescriptive correction. In theory, this fosters deeper conceptual understanding and active learner engagement.

Discussion. Despite the promise of LLM-based generation in English Education, multiple uncertainties persist. *Determining how to provide automatic feedback that genuinely maximizes learning outcomes* is an ongoing challenge (Stamper et al., 2024), particularly given education's risk-averse culture and high accountability standards (Xiao et al., 2024). Moreover, while LLMs like SocraticLM have demonstrated success in domains like mathematics, their applicability to English Education contexts has not been thoroughly validated (Liu et al., 2024b). As such, the design of strategies and follow-up queries remains an open question in ensuring these systems track and respond to learners' cognitive states.

5.3 Mixed Task Predictors

Mixed tasks integrate discriminative and generative elements, requiring LLMs to evaluate learner inputs and generate meaningful feedback or suggestions. These tasks are particularly valuable in fostering an interactive and adaptive learning experience, as they bridge the gap between evaluation and instruction.

Automated Assessment with Feedback. While discriminative systems for automated essay scoring and speech evaluation primarily focus on assigning grades, LLMs extend these capabilities by simultaneously generating formative feedback (Katuka et al., 2024; Stahl et al., 2024b). For example, an LLM can evaluate the coherence and lexical diversity of a written essay, then offer specific revision strategies. In speaking practice, it can measure fluency and pronunciation accuracy while suggesting drills to refine intonation or stress patterns. Through this combination of scoring and tailored advice, learners gain a deeper understanding of their strengths and areas for improvement.

Error Analysis. Error Analysis systematically uncovers and categorizes learners' missteps, from syntactic lapses in writing to flawed pronunciations in speaking (James, 2013; Erdoğan, 2005). LLMs functioning in a mixed capacity can classify these errors and generate corrective guidance, providing revised sentences, clarifications of grammatical rules, or remediation exercises for identified weaknesses (Myles, 2002; Mashoor and Abdullah, 2020). Such insight facilitates targeted interventions that enhance language proficiency across modalities, including reading and listening.

Discussion. Mixed-task systems hold promise by combining assessment and feedback generation, but they face notable challenges. One major issue is the *weak alignment* between scoring mechanisms and the quality of feedback provided (Stahl et al., 2024b). For example, while essay scoring systems may deliver comprehensive evaluations, the feedback often lacks specificity, limiting its instructional value. Additionally, although error analysis has potential, *the absence of standardized pedagogical benchmarks*, especially in oral tasks, hampers the reliability and comparability of LLM-based tools (Leu Jr, 1982).

Our position. While LLMs offer scalable solutions for task prediction in English Education, their current limitations—such as misalignment with expert assessments, lack of empathy, and weak alignment between assessment and feedback—require ongoing refinement. *Future research* should focus on improving model transparency, enhancing the cultural and emotional sensitivity of LLMs, and refining task predictors to better reflect long-term learning trajectories and learner motivation. Additionally, developing standardized pedagogical benchmarks for error analysis will help ensure the consistency and reliability of LLM-generated feedback.

6 LLM-empowered Agent

In this section, we delve into the potential of LLMs as intelligent tutoring agents in English Education. LLMs can act as catalysts for personalized learning, addressing the long-standing scalability, adaptability, and inclusivity challenges in traditional teaching paradigms.

6.1 Fundamental Abilities

This section highlights five key abilities of LLM-empowered agents that enable them to function as adaptive tutors.

Knowledge Integration. LLMs excel at merging structured educational knowledge graphs (Abu-Rasheed et al., 2024; Hu and Wang, 2024) with unstructured textual data (Li et al., 2024c; Modran et al., 2024), providing rich, contextualized information on linguistic constructs and cultural nuances. Their ability to perform real-time knowledge editing (Wang et al., 2024d; Zhang et al., 2024a) ensures learners receive content aligned with evolving language usage, addressing the inherent limitations of static materials.

Pedagogical Alignment. LLMs require embedding with pedagogical principles to facilitate genuine learning experiences (Carroll, 1965; Taneja, 1995). Recent work incorporates theoretical frameworks, such as Bloom's taxonomy (Bloom et al., 1956), to guide LLMs in systematically addressing different cognitive levels (Jiang et al., 2024b). Approaches like *Pedagogical Chain of Thought* (Jiang et al., 2024b) and *preference learning* (Sonkar et al., 2024; Rafailov et al., 2024) focus on aligning model responses with educational objectives.

Planning. By assisting in crafting teaching objectives and lesson designs, LLMs can handle complex tasks such as differentiated instruction (Hu et al., 2024). LessonPlanner (Fan et al., 2024) has been proposed to assist novice teachers in preparing lesson plans, with expert interviews confirming its effectiveness. Zheng et al. (2024) propose a three-stage process to produce customized lesson plans, using Retrieval-Augmented Generation (RAG), self-critique, and subsequent refinement.

Memory. Effective tutoring systems track learner histories and tailor subsequent interactions accordingly (Jiang et al., 2024a; Chen et al., 2024). When serving as memory-augmented agents, LLMs can retain individualized data—such as repeated grammar mistakes or overlooked vocabulary—thereby improving continuity and enabling consistent scaffolding of future learning tasks.

Tool Using. Beyond textual interactions, LLM-based agents can integrate specialized tools to streamline the educational ecosystem, from cognitive diagnosis modules (Ma and Guo, 2019) to report generators (Zhou et al., 2025). By orchestrating these resources, LLMs seamlessly unify diverse utilities under a single interface, enhancing learner experience and instructional efficiency.

6.2 Applications

Although still in its early stages, LLM-empowered agents have already started to show promising applications in English Education.

Classroom Simulation. Classroom simulation leverages LLM-empowered agents to recreate complex, interactive learning settings without the logistical hurdles of organizing physical classrooms (Zhang et al., 2024b). By simulating virtual students and tutors, researchers can study pedagogical strategies at scale, generate diverse learner interactions, and refine teaching techniques. Moreover, this virtual data can be used to fine-tune LLMs for specific educational contexts and learner profiles (Liu et al., 2024b), offering a cost-effective and adaptable approach to language instruction.

Intelligent Tutoring System (ITS). LLM-based agents have demonstrated the capacity to provide dynamic, personalized tutoring experiences (Wang et al., 2025; Kwon et al., 2024), effectively identifying learner weaknesses through large-scale linguistic analysis (Caines et al., 2023). This makes

them promising for delivering individualized instruction at scale. Although current ITS applications in mathematics (Pal Chowdhury et al., 2024) and science (Stamper et al., 2024) have shown success, the extension to English Education requires nuanced handling of cultural and contextual elements, as well as the unpredictability of human language usage.

Discussion. Despite the promise of these applications, critical challenges remain. Existing classroom simulation frameworks often *lack standardized benchmarks for English Education*, making it difficult to assess the efficacy and generalizability of developed systems (Zhang et al., 2024b). In addition, evaluating language-specific tutoring strategies, including real-time conversational practice and holistic skill integration, remains an underexplored frontier. Addressing these gaps requires *new datasets and metrics* centered on holistic skill development and interdisciplinary collaboration.

Our position. We argue that *future research* should focus on integrating multimodal learning tasks (Sonlu et al., 2024) and developing standardized frameworks for evaluating English Education simulations. Moreover, LLMs should evolve beyond text-based capabilities to provide real-time, context-sensitive feedback, particularly in speaking and listening. Interdisciplinary collaboration and the creation of new datasets tailored to English Education are crucial for refining these systems and ensuring their scalability and inclusivity in language instruction.

7 Challenges

While we posit that LLMs have the potential to revolutionize English Education, realizing their full promise requires addressing key challenges. This section offers a concise overview of these challenges, followed by directions that could guide future research and deployment.

Ensuring Reliability and Mitigating Hallucinations. LLMs may produce hallucinations (Huang et al., 2023) that can mislead learners and undermine pedagogical goals. This risk intensifies in high-stakes educational environments, where trust and correctness are paramount. Future directions include enhancing data quality and diversity for training (Long et al., 2024), developing techniques to integrate LLM outputs with structured domain knowledge and pedagogical rules, and employing rigorous automated and human-in-the-loop validation mechanisms to minimize such detrimental outcomes and improve the factual grounding of LLM-generated educational content.

Addressing Bias and Ethical Considerations. As LLMs inherit biases from their training data, these systems may produce culturally insensitive, stereotypical, or unfair responses, potentially harming students from diverse linguistic and sociocultural backgrounds. Moreover, significant privacy concerns emerge when collecting and using learner data to personalize instruction, particularly for K-12 students. Future research must focus on developing robust governance frameworks, transparent documentation of data sources and model behaviors, and advanced bias detection and mitigation strategies (Borah and Mihalcea, 2024; He and Li, 2024) to ensure that LLM-based tools for English Education are equitable, fair, and uphold stringent data protection standards.

Aligning With Pedagogical Principles. LLMs excel at generating fluent language but often lack deep pedagogical alignment, particularly for tasks requiring developmental sensitivity, learner motivation strategies, or differentiated instruction tailored to individual learning needs. Their general-purpose nature means they do not inherently account for established language acquisition theories or specific curricular standards (Razafinirina et al., 2024). A crucial future direction is the development of methodologies to better imbue LLMs with pedagogical intelligence. This includes co-designing LLM applications with educators, fine-tuning models on high-quality pedagogical interaction data, and creating architectures that can dynamically adapt to learners' cognitive states and developmental needs in English language learning.

8 Conclusion

This paper emphasizes the transformative potential of LLMs in English Education, positioning them as valuable tutors to complement traditional teaching methods. Through their roles as data enhancers, task predictors, and agents, LLMs can provide adaptive learning experiences across the core skills of listening, speaking, reading, and writing. This paper encourages continuing dialogue and interdisciplinary collaboration to responsibly integrate LLMs into educational ecosystems.

Limitations

Emphasis on potential over practical implementation barriers. This paper primarily focuses on the potential of LLMs to serve as effective tutors in English Education, outlining beneficial roles as data enhancers, task predictors, and agents. While we acknowledge the existence of challenges (to be discussed in Appendix 7), a limitation of this position is that the main arguments may not fully capture the considerable practical, socio-economic, and infrastructural hurdles that could impede the equitable and effective implementation of these LLM roles across diverse global educational contexts and resource settings.

Generalizability and contextual adaptation of proposed roles. We propose three broad roles for LLMs in English Education. However, this paper does not provide an exhaustive analysis of how the efficacy and suitability of LLMs in these roles might vary significantly across different target languages (especially low-resource languages), specific learner demographics (e.g., preschoolers vs. K-12 vs. adult learners, learners with disabilities), diverse cultural contexts, or varying pedagogical philosophies. The general framework presented may require substantial adaptation and further research to be effectively applied in specific English Education scenarios.

Nuances of human-LLM pedagogical interaction. While advocating for LLMs as tutors that can complement human expertise, this position paper does not delve deeply into the complex dynamics of the pedagogical interactions between learners, LLM-based tutors, human educators, and parents. Critical aspects such as optimizing the collaborative model, designing effective training for educators to leverage LLMs, mitigating risks of learner over-reliance, and ensuring that LLM interactions foster deep learning rather than superficial engagement are multifaceted issues that warrant more extensive investigation than afforded by the scope of this paper.

Acknowledgements

This research is supported in part by NSF under grants III-2106758, and POSE-2346158. This research is also supported by National Natural Science Foundation of China (Grant No.62276154); Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No.2023A1515012914 and 44030024103310080177O); Basic Research Fund of Shenzhen City (Grant No.JCYJ20210324120012033, JCYJ20240813112009013 and GJHZ20240218113603006); The Major Key Project of PCL for Experiments and Applications (PCL2024A08).

References

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. *arXiv preprint arXiv:2403.03008*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Bashar Alhafni, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. Llms in education: Novel perspectives, challenges, and opportunities. *arXiv preprint arXiv:2409.11917*.

Eman Alhusaiyan. 2024. A systematic review of current trends in artificial intelligence in foreign language learning. *Saudi Journal of Language Studies*.

Eman Alhusaiyan. 2025. A systematic review of current trends in artificial intelligence in foreign language learning. *Saudi Journal of Language Studies*, 5(1):1–16.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Preeti Anand. 2023. Khan academy creates gpt-4 based helper khanmigo marking formal entry of ai into education.

Ekaterina Artemova, Akim Tsvigun, Dominik Schlechtweg, Natalia Fedorova, Sergei Tilga, and Boris Obmoroshev. 2024. Hands-on tutorial: Labeling with llm and human-in-the-loop. *arXiv preprint arXiv:2411.04637*.

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. Evaluating LLMs for targeted concept simplification for domain-specific texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226, Miami, Florida, USA. Association for Computational Linguistics.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107, Bali, Indonesia. Association for Computational Linguistics.

Christopher Blair. 1997. Dragon–naturallyspeaking. *Journal of Osteopathic Medicine*, 97(12):711–711.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and 1 others. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. *arXiv preprint arXiv:2410.02584*.

Beatriz Borges, Niket Tandon, Tanja Käser, and Antoine Bosselut. 2024. Let me teach you: Pedagogical foundations of feedback for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12082–12104, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

M Byram. 1989. Cultural studies in foreign language education. *Multilingual Matters*, 61.

Michael Byram. 2008. *From foreign language education to education for intercultural citizenship: Essays and reflections*, volume 17. Multilingual matters.

Marios C Angelides and Isabel Garcia. 1993. Towards an intelligent knowledge based tutoring system for foreign language learning. *Journal of computing and information technology*, 1(1):15–28.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, and 1 others. 2023. On the application of large language models for language teaching and assessment technology. *arXiv preprint arXiv:2307.08393*.

John B Carroll. 1965. The contributions of psychological theory and educational research to the teaching of foreign languages. *The modern language journal*, 49(5):273–281.

Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 354–364.

Olga Cherednichenko, Olha Yanholenko, Antonina Badan, Nataliia Onishchenko, and Nunu Akopiants. 2024. Large language models for foreign language acquisition.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513, Miami, Florida, USA. Association for Computational Linguistics.

Yongwan Cho, Rabia Emhamed AlMamlook, and Tasnim Gharaibeh. 2024. A systematic review of knowledge tracing and large language models in education: Opportunities, issues, and future research. *arXiv preprint arXiv:2412.09248*.

Keith Cochran, Clayton Cohn, Nicole Hutchins, Gautam Biswas, and Peter Hastings. 2022. Improving automated evaluation of formative assessments with text data augmentation. In *International Conference on Artificial Intelligence in Education*, pages 390–401. Springer.

Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. 2023. Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. In *International Conference on Artificial Intelligence in Education*, pages 217–228. Springer.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning*, pages 4558–4586. PMLR.

Stephanie L. Day, Jacapo Cirica, Steven R. Clapp, Veronika Penkova, Amy E. Giroux, Abbey Banta, Catherine Bordeau, Poojitha Mutteneni, and Ben D. Sawyer. 2025. Evaluating genai for simplifying texts for education: Improving accuracy and consistency for enhanced readability. *Preprint*, arXiv:2501.09158.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Sarah Elaine Eaton. 2010. *Global Trends in Language Learning in the 21st Century.* ERIC.

Ronald G Ehrenberg, Dominic J Brewer, Adam Gamoran, and J Douglas Willms. 2001. Class size and student achievement. *Psychological science in the public interest*, 2(1):1–30.

Vacide Erdoğan. 2005. Contribution of error analysis to foreign language teaching. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 1(2).

Haoxiang Fan, Guanzheng Chen, Xingbo Wang, and Zhenhui Peng. 2024. Lessonplanner: Assisting novice teachers to prepare pedagogy-driven lesson plans with large language models. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–20.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a socratic chatbot. *arXiv preprint arXiv:2409.05511*.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Tira Nur Fitria. 2021. Grammarly as ai-powered english writing assistant: Students' alternative for writing english. *Metathesis: Journal of English Language, Literature, and Teaching*, 5(1):65–78.

Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and large language models: on the ethical dimensions of llm-based text simplification. *Ethics and Information Technology*, 26(3):50.

Kaiqi Fu, Linkai Peng, Nan Yang, and Shuran Zhou. 2024. Pronunciation assessment with multimodal large language models. *arXiv preprint arXiv:2407.09209*.

Yang Gao, Qikai Wang, and Xiaochen Wang. 2024. Exploring efl university teachers' beliefs in integrating chatgpt and other large language models in language education: a study in china. *Asia Pacific Journal of Education*, 44(1):29–44.

Yingming Gao, Baorian Nuchged, Ya Li, and Linkai Peng. 2023. An investigation of applying large language models to spoken language learning. *Applied Sciences*, 14(1):224.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*.

Crina Grosan, Ajith Abraham, Crina Grosan, and Ajith Abraham. 2011. Rule-based expert systems. *Intelligent systems: A modern approach*, pages 149–185.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and 1 others. 2023a. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the tenth ACM conference on learning@ scale*, pages 416–420.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA. Association for Computational Linguistics.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and 1 others. 2023b. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*.

Robert Hart. 1981. Language study and the plato system. *Studies in language learning*, 3(1):1–24.

Lin He and Keqin Li. 2024. Mitigating hallucinations in llm using k-means clustering of synonym semantic relevance. *Authorea Preprints*.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 300–304.

Huu-Tuong Ho, Duc-Tin Ly, and Luong Vuong Nguyen. 2024. Mitigating hallucinations in large language models for educational application. In *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4. IEEE.

Yanxia Hou. 2020. Foreign language education in the era of artificial intelligence. In *Big Data Analytics for Cyber-Physical System in Smart City: BDCPS 2019, 28-29 December 2019, Shenyang, China*, pages 937–944. Springer.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.

Silan Hu and Xiaoning Wang. 2024. Foke: A personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering. In *China National Conference on Big Data and Social Computing*, pages 399–411. Springer.

Chieh-Yang Huang, Jing Wei, and Ting-Hao Kenneth Huang. 2024a. Generating educational materials with different levels of readability using llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024b. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Ken Hyland and Fiona Hyland. 2006. Feedback on second language students' writing. *Language teaching*, 39(2):83–101.

Joseph Marvin Imperial, Gail Forey, and Harish Tayyar Madabushi. 2024. Standardize: Aligning language models with expert-defined standards for content generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA. Association for Computational Linguistics.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On limitations of llm as annotator for low resource languages. *arXiv preprint arXiv:2411.17637*.

Carl James. 2013. *Errors in language learning and use: Exploring error analysis*. Routledge.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892.

Yuan-Hao Jiang, Ruijia Li, Yizhou Zhou, Changyong Qi, Hanglei Hu, Yuang Wei, Bo Jiang, and Yonghe Wu. 2024a. Ai agent for education: von neumann multi-agent system framework. *arXiv preprint arXiv:2501.00083*.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024b. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *arXiv preprint arXiv:2405.06705*.

Heeseok Jung, Jaesang Yoo, Yohaan Yoon, and Yeonju Jang. 2024. Clst: Cold-start mitigation in knowledge tracing by aligning a generative language model as a students' knowledge tracer. *arXiv preprint arXiv:2406.10296*.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Haoran Ranran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. 2024. Evaluating LLMs at detecting errors in LLM responses. In *First Conference on Language Modeling*.

Fatih Karataş, Faramarz Yaşar Abedi, Filiz Ozek Gunyel, Derya Karadeniz, and Yasemin Kuzgun. 2024. Incorporating ai in foreign language education: An investigation into chatgpt's effect on foreign language learners. *Education and Information Technologies*, pages 1–24.

Tanja Käser and Giora Alexandron. 2024. Simulated learners in educational technology: A systematic literature review and a turing-like test. *International Journal of Artificial Intelligence in Education*, 34(2):545–585.

Anisia Katinskaia. 2025. An overview of artificial intelligence in computer-assisted language learning. *arXiv preprint arXiv:2505.02032*.

Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. 2024. Investigating automatic scoring and feedback using large language models. *arXiv preprint arXiv:2405.00602*.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.

HYEJI KIM, Jongyoul Park, Hyeongbae Jeon, Sidney S Fels, Samuel Dodson, and Kyoungwon Seo. 2025. Augmented educators and ai: Shaping the future of human-ai collaboration in learning. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders? *arXiv preprint arXiv:2409.13120*.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Soonwoo Kwon, Sojung Kim, Minju Park, Seunghyun Lee, and Kyuseok Kim. 2024. BIPED: Pedagogically informed tutoring system for ESL education. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3389–3414, Bangkok, Thailand. Association for Computational Linguistics.

Jeongmin Lee, Jin-Xia Huang, Minsoo Cho, Yoon-Hyung Roh, Oh-Woog Kwon, and Yunkeun Lee. 2024a. Developing conversational intelligent tutoring for speaking skills in second language learning. In *International Conference on Intelligent Tutoring Systems*, pages 131–148. Springer.

Minhwa Lee, Zae Myung Kim, Vivek Khetan, and Dongyeop Kang. 2024b. Human-ai collaborative taxonomy construction: A case study in profession-specific writing assistants. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 51–57.

Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2024c. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9):11483–11515.

Donald J Leu Jr. 1982. Oral reading error analysis: A critical review of research and application. *Reading Research Quarterly*, pages 420–437.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An LLM-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand. Association for Computational Linguistics.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Xiu Li, Aron Henriksson, Martin Duneld, Jalal Nouri, and Yongchao Wu. 2024c. Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals: Using textbook content as context for retrieval-augmented generation with large language models. In *International Conference on Artificial Intelligence in Education*, pages 118–132. Springer.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024b. Socraticlm: Exploring socratic personalized teaching with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and 1 others. 2024c. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F Chen. 2024d. Personality-aware student simulation for conversational intelligent tutoring systems. *arXiv preprint arXiv:2404.06762*.

Zhexiong Liu, Diane Litman, Elaine Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. erevise+ rf: A writing evaluation system for assessing student essay revisions and providing formative feedback. *arXiv preprint arXiv:2501.00715*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. *arXiv preprint arXiv:2405.11591*.

Wenchao Ma and Wenjing Guo. 2019. Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology*, 72(2):370–392.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*.

Bakheet Bayan Nayif Mashoor and ATH bin Abdullah. 2020. Error analysis of spoken english language among jordanian secondary school students. *International Journal of Education and Research*, 8(5):75–82.

Kaushal Kumar Maurya, KV Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2024. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. *arXiv preprint arXiv:2412.09416*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Horia Modran, Ioana Corina Bogdan, Doru Ursuțiu, Cornel Samoila, and Paul Livius Modran. 2024. Llm intelligent agent tutoring in higher education courses using a rag approach. *Preprints 2024*, 2024070519.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Johanne Myles. 2002. Second language writing and research: The writing process and error analysis in student texts. *Tesl-ej*, 6(2):1–20.

Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of varying shades: How warning affects human perception and engagement regarding LLM hallucinations. In *First Conference on Language Modeling*.

Seyed Parsa Neshaei, Richard Lee Davis, Adam Hazimeh, Bojan Lazarevski, Pierre Dillenbourg, and Tanja Käser. 2024. Towards modeling learner performance with large language models. *arXiv preprint arXiv:2403.14661*.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The write & improve corpus 2024: Error-annotated and cefr-labelled essays by learners of english.

David Nunan. 1989. *Designing tasks for the communicative classroom*. Cambridge university press.

Franz Och. 2006. Statistical machine translation live.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Richard Paul and Linda Elder. 2007. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36.

Andrew Radford, Martin Atkinson, David Britain, Harald Clahsen, and Andrew Spencer. 2009. *Linguistics: an introduction*. Cambridge University Press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Oybek Rashov. 2024. Modern methods of teaching foreign languages. In *International Scientific and Current Research Conferences*, pages 158–164.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.

Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. 2024. Pedagogical alignment of large language models (llm) for personalized

learning: A survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16(4):448–480.

Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Scarlatos and Andrew Lan. 2024. Exploring knowledge tracing in tutor-student dialogues. *arXiv preprint arXiv:2409.16490*.

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. Training llm-based tutors to improve student learning outcomes in dialogues. *arXiv preprint arXiv:2503.06424*.

Torben Schmidt and Thomas Strasser. 2022. Artificial intelligence in foreign language learning and teaching: a call for intelligent practice. *Anglistik: International Journal of English Studies*, 33(1):165–184.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 75–90. Springer.

Johannes Schneider, Bernd Schenk, and Christina Niklaus. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *arXiv preprint arXiv:2411.16337*.

Zekai Shao, Siyu Yuan, Lin Gao, Yixuan He, Deqing Yang, and Siming Chen. 2025. Unlocking scientific concepts: How effective are llm-generated analogies for student understanding and classroom practice? *arXiv preprint arXiv:2502.16895*.

Shikhar Sharma, Manas Mhasakar, Apurv Mehra, Utkarsh Venaik, Ujjwal Singhal, Dhruv Kumar, and Kashish Mittal. 2024. Comuniqa: Exploring large language models for improving english speaking skills. In *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 256–267.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024a. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. 2024b. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*.

Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. *arXiv preprint arXiv:2504.14928*.

Mostafa Faghih Shojaei, Rahul Gulati, Benjamin A Jasperson, Shangshang Wang, Simone Cimolato, Dangli Cao, Willie Neiswanger, and Krishna Garikipati. 2025. Ai-university: An llm-based platform for instructional alignment to scientific classrooms. *arXiv preprint arXiv:2504.08846*.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024. EDEN: Empathetic dialogues for English learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3492–3511, Miami, Florida, USA. Association for Computational Linguistics.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2024a. Multilingual blending: Llm safety alignment evaluation with language mixture. *arXiv preprint arXiv:2407.07342*.

SeungWoo Song, Junghun Yuk, ChangSu Choi, HanGyeol Yoo, HyeonSeok Lim, KyungTae Lim, and Jungyeul Park. 2025. Unified automated essay scoring and grammatical error correction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4412–4426, Albuquerque, New Mexico. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024b. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA. Association for Computational Linguistics.

Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2024. The effects of embodiment and personality expression on learning in llm-based educational agents. *arXiv preprint arXiv:2407.10993*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024a. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on*

Innovative Use of NLP for Building Educational Applications (BEA 2024), pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024b. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.

Danny D Steinberg and Natalia V Sciarini. 2013. *An introduction to psycholinguistics*. Routledge.

Anand Syamkumar, Nora Tseng, Kaycie Barron, Shanglin Yang, Shamya Karumbaiah, Rheeya Uppal, and Junjie Hu. 2024. Improving bilingual capabilities of language models to support diverse linguistic practices in education. *arXiv preprint arXiv:2411.04308*.

Alvin Tan, Chunhua Yu, Bria Long, Wanjing Ma, Tonya Murray, Rebecca Silverman, Jason Yeatman, and Michael C Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *Advances in Neural Information Processing Systems*, 37:77445–77467.

Vidya Ratna Taneja. 1995. *Educational thought and practice*. Sterling Publishers Pvt. Ltd.

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. Pdfchatannotator: A human-llm collaborative multimodal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 419–430.

Peter Toma. 1977. Systran as a multilingual machine translation system. In *Proceedings of the Third European Congress on Information Systems and Networks, overcoming the language barrier*, pages 569–581.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Roumen Vesselinov and John Grego. 2012. Duolingo effectiveness study. *City University of New York, USA*, 28(1-25).

Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2024a. Challenges and opportunities in translating ethical ai principles into practice for children. *Nature Machine Intelligence*, 6(3):265–270.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024b. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and 1 others. 2024c. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024d. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. 2025. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. *arXiv preprint arXiv:2501.15749*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024e. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

John L Watzke. 2003. *Lasting change in foreign language education: A historical case for change in national policy*. Bloomsbury Publishing USA.

Marion Williams, Robert Burden, Gérard Poulet, and Ian Maun. 2004. Learners' perceptions of their successes and failures in foreign language learning. *Language Learning Journal*, 30(1):19–29.

Jane Willis. 2021. *A framework for task-based learning*. Intrinsic Books Ltd.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. Human-ai collaborative essay scoring: A dual-process framework with llms. *Preprint*, arXiv:2401.06431.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Bihan Xu, Zhenya Huang, Jiayu Liu, Shuanghong Shen, Qi Liu, Enhong Chen, Jinze Wu, and Shijin Wang. 2023. Learning behavior-oriented knowledge tracing. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2789–2800.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2024. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.13191*.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023. MixEdit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10161–10175, Singapore. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024. Excgec: A benchmark of edit-wise explainable chinese grammatical error correction. *arXiv preprint arXiv:2407.00924*.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Murong Yue, Wijdane Mifdal, Yixuan Zhang, Jennifer Suh, and Ziyu Yao. 2024. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *arXiv preprint arXiv:2404.06711*.

Weihao Zeng, Lulu Zhao, Keqing He, Ruotong Geng, Jingang Wang, Wei Wu, and Weiran Xu. 2023. Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14179–14196, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*.

Bojun Zhan, Teng Guo, Xueyi Li, Mingliang Hou, Qianru Liang, Boyu Gao, Weiqi Luo, and Zitao Liu. 2024. Knowledge tracing as language processing: A large-scale autoregressive paradigm. In *International Conference on Artificial Intelligence in Education*, pages 177–191. Springer.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024a. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024b. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Daniel Zhang-Li, Zheyuan Zhang, Jifan Yu, Joy Lim Jia Yin, Shangqing Tu, Linlu Gong, Haohua Wang, Zhiyuan Liu, Huiqin Liu, Lei Hou, and 1 others. 2024. Awaking the slides: A tuning-free and knowledge-regulated ai tutoring system via language model coordination. *arXiv preprint arXiv:2409.07372*.

Ying Zheng, Xueyi Li, Yaying Huang, Qianru Liang, Teng Guo, Mingliang Hou, Boyu Gao, Mi Tian, Zitao Liu, and Weiqi Luo. 2024. Automatic lesson plan generation via large language models with self-critique prompting. In *International Conference on Artificial Intelligence in Education*, pages 163–178. Springer.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Yizhou Zhou, Mengqiao Zhang, Yuan-Hao Jiang, Xinyu Gao, Naijie Liu, and Bo Jiang. 2025. A study on educational data analysis and personalized feedback report generation based on tags and chatgpt. *arXiv preprint arXiv:2501.06819*.

Xinlin Zhuang, Hongyi Wu, Xinshu Shen, Peimin Yu, Gaowei Yi, Xinhao Chen, Tu Hu, Yang Chen, Yupei Ren, Yadong Zhang, Youqi Song, Binxuan Liu, and Man Lan. 2024. TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education. In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 5749–5765, Bangkok, Thailand. Association for Computational Linguistics.

A Literature Review

We provide an overview of LLM-centric research of English Education presented in Figure 4.

B Future Directions

Establishing Robust Evaluation Frameworks. A significant challenge in leveraging LLMs for English Education is the current lack of widely accepted and easily implementable evaluation frameworks to assess the quality of LLM-based teaching interactions and outcomes. Existing metrics often focus on linguistic correctness or task completion (Tan et al., 2024; Macina et al., 2025) rather than pedagogical efficacy or impact on learning (Chiang et al., 2024). Future work should prioritize the development of standardized evaluation methodologies, including comprehensive benchmarks and nuanced metrics that capture both the accuracy of linguistic information and the pedagogical value of LLM interventions. This will be essential for comparing different systems and guiding iterative improvements.

Integrating with Modern Standardized Educational Frameworks. English language learning is often governed by established standards and frameworks, such as the Common European Framework of Reference for Languages (CEFR)[1] or Common Core State Standards (CCSS)[2]. For LLM-based tools to be truly effective and gain acceptance, their outputs and interaction patterns should align with these existing frameworks. Future technical development should focus on enabling LLMs to reference, interpret, and operate consistently within these standards (Nicholls et al., 2024; Imperial et al., 2024). This includes generating proficiency-level-appropriate content, providing feedback that corresponds to specific framework descriptors, and assisting learners in achieving standardized learning objectives, thereby enhancing usability, conformity, and trustworthiness among educators and learners.

Fostering Human-AI Collaboration in Pedagogy. While LLMs offer transformative potential, it is unlikely they will completely replace human teachers in English Education in the foreseeable future. Instead, the most promising path involves developing sophisticated human-AI collaborative educational technologies (KIM et al., 2025). Future research should explore how LLMs can best function as assistive tools that augment, rather than supplant, the capabilities of human educators (Shojaei et al., 2025). This includes designing intuitive interfaces for teachers to guide, customize, and oversee LLM-driven activities, investigating teachers' perspectives on integrating LLMs into their practice, and defining technical benchmarks for when an LLM possesses sufficient acquired skills to reliably assist teachers. The focus must be on a synergistic model where LLMs handle scalable tasks while human teachers provide the crucial elements of empathy, nuanced understanding, and holistic student development.

C Alternative Views

While this paper supports the use of LLMs in English Education, it is essential to consider alternative perspectives. Below, we discuss two key opposing views and provide counterarguments.

C.1 Task-Specific or Language-Specific Models as Better Alternatives

Some argue that specialized or language-specific models, including classical ML systems with carefully engineered features, can outperform general-purpose LLMs in narrowly defined tasks (e.g., phonetics or grammar drills (Fang et al., 2023)). By focusing on limited objectives, such models avoid the computational overhead and potential inaccuracies of LLMs, which aim to handle a broader range of inputs and contexts (Shen et al., 2024a).

Counterargument. While specialized models may excel in isolated tasks, they lack the flexibility required for comprehensive English Education, which involves cultural nuances, conversations, and evolving learner needs. In contrast, LLMs can be fine-tuned for specific goals while still offering broader linguistic competence (Song et al., 2024a). Additionally, relying on multiple specialized models can be resource-intensive, whereas a well-configured LLM provides a unified framework that balances specialization and scalability.

C.2 Concerns About Over-Reliance on LLMs

Critics warn that over-reliance on LLMs may lead to problems such as generating misleading out-

[1] https://www.coe.int/en/web/portal/home
[2] https://corestandards.org/

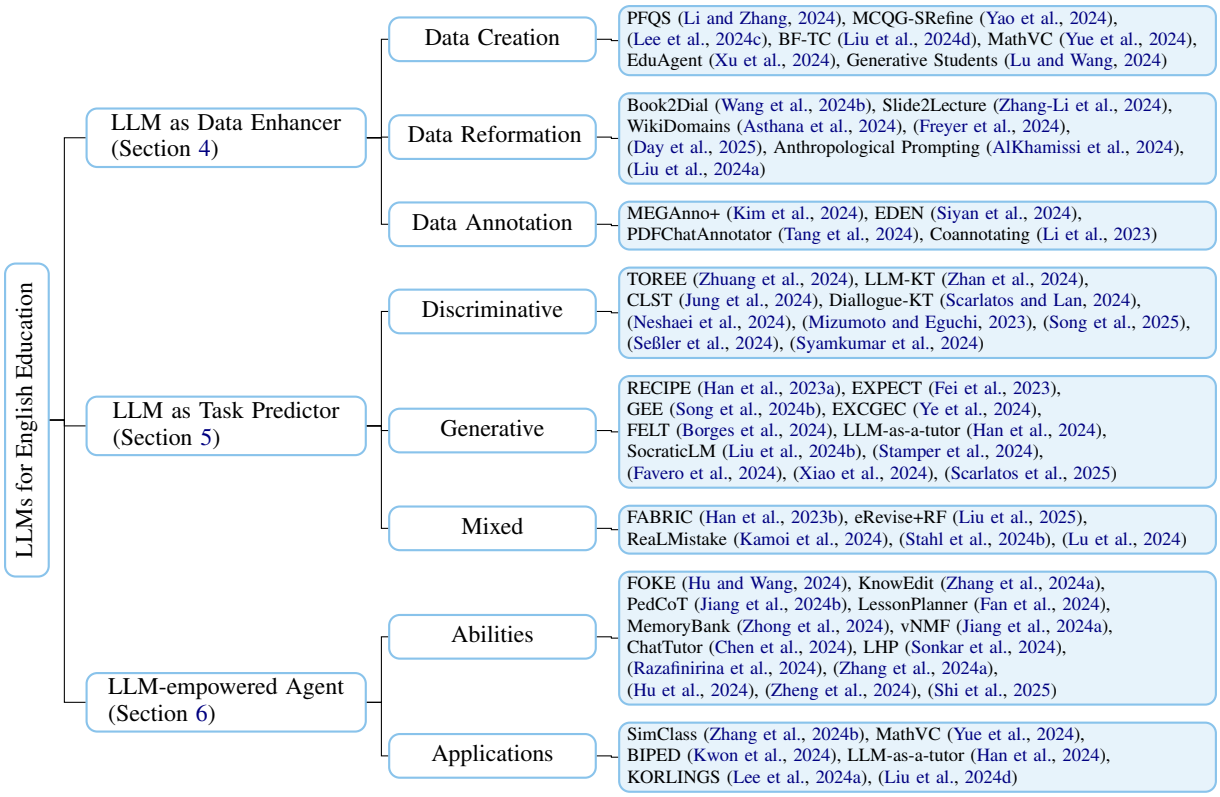

Figure 4: An overview of LLM-centric research of FLE.

puts (Nahar et al., 2024), reducing human interaction, and over-standardizing teaching methods. These issues could undermine the interpersonal and motivational aspects of language learning.

Counterargument. These risks highlight the need for balanced integration rather than the replacement of human tutors. LLMs can complement educators by automating repetitive tasks, allowing teachers to focus on individualized support and motivation. Advances in AI safety, such as feedback loops (Tong et al., 2024) and human-in-the-loop systems (Wu et al., 2022), can help minimize inaccuracies (Ho et al., 2024). Additionally, the fine-tuning capabilities of LLMs ensure adaptability, supporting diverse and inclusive learning experiences (Lee et al., 2024b).