Opportunities and Challenges of LLMs in Education: An NLP Perspective

Anonymous ACL submission

Abstract

Interest in the role of large language models (LLMs) in education is increasing, considering the new opportunities they offer for teaching, learning, and assessment. In this paper, we examine the impact of LLMs on educational NLP in the context of two main application scenarios: *assistance* and *assessment*, grounding them along the four dimensions – reading, writing, speaking, and tutoring. We then present the new directions enabled by LLMs, and the key challenges to address. We envision that this holistic overview would be useful for NLP researchers and practitioners interested in exploring the role of LLMs in developing language-focused and NLP-enabled educational applications of the future.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various tasks within and beyond NLP. The rapid adoption of LLMs and generative AI by EdTech companies such as Duolingo (Naismith et al., 2023a) and Grammarly (Raheja et al., 2023, 2024) and the development of fine-tuned models for educational use cases such as LearnLM (Team et al., 2024) are some examples of real-world impact in Education domain. The NLP community has a long history in this area, especially on problems such as automated essay scoring, grammatical error correction, and text simplification, to name a few. Naturally, there is a huge interest in using LLMs for educational applications within the community. While LLMs have undoubtedly caused a paradigm shift in this area, enabling new opportunities in writing assistance, personalization, and interactive teaching and learning, among other tasks, they also present novel challenges. In this paper, we delve into the opportunities and challenges presented by LLMs for educational applications by considering the use cases involving language, and instruction in natural language, and connect the recent developments to past NLP research in this area, outlining the path ahead.

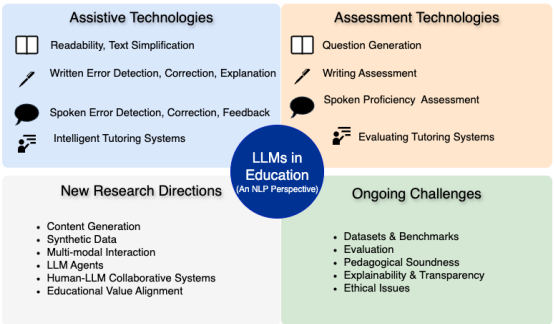

Figure 1: Overview of the paper.

We group the state of the art into two main topics: *assistive technologies* – meant to support students and teachers (§2) – and *assessment technologies* – meant to assess the performance of students (§3). Within each, we discuss the role of NLP and LLMs across specific aspects of education – reading, writing, speaking, and general tutoring. We then turn to some of the new directions enabled by LLMs in NLP in this area (§4), point to some ongoing challenges (§5), and summarize our key insights (§6).

While there are other related surveys on this topic in other disciplines (e.g., Yan et al., 2024), we focus specifically on the development of language-based educational applications, and how other existing NLP methods fit into the current scenario (See Appendix A for a discussion on related recent surveys outside NLP). In terms of scope, as we focus on language use, we exclude topics such as learning analytics, development of student models, measuring long-term educational outcomes, interactive classroom technologies, user studies, and similar.

2 Assistive Technologies

We refer to the NLP problems focused on supporting learners and/or instructors as *assistive technologies*, and discuss them by splitting them into four groups: writing, speaking, reading, and general tutoring. Note that we focus on the recent developments and refer to the relevant surveys for the pre-LLM research, where needed.

2.1 Writing

Assistive technologies for writing primarily focus on **Grammatical Error Detection (GED) and Correction (GEC)**, both of which have a long-standing pedagogical value in writing assistance tool development. GEC has witnessed significant progress over the past two decades through the organization of several shared tasks (Ng et al., 2014; Bryant et al., 2019; Masciolini et al., 2025, *inter alia*). For a comprehensive overview of the GEC literature, see the survey by Bryant et al. (2023). While GEC has received much of the attention, GED has also evolved as a stand-alone task (Tetreault and Chodorow, 2008; Leacock et al., 2014; Rei and Yannakoudakis, 2016, *inter alia*).

Several recent studies have applied LLMs to (mainly English) GEC, comparing prompting methods along two dimensions: strategy (e.g., zero-shot, few-shot, chain-of-thought) and design (e.g., fluency-oriented vs. minimal edits). So far, few-shot prompting tends to outperform zero-shot, while chain-of-thought shows no clear benefit (Fang et al., 2023; Coyne et al., 2023; Wu et al., 2023a; Loem et al., 2023; Davis et al., 2024; Katinskaia and Yangarber, 2024; Omelianchuk et al., 2024). In terms of performance, LLMs often outperform state-of-the-art models on some benchmarks such as JFLEG (Napoles et al., 2017) due to their strength in generating fluent rewrites, but underperform on larger benchmarks like CoNLL-2014 (Ng et al., 2014) and BEA-2019 (Bryant et al., 2019), which prioritize precision and minimal edits. This reflects the difficulty of controlling LLMs to make minimal, targeted corrections, which is essential in educational applications where the goal is to guide learners in revising their own errors while preserving intent (Nicholls, 2003).

LLMs have also been leveraged for **Grammatical Error Explanation (GEE)**, a task that combines GED and GEC to generate natural language explanations of learner errors. Recent work has introduced methods to guide LLMs in producing such explanations using detected edits (Kaneko and Okazaki, 2024; López Cortez et al., 2024). Song et al. (2024b) evaluated LLMs on GEE in English, German, and Chinese, showing that models often struggle to identify and explain errors, though performance improves when edits are included in the prompt. There is a growing interest in GEE for other languages as well (Ye et al., 2025; Maity and Deroy, 2025). LLMs have also been shown to be useful in providing feedback on other aspects of language assessment such as vocabulary usage (Ortiz-Zambrano et al., 2024; Bannò et al., 2025a), discourse coherence (Naismith et al., 2023b) and analytical assessment of written texts (Bannò et al., 2024b; Stahl et al., 2024), indicating the growing interest in this direction. Note that some previous work on feedback comment generation for writing also pursued similar goals but preceded the widespread adoption of LLMs (Nagata, 2019; Nagata et al., 2020; Hanawa et al., 2021; Nagata et al., 2021).

2.2 Speaking

As with writing, a common application for supporting learners in speech is **spoken GEC**. However, compared to written GEC which typically works with well-formed inputs where punctuation and capitalization can aid in error detection, spoken GEC presents a distinct set of challenges which significantly complicate the task of identifying and correcting grammatical errors in speech compared to written text. . Traditionally, spoken GEC systems have adopted a cascaded pipeline architecture, typically consisting of an automatic speech recognition (ASR) module to transcribe audio into text, followed by a disfluency detection module to produce fluent transcriptions, and finally a GEC module to correct grammatical errors (Lu et al., 2020, 2022). While this approach has shown some effectiveness, it is often hindered by error propagation across stages, which can degrade overall system performance.

This was followed by end-to-end approaches powered by large speech foundation models which promise to decrease the number of compounded errors (Bannò et al., 2024a). To address the problem of the scale of data needed to build such systems, Qian et al. (2025a) explored data augmentation, and Qian et al. (2025b) describe a novel reference alignment process to reduce transcription errors which outperforms a fine-tuned multimodal LLM Lu et al. (2025) for this task.

Despite these advances, generating accurate and meaningful feedback from spoken input continues to be a significant challenge. The recent release of the Speak & Improve Corpus (Knill et al., 2024), the first publicly available speech dataset annotated for grammatical errors, and its associated challenge (Qian et al., 2024) represents a major milestone and is expected to catalyze further progress and innovation in the field.

2.3 Reading

Assistive technologies for reading in NLP primarily focus on **Automatic Readability Assessment** and **Automatic Text Simplification**. Readability Assessment refers to the task of assigning a reading level to a given text based on its language difficulty, to various target readers. Interest in this topic is almost a century old among the education researchers (e.g., Vogel and Washburne, 1928) while the NLP research has an over two decade history (Kevyn, 2014; Vajjala, 2022), and different approaches from feature based machine learning to deep learning methods have been studied. Recent adaptation of LLMs to this problem so far seems to indicate that task-specific fine-tuned models achieve better results than zero- or few-shot prompting of LLMs (Naous et al., 2024; Wang et al., 2024f; Smădu et al., 2024). However, other work demonstrates better agreement between LLM-generated reading level judgments and human evaluations (Trott and Rivière, 2024), and Rooein et al. (2024) argue for new prompt-based evaluation metrics switching from the traditional static evaluation metrics while using LLMs.

Text Simplification refers to the task of generating text in a simpler, easier to understand language, given a more complex text (typically sentence-to-sentence). It is a well-studied area of research in NLP (Alva-Manchego et al., 2020; Štajner, 2021; Chi et al., 2023; Huang and Kochmar, 2024, *inter alia*) and the advent of LLMs resulted in a natural extension of this research. While Engelmann et al. (2024) propose to use LLMs to create datasets for text simplification research, several groups showed the effectiveness of few-shot, in-context learning for generating diverse simplifications in multiple languages (Kew et al., 2023a; Nozza and Attanasio, 2023a; Scalercio et al., 2024). Human user studies show better comprehension with LLM simplified text (Guidroz et al., 2025) but also substantial variation among human judgements (Trott and Rivière, 2024). In terms of modeling, some recent approaches utilize LLMs and multi-agentic workflows to explore document-level simplification, showing promising early results (Mo and Hu, 2024; Fang et al., 2025; Qiang et al., 2025). There is also a growing interest in personalizing text simplification through preference learning (Gao et al., 2025), generating texts at multiple levels of simplification (Farajidizaji et al., 2024; Barayan et al., 2025), domain specific simplification (Zečević et al., 2024), and elaborative simplification (Hewett et al., 2024).

2.4 Tutoring

Within the domain of general knowledge acquisition and tutoring, one of the most effective NLP-enabled tools are **Intelligent Tutoring Systems (ITS)**, in particular, dialogue-based ITSs. ITS are defined as computerized learning environments that incorporate computational models and provide feedback based on students' learning progress (Graesser et al., 2001); for dialogue-based systems, such feedback and communication with the student are empowered by NLP models. Lack of individualized tutoring has been linked to less effective learning and increased learner dissatisfaction (Brinton et al., 2014; Eom et al., 2006; Hone and El Said, 2016), particularly in large classroom settings. This has led to the development of pre-LLM ITSs (Paladines and Ramirez, 2020), including systems focused on misconception identification (Graesser et al., 1999; Rus et al., 2013), model-tracing tutors (Rickel et al., 2002; Heffernan et al., 2008), constraint-based models (Mitrovic, 2005), and Bayesian network models (Pon-Barry et al., 2004) across educational levels. LLM-powered ITS systems promise to offer more personalized, one-on-one tutoring, enabling equitable and pedagogically sound learning experiences, which have long been known to lead to substantial learning gains (Bloom, 1984). Methods such as prompting (Wang et al., 2024c), fine-tuning (Jurenka et al., 2024), and Reinforcement Learning from Human Feedback (RLHF) (Team et al., 2024) have been used in state-of-the-art LLM-based ITSs, as they help to overcome the limitations of traditional systems by enabling more adaptive, generalizable, and effective tutoring models.

One of the key limitations for ITSs is the scope and size of current educational datasets (Macina et al., 2023b; Wang et al., 2024c; Stasaski et al., 2020a). Thus, building large-scale, publicly available educational datasets for LLM pre-training and

fine-tuning should be prioritized in the near future. The focus on domain-specific models optimized for educational tasks and methods and the development of methods to assess the long-term impact of LLM-driven tutoring on learners and educators, including analysis of pedagogical effectiveness and bias, should also be considered more closely.

3 Assessment Technologies

The assessment of writing, speaking, reading, and tutoring relies on a set of overlapping principles and techniques. Although each modality has its own unique features, they are deeply intertwined, particularly in the use of textual analysis and information extraction methods. Many of these techniques, initially developed within the domain of writing assessment, have been adapted for use in both speaking, reading and tutoring contexts, which we summarize in this section.

3.1 Writing

The origins of **automated writing assessment (AWA)** date back to the 1960s with the introduction of Project Essay Grade, (Page, 1966, 1968). with notable progress in the 1990s and early 2000s that saw the emergence of the commercial systems such as e-rater® (Burstein, 2002), IntelliMetric™ (Rudner et al., 2006), and the Intelligent Essay Assessor™ (Landauer et al., 2002). In later years, Deep Neural Network (DNN) approaches have led to substantial progress (Alikaniotis et al., 2016). In particular, transformer-based models have achieved performance levels that surpass even human inter-annotator agreement (Rodriguez et al., 2019). Comprehensive overviews on AWA can be found in Beigman Klebanov and Madnani (2022) and Li and Ng (2024a,b).

Recent studies looked into the usefulness of LLMs for the assessment of second language (L2) writing, obtaining promising results (Mizumoto and Eguchi, 2023; Yancey et al., 2023a). In line with Liusie et al. (2024)'s observation that LLMs tend to perform better at comparative rather than absolute assessment, Cai et al. (2025) proposed a combined ranking-and-scoring framework that outperforms standard prompt-based approaches. While most of the writing assessment research focused on evaluating the language proficiency aspect, a substantial amount of NLP research also focused on content assessment, in the form of **short answer scoring** (Burrows et al., 2015). LLM based research on this topic is still emerging and recent studies so far conclude that zero/few-shot learning with LLMs fares poorly compared to traditional fine-tuning approaches for this task (Chamieh et al., 2024; Ferreira Mello et al., 2025).

3.2 Speaking

Research on **automated speaking assessment (ASA)** began with relatively simple tasks, such as evaluating learners' ability to read individual words or sentences (Bernstein et al., 1990; Cucchiarini et al., 1997; Franco et al., 2000). A significant milestone in this field was the development of SpeechRater system, which broadened the scope of automated assessment to include both spontaneous and read speech (Xi et al., 2008). Recent years have seen significant advancements in the field through the adoption of DNN approaches (Qian et al., 2012), and end-to-end neural-based methods have outperformed traditional systems such as SpeechRater (Chen et al., 2018b). A comprehensive survey of ASA can be found in Zechner and Evanini (2019).

Pre-trained language models have contributed to further progress in ASA (Raina et al., 2020; Wang et al., 2021) and recent research has explored speech embedding representations for applications such as mispronunciation detection and diagnosis (Wu et al., 2021b; Xu et al., 2021), automatic pronunciation assessment (Kim et al., 2022), and the evaluation of proficiency across both monologic (Bannò et al., 2023a; Park and Ubale, 2023) and conversational (McKnight et al., 2023) data.

The application of speech-based LLMs in this domain is still in its early stages. Fu et al. (2024) developed a speech LLM for L2 assessment that achieved competitive performance, albeit limited to the specific task of pronunciation scoring. With respect to holistic assessment, Ma et al. (2025) recently explored the application of Qwen2-Audio (Chu et al., 2024), for ASA in both zero-shot and fine-tuned settings. In a recent related study, Bannò et al. (2025b) demonstrated that integrating analytic proficiency descriptors with a zero-shot, text-based LLM applied to automatic transcriptions outperforms a BERT-based grader fine-tuned for the task, and achieves competitive performance compared to fine-tuned speech-based LLMs. This appears to be a promising direction to pursue in future research on ASA.

3.3 Reading

One commonly studied problem in NLP in the area of reading assessment is the generation of reading comprehension questions. Question generation research in educational NLP and AI community in general addressed different scenarios from form-focused questions (e.g., to check grammatical knowledge) to more content-focused reading comprehension questions, using a range of methods from syntactic structures to neural language models (Kurdi et al., 2020; Perkoff et al., 2023; Uto et al., 2023; Al Faraby et al., 2024a). LLMs were used for question generation in math domain (Christ et al., 2024; Scarlatos et al., 2024) and for personalized question generation in general (Xiao et al., 2023a; Säuberli and Clematide, 2024a). Although English is the dominant language for research on this topic, cross-lingual transfer approaches have also been explored, and Hwang et al. (2024) show that smaller fine-tuned language models can achieve comparable performance to larger language models on this task.

While the past research was restricted to a smaller set of datasets, the advent of LLMs resulted in approaches to benchmark construction and generation of questions at various difficulty levels according to a pre-existing taxonomy (Chen et al., 2024; Scaria et al., 2024b), and towards the development of novel evaluation approaches for automatically generated questions (Moon et al., 2024; Deroy et al., 2025). Flor (2025) presents an elaborate summary of automatic question generation research from traditional rule based methods to generative AI in a series of articles, which can serve as a good reference for those interested in further study. Item Response Theory (IRT, Lalor et al. (2024)), which has been widely used in psychometrics, has been explored in the context of vocabulary assessment (Ehara et al., 2012; Ehara, 2018) and developing questions for general language assessment (Settles et al., 2020) in the past in NLP research. However, research combining automatic question generation with IRT is yet to emerge in NLP research, and appears to be a promising area where LLMs can help.

3.4 Tutoring

Tutoring systems have long served as embedded assessment technologies, using learner interactions to evaluate understanding and guide instruction. Early systems like PLATO used rule-based feedback and simple branching logic for assessment and remediation (Woolf, 2010). Later, ITSs incorporated expert system models and student diagnostic models to reason about domain knowledge and identify misconceptions (Clancey, 1987). By the 1990s, cognitive tutors like the Algebra Tutor employed cognitive models combined with model tracing and knowledge tracing to perform fine-grained, real-time skill assessment (Anderson et al., 1995). Other ITS approaches such as AutoTutor utilize NLP models and dialogue-based reasoning to assess deeper conceptual understanding (Nye et al., 2014).

Recent advances in LLMs have transformed tutoring systems into flexible, multi-modal assessment environments. LLM-based platforms like Khanmigo (Shetye, 2024) and Google's LearnLM (Jurenka et al., 2024; Team et al., 2024) leverage generative AI to assess learner responses in natural language, interpret comprehension across reading, writing, and speaking tasks, and adapt instruction accordingly. Unlike traditional ITSs, LLMs enable open-ended, personalized feedback across diverse learning tasks, integrating instruction and assessment seamlessly (Venugopalan et al., 2025; Wang et al., 2025).

Despite their potential, LLM-based tutoring systems often lack rigorous validation linking their assessments to learning outcomes (Macina et al., 2023c). Few studies have examined their diagnostic accuracy (Maurya et al., 2025), adaptability across diverse learners (Wang et al., 2024d), or long-term impact on knowledge retention (Kosmyna et al., 2025). Ethical concerns such as feedback bias and transparency also remain underexplored (Mvondo et al., 2023). Future research should develop standardized evaluation frameworks and investigate how LLM-driven assessments can be aligned with pedagogical goals.

Compared to assistive technologies, it appears that there are relatively fewer cases of LLMs' integration into assessment approaches, although it is clearly increasing. One reason could be that assessment is likely subject to more questions around reliability and validity of the models, considering the potential high stakes of the outcomes. Despite that, what we have seen so far shows how LLMs are increasingly being used in some of the common educational tasks traditionally studied in the NLP community.

4 New Directions Enabled by LLMs

In this section, we turn to previously under-explored or new use cases enabled by LLMs across the four aspects (writing, speaking, reading, tutoring), for both assistive and assessment use cases.

Content Generation: A relatively new task, introduced with the advent of LLMs capable of fluent text generation in multiple languages, is educational content generation according to expert defined standards (Imperial et al., 2024), for a specified grade level (Bezirhan and von Davier, 2023; Jin et al., 2025), or for creation of evaluation and scaffolding exercises for different subjects (Xiao et al., 2023a; Malik et al., 2025). One interesting question to extend this line of research further could be on-the-fly content generation given a topic, grade and standard specification, and target audience.

Multi-modal Interaction: Text has been the dominant form of input in the development of educational technology applications. However, with multi-modal LLMs, some recent research explored other modes of interaction. Curating multi-modal content for education (Chaturvedi, 2024), multi-modal question generation (Luo et al., 2024), end-to-end spoken language grammatical error correction (Bannò et al., 2024a), low-resource language learning app development (Chu et al., 2025a), supporting listening assessment (Aryadoust et al., 2024), and evaluating handwritten exams (Liu et al., 2024) are some recent examples. Given these diverse use cases, and given that human learning can be considered multi-modal as we gain information from multiple forms of content, modeling of multi-modal interactions in human learning and multi-modal content generation can be considered challenging and useful future possibilities to study.

Synthetic Data Generation for Fine-tuning: Synthetic data is increasingly being used at various stages of LLM training and fine-tuning pipelines, and education domain also started to see some new use cases for synthetic data such as aiding the development of educational chatbots and tutoring systems (Wang et al., 2024a; Fateen and Mine, 2024), spoken GEC (Karanasou et al., 2025), development of benchmark datasets for educational applications (Engelmann et al., 2024; Xu et al., 2025), and using LLMs as proxies for piloting educational assessments (Säuberli et al., 2025). Considering the advantages synthetic data provides in terms of alleviating the need for labeled training data, exploring the limits and limitations of LLM-based synthetic data generation approaches for educational applications would be an important direction for the future.

LLM Agents for Education: When LLMs are combined with components such as memory, tool use, and planning to solve complex tasks, they are referred to as *LLM agents* (Chu et al., 2025b; Tran et al., 2025). In an educational context, these additional components enable real-time adaptation, access to external resources, and planning of tailored learning paths, among other capabilities. At a high level, such agents function either as *pedagogical agents* or *domain-specific educational agents* (Chu et al., 2025b). Pedagogical agents imitate tutors to assist students or instructors in tutoring sessions and simulate students for tasks such as piloting exam questions or training tutors. Furthermore, multiple agents can operate simultaneously in multi-agent setups like in CAMEL (Li et al., 2023), AutoGen (Wu et al., 2023b), and PitchQuest (Mollick et al., 2024) to develop educational prototypes or solve complex problems. Domain-specific educational agents assist with learning in subjects such as science, languages, or professional development for specific domains. However, beyond general risks such as safety, hallucinations, and bias, the responses of the current state-of-the-art models are often not grade-appropriate (Srivatsa et al., 2025), may diverge from the learning path, conflate user roles, or enter conversational loops (amplified in multi-agent settings) (Li et al., 2023; Chu et al., 2025b; Tran et al., 2025). In summary, this research direction holds huge promise, but key limitations must be addressed when deploying these systems in sensitive domains like education.

Educational Human-LLM Collaborative Systems: Human-LLM collaborative systems leverage the complementary strengths of humans and LLMs to improve performance in tasks such as data annotation, problem-solving, and decision-making across domains like education and healthcare (Yang et al., 2024; Fragiadakis et al., 2024). In education, LLM-powered systems have been employed to support both *single-turn interactions* (e.g., answering questions, explaining steps) (Gao et al., 2024; Hashir et al., 2024) and *multi-turn interactions* (e.g., Tutor-Copilot (Wang et al., 2024d), GPTeach (Markel et al., 2023)). These systems

can deepen learner understanding and assist novice tutors in improving their teaching skills and qualities. They are not free from challenges, though. Such systems often lack interpretability, making it hard to trust AI outputs (Yang et al., 2024). They may prioritize correctness over pedagogical goals like conceptual understanding and learner support (Macina et al., 2023c). LLMs also struggle with ambiguity, personalization, and maintaining context in extended interactions, and they rarely offer adaptive feedback tailored to learners' evolving needs or emotions (Maurya et al., 2025). Addressing these challenges is essential for building reliable and effective educational human-LLM collaborative systems.

Educational Value Alignment: Alignment with human preferences is a key driver to the success of state-of-the-art LLMs (Ouyang et al., 2022; Yao et al., 2023). This ranges from the development of general values-based LLMs (Guo et al., 2025; Team et al., 2024) to models tailored to specific age groups or domains (Nayeem and Rafiei, 2024; Chen et al., 2023). These advancements have also significantly influenced the educational domain, leading to the development of education-specific LLMs such as LearnLM (Team et al., 2024; Jurenka et al., 2024) and pedagogical tutors (Dinucu-Jianu et al., 2025). These LLMs are grounded in pedagogical values (Team et al., 2024; Maurya et al., 2025) and draw on decades of research in the learning sciences to generate pedagogically rich datasets, which are subsequently used for instruction tuning and fine-tuning. These specialized models have proven effective across a wide range of educational applications. However, an open research question remains – "*What should we align with?*" (Yao et al., 2023), which directly affects LLM performance.

Specifically, in the case of tutor LLMs, there is currently no consensus among researchers regarding the key pedagogical principles and associated teacher moves that lead to effective learning (Team et al., 2024). Future research should explore the core educational values that need to be integrated to enable the development of more effective educational models.

5 Ongoing Challenges

So far, we have seen how LLMs enabled existing NLP research on educational applications, and also paved way for new use cases. With the growth in their usage in real-world educational scenarios and the potential for personalized education, a discussion about the challenges involved becomes inevitable. In this section, we discuss some of the the technical as well as broader application-related challenges in this area.

5.1 Datasets

A lot of NLP research on educational applications relies on the existence of labeled datasets. For most of the tasks, such datasets are created by re-purposing existing online resources (e.g., using Wikipedia and Simple Wikipedia, websites such as Newsela for automatic text simplification), and this is not an exception compared to other areas of research in NLP. Carefully crafted datasets that are specifically developed for a particular task (e.g., grammatical error detection) are not rare, but hard to develop on a large scale. Datasets that consider target user input (e.g., those that contain learner feedback or outcome information) are even rare. Adding multilingual support to the mix makes dataset development across educational NLP tasks still more challenging.

Although LLMs could offer better zero-shot, off-the-shelf performance for many tasks and languages today, and synthetic data generation with LLMs can address the data scarcity across languages to some extent, we would still need concerted efforts to build high quality educational datasets to develop and evaluate educational support LLMs across languages. Some recent research also reports poorer performance of LLMs across four education-related tasks beyond English and recommends verifying the LLM performance in the target language before deployment (Gupta et al., 2025). Imperial et al. (2025)'s recent effort to consolidate multilingual language proficiency assessment datasets under one unified format and license is a welcome step in this direction.

5.2 Evaluation

Across different NLP tasks involving the use of LLMs in the education domain, evaluation challenges have been widely discussed, along with a comparison between automated and human evaluation (Horbach et al., 2020; Vásquez-Rodríguez et al., 2021; Agrawal and Carpuat, 2024; Kobayashi et al., 2024). While most of the discussions around evaluation focused on the task specific aspects, for technologies such as tutoring systems, a multi-dimensional view of evaluation is necessary.

Traditional evaluation of teacher effectiveness has relied on artifacts, portfolios, self-reports, and student feedback (Goe et al., 2008). More recently, text generation metrics are being explored to assess ITS or AI tutor responses. However, while effective for measuring coherence and fluency, these domain-agnostic metrics often miss deeper pedagogical aspects, depend on gold references, and can be gamed by generic responses (Tack and Piech, 2022). Efforts to capture pedagogical effectiveness more directly have included human evaluations and tailored frameworks defining specific strategies, but these face challenges such as subjectivity, lack of standardization, and limited scope (Jurenka et al., 2024). Some potential directions for evaluating tutor responses were recently proposed by Maurya et al. (2025), where a unified taxonomy was introduced to measure the quality and appropriateness of these responses. However, these rely on human evaluation, which is non-scalable and is typically conducted at the utterance level rather than the conversation level. So far, using LLMs as proxy judges shows promise but still falls short in reliably evaluating complex pedagogical traits (Gu et al., 2024).

Overall, NLP research is understandably model-focused and that impacts the way we evaluate. But, user focused evaluations are emerging. For example, some recent research points to the *mismatch between the user needs and model availability* in the context of graded content generation (Asthana et al., 2024; Kim et al., 2024b). Future NLP research could consider a user-first rather than a dataset- and model-first approach in developing standardized evaluation methodologies for LLM-based educational applications.

5.3 Ethical Issues

We found the discussion around the ethics of using LLMs in education emerging only recently in the NLP community (e.g., Hämäläinen, 2024) but there has been some thought in this direction in the broader education technology and assessment community. Yan et al. (2024) discuss the ethical implications of the increasing use of LLMs in education considering a range of use cases, and identify transparency for educational stakeholders (teachers, students, parents), privacy, support across languages, and fairness across population groups as the main ethical concerns surrounding the use of LLMs in education, calling for better reporting standards from empirical research that uses LLMs to develop new solutions. This issue of reporting standards is perhaps of most direct relevance to the NLP community.

From an assessment perspective, some recent work discussed the implications of the usage of LLMs in education to academic integrity (de Winter, 2024; Leppänen et al., 2025) and fairness (Yamashita, 2025) and the language assessment community calls for a collaboration between model developers, test creators and subject matter experts, psychometricians and the AI research community to develop education-specific standards for using AI in assessment to ensure reliability and fairness (Bolender et al., 2023; Voss et al., 2023; Xi, 2023).

Hallucination is a well-known concern with LLMs, and educational use cases are not immune to that. Some recent research on text simplification (Hewett et al., 2024; Zečević et al., 2024) pointed to how the tendency to hallucinate increases as the task gets more complex such as generating in a specific domain or in a new language, for example. While this discussion about the challenges is non-exhaustive, it broadly highlights some of the general task-agnostic issues related to the use of LLMs and NLP in education.

6 Conclusions

We presented how LLMs are integrated into existing research on the NLP-driven educational applications, and how they opened up new directions of research. Our study shows that LLMs lead to several interesting new developments which hold a lot of promise for the future in terms of both effective performance as well as inclusive development of applications addressing different languages and population groups. However, there are also several ongoing challenges related to available data, evaluation, and ethical concerns. As suggested by others, we envision an increase in inter-disciplinary collaboration between NLP researchers, domain experts and educators in leading to the development of better assistive and assessment technologies to support students and teachers in the future. We hope this paper would serve as a good starting point for NLP researchers about the state of the art in the educational applications of NLP using LLMs and what lies ahead.

Limitations

We perceive two primary limitations to this paper: (a) Since our goal in this paper is to provide

an overview of what lies ahead, we did not provide an exhaustive survey of the current state of the art. We focused largely on post-LLM research in this area, pointing to relevant surveys for pre-LLM approaches; and (b) We have also primarily restricted ourselves to NLP publication venues, citing research from other related disciplines to a much smaller extent. Our observations and conclusions drawn in this paper should be considered along with these limitations. However, we provide an extensive, although by no means an exhaustive, list of additional readings grouped by the four dimensions – writing, speaking, reading and tutoring – in the Appendix Tables 1–4, for those interested in exploring these topics further.

Ethical Considerations

The study does not involve the use of any datasets with ethical concerns or training of AI models with potential ethical issues. Hence, we do not anticipate any significant risks associated with this work.

References

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Said Al Faraby, Adiwijaya Adiwijaya, and Ade Romadhony. 2024a. Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, 34(3):1008–1045.

Said Al Faraby, Ade Romadhony, and 1 others. 2024b. Analysis of LLMs for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298.

Bashar Alhafni and Nizar Habash. 2025. Enhancing text editing for grammatical error correction: Arabic as a case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17892–17914, Vienna, Austria. Association for Computational Linguistics.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The SAMER Arabic text simplification corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093, Torino, Italia. ELRA and ICCL.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic Text Scoring Using Neural Networks. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 715–725. Association for Computational Linguistics (ACL).

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207.

Vahid Aryadoust, Azrifah Zakaria, and Yichen Jia. 2024. Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, 6:100204.

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.

John Atkinson and Diego Palma. 2025. An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1):14551.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Stefano Banno, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Stefano Bannò, Kate Knill, and Mark Gales. 2025a. Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs. *Preprint*, arXiv:2506.02758.

Stefano Bannò, Katherine M Knill, Marco Matassoni, Vyas Raina, and Mark Gales. 2023a. Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 126–130.

Stefano Bannò, Rao Ma, Mengjie Qian, Kate M. Knill, and Mark J. F. Gales. 2024a. Towards End-to-End Spoken Grammatical Error Correction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10791–10795.

Stefano Bannò, Rao Ma, Mengjie Qian, Siyuan Tang, Kate Knill, and Mark Gales. 2025b. Natural Language-based Assessment of L2 Oral Proficiency using LLMs. *Preprint*, arXiv:2507.10200.

Stefano Bannò, Michela Rais, and Marco Matassoni. 2023b. Grammatical Error Correction for L2 Speech Using Publicly Available Data. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 136–140.

Stefano Bannò, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024b. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Helmer Rayner, and Xizi Wei. 2019. Overview of the 2019 Spoken CALL Shared Task. In *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*. ISCA.

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2018. Overview of the 2018 spoken CALL shared task. In *Proceedings of Interspeech 2018*, pages 2354–2358.

Claudia Baur, Cathy Chua, Johanna Gerlach, Emmanuel Rayner, Martin Russel, Helmer Strik, and Xizi Wei. 2017. Overview of the 2017 spoken CALL shared task. In *Proceedings of the 7th Workshop on Speech and Language Technology in Education (SLaTE 2017)*.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. 1990. Automatic evaluation and training in English pronunciation. In *First International Conference on Spoken Language Processing*.

Ummugul Bezirhan and Matthias von Davier. 2023. Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5:100161.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.

Brad Bolender, Charles Foster, and Sara Vispoel. 2023. The criticality of implementing principled design when using AI technologies in test development. *Language Assessment Quarterly*, 20(4-5):512–519.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher G Brinton, Ruediger Rill, Sangtae Ha, Mung Chiang, Robert Smith, and William Ju. 2014. Individualization for education at scale: MIIC design and preliminary evaluation. *IEEE Transactions on Learning Technologies*, 8(1):136–148.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. 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*, pages 643–701.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1):60–117.

Jill Burstein. 2002. The e-rater scoring engine: automated essay scoring with natural language processing. In M. D. Shermis and J. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 113–122. Routledge, New York.

Jill Burstein. 2009. Opportunities for natural language processing research in education. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 6–27. Springer.

Heather Buzick, Maria Elena Oliveri, Yigal Attali, and Michael Flor. 2016. Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education*, 29(3):161–172.

Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. 2025. Rank-Then-Score: Enhancing Large Language Models for Automated Essay Scoring. *Preprint*, arXiv:2504.05736.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein E 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. In *LLM@ AIED*.

Andrew Caines, Christian Bentz, Calbert Graham, Tim Polzehl, and Paula Buttery. 2016. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the CrowdIS corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2145–2152, Portorož, Slovenia. European Language Resources Association (ELRA).

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Naina Chaturvedi. 2024. LLMs and NLP for generalized learning in AI-enhanced educational videos and powering curated videos with generative intelligence. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 148–154, Miami, FL, USA. Association for Computational Linguistics.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018a. LearningQ: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018b. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, pages 6234–6238.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024. Dr.Academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3138–3167, Bangkok, Thailand. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to Paraphrase Sentences to Different Complexity Levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic. Association for Computational Linguistics.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, and Omri Abend. 2020. Classifying syntactic errors in learner language. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 97–107, Online. Association for Computational Linguistics.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails. *Preprint*, arXiv:2402.09216.

Bryan R Christ, Jonathan Kropko, and Thomas Hartvigsen. 2024. MATHWELL: Generating educational math word problems using teacher annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11914–11938, Miami, Florida, USA. Association for Computational Linguistics.

Yun-Hsin Chu, Shuai Zhu, Shou-Yi Hung, Bo-Ting Lin, En-Shiun Annie Lee, and Richard Tzong-Han Tsai. 2025a. ATAIGI: An AI-Powered Multimodal Learning App Leveraging Generative Models for Low-Resource Taiwanese Hokkien. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 11–19.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025b. LLM agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

William J. Clancey. 1987. Knowledge-Based Tutoring: The GUIDON Program.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Sylvain Coulange, Marie-Hélène Fries, Monica Masperi, and Solange Rossato. 2024. A corpus of spontaneous L2 English speech for real-situation speaking assessment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 293–297, Torino, Italia. ELRA and ICCL.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction. *Preprint*, arXiv:2303.14342.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269, Online. Association for Computational Linguistics.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The English Language Learning Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 622–629.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334, New Orleans, Louisiana. Association for Computational Linguistics.

Joost CF de Winter. 2024. Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*, 34(3):915–930.

Juan Dempere, Kennedy Modugu, Allam Hesham, and Lakshmana Kumar Ramasamy. 2023. The impact of ChatGPT on higher education. In *Frontiers in Education*, volume 8, page 1206936. Frontiers Media SA.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Paul Denny, Sumit Gulwani, Neil T. Heffernan, Tanja Käser, Steven Moore, Anna N. Rafferty, and Adish Singla. 2024. Generative AI for Education (GAIED): Advances, Opportunities, and Challenges. *Preprint*, arXiv:2402.01580.

Aniket Deroy, Subhankar Maity, and Sudeshna Sarkar. 2025. MIRROR: A Novel Approach for the Automated Evaluation of Open-Ended Question Generation. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 3–32. PMLR.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. *arXiv preprint arXiv:2505.15607*.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398.

Myroslava O. Dzikovska, Johanna D. Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B. Callaway. 2010. Beetle II: A system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, pages 13–18, Uppsala, Sweden. Association for Computational Linguistics.

Yo Ehara. 2018. Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Sean B Eom, H Joseph Wen, and Nicholas Ashill. 2006. The determinants of students' perceived learning outcomes and satisfaction in university online education: An empirical investigation. *Decision Sciences Journal of Innovative Education*, 4(2):215–235.

Keelan Evanini, Matthew Mulholland, Rutuja Ubale, Yao Qian, Robert A Pugh, Vikram Ramanarayanan,

and Aoife Cahill. 2018. Improvements to an Automated Content Scoring System for Spoken CALL Responses: the ETS Submission to the Second Spoken CALL Shared Task. In *Proc. Interspeech 2018*, pages 2379–2383.

Marc Evers and Anton Nijholt. 2000. Jacob - An Animated Instruction Agent in Virtual Reality. In *Advances in Multimodal Interfaces — ICMI 2000*, pages 526–533, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

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. *Preprint*, arXiv:2304.01746.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Menna Fateen and Tsunenori Mine. 2024. Developing a tutoring dialog dataset to optimize LLMs for educational use. *arXiv preprint arXiv:2410.19231*.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587, Denver, Colorado. Association for Computational Linguistics.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, Baltimore, Maryland. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 93–103, New York, NY, USA. Association for Computing Machinery.

Michael Flor. 2025. *Automatic Question Generation*. Springer, Switzerland.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. Evaluating human-AI collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*.

Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari. 2000. The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTILL*, pages 123–128.

Reva Freedman. 2000. Plan-based dialogue management in a physics tutor. In *Sixth Applied Natural Language Processing Conference*, pages 52–59, Seattle, Washington, USA. Association for Computational Linguistics.

Kaiqi Fu, Linkai Peng, Nan Yang, and Shuran Zhou. 2024. Pronunciation Assessment with Multi-modal Large Language Models. *Preprint*, arXiv:2407.09209.

Kevin Fuchs. 2023. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? In *Frontiers in education*, volume 8, page 1166682. Frontiers Media SA.

Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. 2024. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*.

Yingqiang Gao, Kaede Johnson, David Froehlich, Luisa Carrer, and Sarah Ebling. 2025. Evaluating the Effectiveness of Direct Preference Optimization for Personalizing German Automatic Text Simplifications for Persons with Intellectual Disabilities. *Preprint*, arXiv:2507.01479.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254.

Aivars Glaznieks, Jennifer-Carmen Frey, Andrea Abel, Lionel Nicolas, and Chiara Vettori. 2023. The Kolipsi corpus Family: Resources for learner corpus research in Italian and German. *IJCoL. Italian Journal of Computational Linguistics*, 9(9-2).

Laura Goe, Courtney Bell, and Olivia Little. 2008. Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. *National Comprehensive Center for Teacher Quality*.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Guher Gorgun and Okan Bulut. 2024. Current Evaluation Methods are a Bottleneck in Automatic Question Generation. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*, volume 257 of *Proceedings of Machine Learning Research*, pages 3–8. PMLR.

Arthur C. Graesser, G Tanner Jackson, Hyun-Jeong Joyce Kim, and Andrew Olney. 2006. AutoTutor 3-D Simulations: Analyzing Users' Actions and Learning Trends. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Arthur C. Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M. Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Arthur C. Graesser, Kurt VanLehn, Carolyn P Rosé, Pamela W Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI magazine*, 22(4):39–39.

Arthur C. Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Roger Kreuz, Tutoring Research Group, and 1 others. 1999. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, and 1 others. 1993. The international corpus of learner English. *English language corpora: Design, analysis and exploitation*, pages 57–71.

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*.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, and 1 others. 2025. LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load. *arXiv preprint arXiv:2505.01980*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. Multilingual Performance Biases of Large Language Models in Education. *arXiv preprint arXiv:2504.17720*.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Mika Hämäläinen. 2024. Legal and ethical considerations that hinder the use of LLMs in a Finnish institution of higher education. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 24–27, Torino, Italia. ELRA and ICCL.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Hashir, Avinash Kumar Sharma, Shah Faisal, and Gourav Rawat. 2024. Automatic Feedback Generation In NLP: A Systematic Review. In *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, pages 1492–1497. IEEE.

Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. 2022. JADES: New text simplification dataset in Japanese targeted at non-native speakers. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 179–187, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Neil T Heffernan, Kenneth R Koedinger, and Leena Razzaq. 2008. Expanding the model-tracing architecture: A 3rd generation intelligent tutor for algebra symbolization. *International Journal of Artificial Intelligence in Education*, 18(2):153–178.

Freya Hewett, Hadi Asghari, and Manfred Stede. 2024. Elaborative simplification for German-language texts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 29–39, Kyoto, Japan. Association for Computational Linguistics.

Kate S Hone and Ghada R El Said. 2016. Exploring the factors affecting MOOC retention: A survey study. *Computers & Education*, 98:157–168.

Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1753–1762, Marseille, France. European Language Resources Association.

Andrea Horbach, Joey Pehlke, Ronja Laarmann-Quante, and Yuning Ding. 2024. Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. *International Journal of Artificial Intelligence in Education*, 34(4):1294–1320.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Yichen Huang and Ekaterina Kochmar. 2024. REFeREE: A REference-FREE model-based metric for text simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13740–13753, Torino, Italia. ELRA and ICCL.

Seonjeong Hwang, Yunsu Kim, and Gary Lee. 2024. Cross-lingual transfer for automatic question generation by learning interrogative structures in target languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3208, Miami, Florida, USA. Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugenio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas Francois, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment. *Preprint*, arXiv:2506.01419.

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.

Shin'ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, 1:91–118.

Shin'ichiro Ishikawa. 2019. The ICNALE Spoken Dialogue: A New Dataset for the Study of Asian Learners' Performance in L2 English Interviews. *English teaching*, 74(4):153–177.

Shin'ichiro Ishikawa. 2014. Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner corpus studies in Asia and the world*, 2:63–76.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 145–148, Sapporo, Japan. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Utimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *Special Issues of International Journal of the Computer*, 1:31–48.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17293–17319, Miami, Florida, USA. Association for Computational Linguistics.

Meiqing Jin, Liam Dugan, and Chris Callison-Burch. 2025. Controlling Difficulty of Generated Text for AI-Assisted Language Learning. *Preprint*, arXiv:2506.04072.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*,

pages 25–33, Baltimore, Maryland. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Austin, Texas. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, and 55 others. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. *Preprint*, arXiv:2407.12687.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.

Masahiro Kaneko and Naoaki Okazaki. 2023. Reducing sequence length by predicting edit spans with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10017–10029, Singapore. Association for Computational Linguistics.

Masahiro Kaneko and Naoaki Okazaki. 2024. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3955–3961, Torino, Italia. ELRA and ICCL.

Penny Karanasou, Mengjie Qian, Stefano Bannò, Kate Knill, and Mark JF Gales. 2025. Data Augmentation for Spoken Grammatical Error Correction. In *To appear in Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE 2025)*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China. Association for Computational Linguistics.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI*, volume 19, pages 6300–6308.

Collins-Thompson Kevyn. 2014. Computational assessment of text readability. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023a. BLESS: Benchmarking Large Language Models on Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023b. BLESS: Benchmarking large language models on sentence

simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. In *Proceedings of Interspeech 2022*, pages 1411–1415.

Haechan Kim, Junho Myung, Seoyoung Kim, Sungpah Lee, Dongyeop Kang, and Juho Kim. 2024a. LearnerVoice: A Dataset of Non-Native English Learners' Spontaneous Speech. In *Proceedings of Interspeech 2025*, pages 2325–2329.

Jenia Kim, Stefan Leijnen, and Lisa Beinborn. 2024b. Considering human interaction and variability in automatic text simplification. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 52–60, Miami, Florida, USA. Association for Computational Linguistics.

Kate Knill, Diane Nicholls, Mark J F Gales, Mengjie Qian, and Pawel Stroinski. 2024. Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback. *arXiv preprint arXiv:2412.11986*.

Kate M. Knill, Mark JF Gales, Potsawee P. Manakul, and Andrew P. Caines. 2019. Automatic Grammatical Error Detection of Non-native Spoken Learner English. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 8127–8131.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Ekaterina Kochmar, Øistein Andersen, and Ted Briscoe. 2012. HOO 2012 error recognition and correction shared task: Cambridge University submission report. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 242–250, Montréal, Canada. Association for Computational Linguistics.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv preprint arXiv:2506.08872*.

Aomi Koyama, Tomoshige Kiyuna, Kenji Kobayashi, Mio Arai, and Mamoru Komachi. 2020. Construction of an evaluation corpus for grammatical error correction for learners of Japanese as a second language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 204–211, Marseille, France. European Language Resources Association.

Nischal Ashok Kumar and Andrew Lan. 2024. Improving socratic question generation using data augmentation and preference optimization. *arXiv preprint arXiv:2403.00199*.

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.

John P. Lalor, Pedro Rodriguez, Joao Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Malta. Association for Computational Linguistics.

Thomas K Landauer, D Laham, and PW Foltz. 2002. Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis and J. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 87–112. Routledge, New York.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*, 2 edition, volume 1 of *Synthesis Lectures on Human Language Technologies*. Springer, Cham.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023. Prompt-based learning for text readability assessment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1819–1824, Dubrovnik, Croatia. Association for Computational Linguistics.

Changyoon Lee, Junho Myung, Jieun Han, Jiho Jin, and Alice Oh. 2023. Learning from Teaching Assistants to Program with Subgoals: Exploring the Potential for AI Teaching Assistants. *Preprint*, arXiv:2309.10419.

Sungjin Lee, Hyungjong Noh, Kyusong Lee, and Gary Geunbae Lee. 2011. Grammatical error detection for corrective feedback provision in oral conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 797–802.

Leo Leppänen, Lili Aunimo, Arto Hellas, Jukka K. Nurminen, and Linda Mannila. 2025. How Large Language Models Are Changing MOOC Essay Answers: A Comparison of Pre- and Post-LLM Responses. *Preprint*, arXiv:2504.13038.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024b. Automated essay scoring: Recent successes and future directions. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea*.

Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. 2024. AI-assisted Automated Short Answer Grading of Handwritten University Level Mathematics Exams. *Preprint*, arXiv:2408.11728.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

S. Magalí López Cortez, Mark Josef Norris, and Steve Duman. 2024. GMEG-EXP: A dataset of human- and LLM-generated explanations of grammatical and fluency edits. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7785–7800, Torino, Italia. ELRA and ICCL.

Hao-Chien Lu, Jhen-Ke Lin, Hong-Yun Lin, Chung-Chun Wang, and Berlin Chen. 2025. Advancing automated speaking assessment leveraging multifaceted relevance and grammar information. *Preprint*, arXiv:2506.16285.

Yiting Lu, Stefano Bannò, and Mark Gales. 2022. On assessing and developing spoken 'grammatical error correction' systems. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 51–60, Seattle, Washington. Association for Computational Linguistics.

Yiting Lu, Mark J. F. Gales, Katherine M. Knill, Potsawee Manakul, and Yu Wang. 2019a. Disfluency detection for spoken learner English. In *Proceedings of the 8th Workshop on Speech and Language Technology for Education (SLaTE)*, pages 74–78.

Yiting Lu, Mark J. F. Gales, and Yu Wang. 2020. Spoken Language 'Grammatical Error Correction'. *Proceedings of Interspeech 2020*, pages 3840–3844.

Yiting Lu, Mark JF Gales, Kate M. Knill, Potsawee Manakul, Linlin Wang, and Yu Wang. 2019b. Impact of ASR Performance on Spoken Grammatical Error Detection. *Proceedings of Interspeech 2019*, pages 1876–1880.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Assessment of L2 Oral Proficiency using Speech Large Language Models. *Preprint*, arXiv:2505.21148.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023c. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3536–3553, Online. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Subhankar Maity and Aniket Deroy. 2025. Leveraging Prompt-Tuning for Bengali Grammatical Error Explanation Using Large Language Models. *arXiv preprint arXiv:2504.05642*.

Rizwaan Malik, Dorna Abdi, Rose Wang, and Dorottya Demszky. 2025. Scaffolding middle school mathematics curricula with large language models. *British Journal of Educational Technology*.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating Uncertainty into Deep Learning for Spoken Language Assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 45–50.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-BR: a Brazilian Corpus of Essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64, Porto Alegre, RS, Brasil. SBC.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 226–236, New York, NY, USA. Association for Computing Machinery.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia. University of Tartu Library.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Simon W McKnight, Arda Civelekoglu, Mark Gales, Stefano Bannò, Adian Liusie, and Katherine M Knill. 2023. Automatic Assessment of Conversational Speaking Tests. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 99–103.

Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE corpus of non-native spoken English. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Stuart Mesham, Christopher Bryant, Marek Rei, and Zheng Yuan. 2023. An extended sequence tagging vocabulary for grammatical error correction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1608–1619, Dubrovnik, Croatia. Association for Computational Linguistics.

Detmar Meurers. 2012. Natural language processing and language learning. *Encyclopedia of applied linguistics*, pages 4193–4205.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scottland, UK. Association for Computational Linguistics.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2024. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 251–265, Mexico City, Mexico. Association for Computational Linguistics.

Antonija Mitrovic. 2005. The Effect of Explaining on Learning: a Case Study with a Data Normalization Tutor. In *AIED*, pages 499–506.

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.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.

Ethan Mollick, Lilach Mollick, Natalie Bach, LJ Ciccarelli, Ben Przystanski, and Daniel Ravipinto. 2024. AI agents and education: Simulated practice at scale. *arXiv preprint arXiv:2407.12796*.

Sneha Mondal, Ritika Ritika, Ashish Agrawal, Preethi Jyothi, and Aravindan Raghuveer. 2024. DIMSIM: Distilled multilingual critics for Indic text simplification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16093–16109, Bangkok, Thailand. Association for Computational Linguistics.

Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, Jaehyung Seo, and Heuiseok Lim. 2024. Generative interpretation: Toward human-like evaluation for educational question-answer pair generation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2185–2196, St. Julian's, Malta. Association for Computational Linguistics.

Gustave Florentin Nkoulou Mvondo, Ben Niu, and Salman Eivazinezhad. 2023. Exploring The Ethical Use Of LLM Chatbots In Higher Education. *Available at SSRN 4548263*.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476, Torino, Italia. ELRA and ICCL.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Ben Naismith, Na-Rae Han, and Alan Juffs. 2022. The university of pittsburgh english language institute corpus (PELIC). *International Journal of Learner Corpus Research*, 8(1):121–138.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023a. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023b. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of*

the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024. KidLM: Advancing Language Models for Children–Early Insights and Future Directions. *arXiv preprint arXiv:2410.03884*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics Conference*, pages 572–581.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English.

Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2025. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances. *Preprint*, arXiv:2407.09975.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2024. Difficult for whom? a study of Japanese lexical complexity. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 69–81, Miami, Florida, USA. Association for Computational Linguistics.

Debora Nozza and Giuseppe Attanasio. 2023a. Is It Really That Simple? Prompting Large Language Models for Automatic Text Simplification in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333.

Debora Nozza and Giuseppe Attanasio. 2023b. Is it really that simple? prompting large language models for automatic text simplification in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333, Venice, Italy. CEUR Workshop Proceedings.

Benjamin D. Nye, Arthur C. Graesser, and Xiaowen Hu. 2014. AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative*

Use of NLP for Building Educational Applications, pages 11–25, Online. Association for Computational Linguistics.

Jenny Alexandra Ortiz-Zambrano, César Humberto Espín-Riofrío, and Arturo Montejo-Ráez. 2024. Enhancing lexical complexity prediction through few-shot learning with gpt-3. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 68–76, Torino, Italia. ELRA and ICCL.

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia. Association for Computational Linguistics.

Leila Ouahrani and Djamal Bennouar. 2020. AR-ASAG an Arabic dataset for automatic short answer grading evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2634–2643.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

E. B. Page. 1968. The use of the computer in analyzing student essays. *International Review of Education*, 14(2):210–225.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

José Paladines and Jaime Ramirez. 2020. A Systematic Literature Review of Intelligent Tutoring Systems With Dialogue in Natural Language. *IEEE Access*, 8:164246–164267.

Seongjin Park and Rutuja Ubale. 2023. Multitask learning model with text and speech representation for fine-grained speech scoring. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhang. 2021. A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Proceedings of Interspeech 2021*, pages 4448–4452.

E. Margaret Perkoff, Abhidip Bhattacharyya, Jon Cai, and Jie Cao. 2023. Comparing neural question generation architectures for reading comprehension. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 556–566, Toronto, Canada. Association for Computational Linguistics.

Nhan Phan, Anna von Zansen, Maria Kautonen, Ekaterina Voskoboinik, Tamás Grósz, Raili Hilden, and Mikko Kurimo. 2024. Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task. In *Proceedings of Interspeech 2024*, pages 317–321.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 332–339, Marseille, France. European Language Resources Association.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Heather Pon-Barry, Brady Clark, Elizabeth Owen Bratt, Karl Schultz, and Stanley Peters. 2004. Evaluating the effectiveness of SCoT: A spoken conversational tutor. In *ITS 2004 Workshop on Dialog-based Intelligent Tutoring Systems*, pages 23–32.

Mengjie Qian, Kate Knill, Stefano Banno, Siyuan Tang, Penny Karanasou, Mark J F Gales, and Diane Nicholls. 2024. Speak & Improve Challenge 2025: Tasks and Baseline Systems. *arXiv preprint arXiv:2412.11985*.

Mengjie Qian, Rao Ma, Stefano Bannò, Kate M Knill, and Mark JF Gales. 2025a. Scaling and Prompting for Improved End-to-End Spoken Grammatical Error Correction. In *Interspeech 2025*.

Mengjie Qian, Rao Ma, Stefano Bannò, Mark J. F. Gales, and Kate M. Knill. 2025b. End-to-End Spoken Grammatical Error Correction. *Preprint*, arXiv:2506.18532.

Xiaojun Qian, Helen Meng, and Frank K Soong. 2012. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. In *Proceedings of Interspeech 2012*.

Jipeng Qiang, Minjiang Huang, Yi Zhu, Yunhao Yuan, Chaowei Zhang, and Kui Yu. 2025. Redefining Simplicity: Benchmarking Large Language Models from Lexical to Document Simplification. *Preprint*, arXiv:2502.08281.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. LSBert: Lexical Simplification Based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical Simplification with Pretrained Encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8649–8656.

Le Qiu, Shanyue Guo, Tak-Sum Wong, Emmanuele Chersoni, John Lee, and Chu-Ren Huang. 2024. CompLex-ZH: A new dataset for lexical complexity prediction in Mandarin and Cantonese. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 20–26, Miami, Florida, USA. Association for Computational Linguistics.

Xin Ying Qiu and Jingshen Zhang. 2024. Label confidence weighted learning for target-level sentence simplification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18004–18019, Miami, Florida, USA. Association for Computational Linguistics.

Vipul Raheja, Dimitris Alikaniotis, Vivek Kulkarni, Bashar Alhafni, and Dhruv Kumar. 2024. mEdIT: Multilingual text editing via instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 979–1001, Mexico City, Mexico. Association for Computational Linguistics.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Vyas Raina, Mark J. F. Gales, and Kate M. Knill. 2020. Universal Adversarial Attacks on Spoken Language Assessment Systems. In *Proceedings of Interspeech 2020*, pages 3855–3859.

Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. Building a learner corpus for Russian. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 66–75, Umeå, Sweden. LiU Electronic Press.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1181–1191, Berlin, Germany. Association for Computational Linguistics.

Jeff Rickel, Neal Lesh, Charles Rich, Candace L Sidner, and Abigail Gertner. 2002. Collaborative discourse theory as a foundation for tutorial dialogue. In *Intelligent Tutoring Systems: 6th International Conference, ITS 2002 Biarritz, France and San Sebastian, Spain, June 2–7, 2002 Proceedings 6*, pages 542–551. Springer.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. *Preprint*, arXiv:1909.09482.

Cristobal Romero and Sebastian Ventura. 2013. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Ourania Rotou and André A Rupp. 2020. Evaluations of automated scoring systems in practice. *ETS Research Report Series*, 2020(1):1–18.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of IntelliMetric™essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Vasile Rus, Sidney D'Mello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Horacio Saggion, Stefan Bott, Sandra Szasz, Nelson Pérez, Saúl Calderón, and Martín Solís. 2024. Lexical complexity prediction and lexical simplification for Catalan and Spanish: Resource creation, quality assessment, and ethical considerations. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 82–94, Miami, Florida, USA. Association for Computational Linguistics.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*,

pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Andreas Säuberli and Simon Clematide. 2024a. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.

Andreas Säuberli and Simon Clematide. 2024b. Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024*, pages 22–37.

Arthur Scalercio, Maria Finatto, and Aline Paes. 2024. Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15076–15091, Bangkok, Thailand. Association for Computational Linguistics.

Nicy Scaria, Suma Chenna, and Deepak Subramani. 2024a. How Good are Modern LLMs in Generating Relevant and High-Quality Questions at Different Bloom's Skill Levels for Indian High School Social Science Curriculum? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024b. How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico. Association for Computational Linguistics.

Alexander Scarlatos, Wanyong Feng, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. SimPA: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Iulian Vlad Serban, Varun Gupta, Ekaterina Kochmar, Dung D. Vu, Robert Belfer, Joelle Pineau, Aaron Courville, Laurent Charlin, and Yoshua Bengio. 2020. A Large-Scale, Open-Domain, Mixed-Interface Dialogue-Based ITS for STEM. *Preprint*, arXiv:2005.06616.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current Applications and New Directions*. Routledge, New York, NY.

Mark D Shermis, Jill Burstein, Derrick Higgins, and Klaus Zechner. 2010. Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education*, 4(1):20–26.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Mark D Shermis and Joshua Wilson. 2024. *The Routledge international handbook of automated essay evaluation*. Routledge, New York, NY.

Shamini Shetye. 2024. An evaluation of Khanmigo, a generative AI tool, as a computer-assisted language learning app. *Studies in Applied Linguistics and TESOL*, 24(1).

Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, and Yevgeni Berzak. 2024. Fine-grained prediction of reading comprehension from eye movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3372–3391, Miami, Florida, USA. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165(2):259–298.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024a. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17:1880–1890.

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, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Kv Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025. Can LLMs Reliably Simulate Real Students' Abilities in Mathematics and Reading Comprehension? In *In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna,Austria. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. 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.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online. Association for Computational Linguistics.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Sanja Štajner and Sergiu Nisioi. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. 2020a. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020b. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Pramuditha Suraweera and Antonija Mitrovic. 2002. KERMIT: A Constraint-Based Tutor for Database Modeling. In *Intelligent Tutoring Systems*, pages 377–387, Berlin, Heidelberg. Springer Berlin Heidelberg.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised

neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. 2025. Do LLMs Give Psychometrically Plausible Responses in Educational Assessments? *Preprint*, arXiv:2506.09796.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *Preprint*, arXiv:2205.07540.

Keren Tan, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. 2024. An LLM-enhanced adversarial editing system for lexical simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1136–1146, Torino, Italia. ELRA and ICCL.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, and 1 others. 2024. LearnLM: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*.

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872, Manchester, UK. Coling 2008 Organizing Committee.

Amalia Todirascu, Thomas François, Delphine Bernhard, Núria Gala, and Anne-Laure Ligozat. 2016. Are cohesive features relevant for text readability evaluation? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 987–997, Osaka, Japan. The COLING 2016 Organizing Committee.

Brent Townshend, Jared Bernstein, Ognjen Todic, and Eryk Warren. 1998. Estimation of spoken language proficiency. In *STiLL-Speech Technology in Language Learning*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

Masaki Uto and Yuto Uchida. 2020. Automated short-answer grading using deep neural networks and item response theory. In *International Conference on Artificial Intelligence in Education*, pages 334–339. Springer.

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International journal of artificial intelligence in education*, 28(1):79–105.

Sowmya Vajjala. 2022. Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377.

Sowmya Vajjala. 2024. Generative Artificial Intelligence and Applied Linguistics. *JALT journal: the research journal of the Japan Association for Language Teaching*, 46(1):55–76.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational*

Linguistics for Linguistic Complexity (CL4LC), pages 38–48, Osaka, Japan. The COLING 2016 Organizing Committee.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.

Justin Vasselli, Christopher Vasselli, Adam Nohejl, and Taro Watanabe. 2023. NAISTeacher: A prompt and rerank approach to generating teacher utterances in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 772–784, Toronto, Canada. Association for Computational Linguistics.

Devika Venugopalan, Ziwen Yan, Conrad Borchers, Jionghao Lin, and Vincent Aleven. 2025. Combining large language models with tutoring system intelligence: A case study in caregiver homework support. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 373–383.

Mabel Vogel and Carleton Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.

Ekaterina Voskoboinik, Nhan Phan, Tamás Grósz, and Mikko Kurimo. 2025. Leveraging Uncertainty for Finnish L2 Speech Scoring with LLMs. In *The Workshop on Automatic Assessment of Atypical Speech*. University of Tartu Library.

Erik Voss, Sara T Cushing, Gary J Ockey, and Xun Yan. 2023. The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly*, 20(4-5):520–532.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. 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.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Rose Wang, Pawan Wirawarn, Omar Khattab, Noah Goodman, and Dorottya Demszky. 2024b. Backtracing: Retrieving the cause of the query. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 722–735, St. Julian's, Malta. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024c. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024d. Tutor CoPilot: A human-AI approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024e. Large Language Models for Education: A Survey and Outlook. *Preprint*, arXiv:2403.18105.

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. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 510–519.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021. Automated scoring of spontaneous speech from young learners of English using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Ziyang Wang, Sanwoo Lee, Hsiu-Yuan Huang, and Yunfang Wu. 2024f. FPT: Feature prompt tuning for few-shot readability assessment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 280–295, Mexico City, Mexico. Association for Computational Linguistics.

Amali Weerasinghe and Antonija Mitrovic. 2006. Individualizing self-explanation support for ill-defined

tasks in constraint-based tutors. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, pages 56–64.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, Volume 4 - 2021.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023a. ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark. *Preprint*, arXiv:2303.13648.

M. Wu, K. Li, W.-K. Leung, and H. Meng. 2021a. Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In *Proceedings of Interspeech 2021*, pages 3954–3958.

Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng. 2021b. Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In *Proceedings of Interspeech 2021*, pages 3954–3958.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023b. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Xiaoming Xi. 2023. Advancing language assessment with AI and ML–Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, 20(4-5):357–376.

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David M. Williamson. 2008. Automated scoring of spontaneous speech using SpeechRaterSM v1.0. Technical Report 2, ETS Research Report Series.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023a. 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.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023b. 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.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. 2025. EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios. *arXiv preprint arXiv:2505.16160*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for Mispronunciation Detection. In *Proceedings of Interspeech 2021*, pages 4428–4432.

Taichi Yamashita. 2025. Exploring potential biases in GPT-4o's ratings of English language learners' essays. *Language Testing*, 42(3):344–358.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review.

British Journal of Educational Technology, 55(1):90–112.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023a. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023b. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Diyi Yang, Sherry Tongshuang Wu, and Marti A. Hearst. 2024. Human-AI interaction in the age of LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 34–38, Mexico City, Mexico. Association for Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values–A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv:2308.12014*.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, and 1 others. 2025. EXCGEC: A Benchmark for Edit-Wise Explainable Chinese Grammatical Error Correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25678–25686.

Su-Youn Yoon, Lisa Pierce, Amanda Huensch, Eric Juul, Samantha Perkins, Richard Sproat, and Mark Hasegawa-Johnson. 2009. Construction of a rated speech corpus of L2 learners' spontaneous speech. *CALICO Journal*, 26(3):662–673.

Zheng Yuan, Felix Stahlberg, Marek Rei, Bill Byrne, and Helen Yannakoudakis. 2019. Neural and FST-based approaches to grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 228–239, Florence, Italy. Association for Computational Linguistics.

Anđelka Zečević, Milica Ćulafić, and Stefan Stojković. 2024. On simplification of discharge summaries in Serbian: Facing the challenges. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 104–108, Torino, Italia. ELRA and ICCL.

Klaus Zechner and Keelan Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado. Association for Computational Linguistics.

Tatsuya Zetsu, Yuki Arase, and Tomoyuki Kajiwara. 2024. Edit-constrained decoding for sentence simplification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7161–7173, Miami, Florida, USA. Association for Computational Linguistics.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *Proceedings of Interspeech 2021*, pages 3710–3714.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Guoguo Zhao, Sinem Sonsaat, Apimee Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018a. L2-ARCTIC: A Non-native English Speech Corpus. In *Proceedings of Interspeech 2018*, pages 2783–2787.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018b. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

Houquan Zhou, Yumeng Liu, Zhenghua Li, Min Zhang, Bo Zhang, Chen Li, Ji Zhang, and Fei Huang. 2023. Improving Seq2Seq grammatical error correction via decoding interventions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7393–7405, Singapore. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 190–200.

A Background

Post ChatGPT, there has been a lot of interest in the role of LLMs, and more generally, generative AI, in education. Several surveys and opinion pieces were published in the past two years outlining the potential use cases, ethical implications, and challenges of using LLMs in education, primarily by research groups focusing on the use of technology in education (Dempere et al., 2023; Fuchs, 2023; Kasneci et al., 2023; Yan et al., 2024). However, none of them take a closer look at methods (LLM-based or not) used to address the different use cases of LLMs in education, which is of primary interest with NLP.

NLP research has a long history of working on problems of relevance to the education domain, such as automated language assessment and grammatical error correction, among others. Burstein (2009) and Meurers (2012) give an overview of how NLP plays a role in the development of computational tools for language learning and assessment, and more recently, Vajjala (2024) has explored the role of generative AI in language learning technology. But all three are primarily directed towards non-NLP audience and we are not aware of any such discussion directed towards an NLP audience, focusing on the recent developments in generative AI and large language models.

Caines et al. (2023) discuss the role of LLMs specifically in language teaching and assessment technologies from an NLP standpoint, with a particular focus on writing assessment, while not addressing speaking assessment. However, the use of language in the educational application space is not limited to language learning alone, and it encompasses other areas such as content learning, as well as assessment and learning support for other subjects. In this paper, we aim to give a broader perspective on the relevance of LLMs for the development of language-based educational applications in general, and how other NLP methods fit into the current scenario. Where relevant, we will also summarize the perspectives on the use of NLP and LLMs in this area from other related disciplines.

B Additional References

We provide more detailed references for additional reading on specific topics in this section, grouping them along the four dimensions: writing, speaking, reading, and tutoring, in the Tables 1–4.

Written GEC & GED	
Surveys	Bryant et al. (2023)
Datasets	Yannakoudakis et al. (2011); Dahlmeier et al. (2013); Dale et al. (2012); Ng et al. (2013, 2014); Mohit et al. (2014); Rozovskaya et al. (2015); Napoles et al. (2017); Bryant et al. (2019); Rozovskaya and Roth (2019); Koyama et al. (2020); Náplava et al. (2022); Masciolini et al. (2025)
Evaluation	Dahlmeier and Ng (2012); Felice and Briscoe (2015); Napoles et al. (2015, 2016); Bryant et al. (2017); Choshen et al. (2020); Belkebir and Habash (2021)
Pre-LLM Approaches	Chodorow et al. (2007); Kochmar et al. (2012); Felice et al. (2014); Junczys-Dowmunt and Grundkiewicz (2014, 2016); Junczys-Dowmunt et al. (2018); Yuan et al. (2019); Malmi et al. (2019); Stahlberg and Kumar (2020); Kaneko et al. (2020); Omelianchuk et al. (2020); Mallinson et al. (2020); Katsumata and Komachi (2020); Mallinson et al. (2022); Alhafni et al. (2023); Zhou et al. (2023); Mesham et al. (2023); Alhafni and Habash (2025)
LLM Approaches	Fang et al. (2023); Coyne et al. (2023); Wu et al. (2023a); Loem et al. (2023); Raheja et al. (2023); Davis et al. (2024); Katinskaia and Yangarber (2024); Omelianchuk et al. (2024); Kaneko and Okazaki (2023); Katinskaia and Yangarber (2024); Raheja et al. (2024); Omelianchuk et al. (2024); Kobayashi et al. (2024); Mita et al. (2024)
GEE	
Datasets	Nagata (2019); Nagata et al. (2020); Pilan et al. (2020); López Cortez et al. (2024); Kobayashi et al. (2024)
Pre-LLM Approaches	Nagata (2019); Pilan et al. (2020)
LLM Approaches	López Cortez et al. (2024); Kobayashi et al. (2024)
Automatic Writing Assessment	
Surveys	Shermis and Burstein (2003); Shermis et al. (2010); Shermis and Burstein (2013); Ke and Ng (2019); Beigman Klebanov and Madnani (2020, 2022); Li and Ng (2024b,a); Shermis and Wilson (2024)
Datasets	Granger et al. (1993); Yannakoudakis et al. (2011); Blanchard et al. (2013); Geertzen et al. (2013); Ishikawa (2013); Östling et al. (2013); Boyd et al. (2014); Rakhilina et al. (2016); Horbach et al. (2017); Mathias and Bhattacharyya (2018); Glaznieks et al. (2023); Marinho et al. (2021); Habash and Palfreyman (2022); Naismith et al. (2022); Crossley et al. (2023); Nicholls et al. (2024); Imperial et al. (2025)
Evaluation	Williamson et al. (2012); Buzick et al. (2016); Rotou and Rupp (2020)
Pre-LLM Approaches	Burstein (2002); Landauer et al. (2002); Rudner et al. (2006); Yannakoudakis et al. (2011); Chen and He (2013); Zesch et al. (2015); Alikaniotis et al. (2016); Vajjala (2018); Rodriguez et al. (2019); Yang et al. (2020); Wang et al. (2022)
LLM Approaches	Mizumoto and Eguchi (2023); Yancey et al. (2023b); Banno et al. (2024); Song et al. (2024a); Stahl et al. (2024); Atkinson and Palma (2025); Cai et al. (2025); Yamashita (2025)
Short Answer Scoring	
Surveys	Ziai et al. (2012); Burrows et al. (2015)
Datasets	Meurers et al. (2011); Ouahrani and Bennouar (2020)
Pre-LLM Approaches	Leacock and Chodorow (2003); Uto and Uchida (2020); Horbach et al. (2024)
LLM Approaches	Chamieh et al. (2024); Ferreira Mello et al. (2025)

Table 1: Additional references for writing tasks (assistive/assessment)

Readability Assessment	
Surveys	Collins-Thompson (2014); Vajjala (2022)
Datasets	Paetzold and Specia (2016); Vajjala and Lučić (2018); Shardlow et al. (2021); Seiffe et al. (2022); Naous et al. (2024)
Evaluation	Vajjala et al. (2016); Todirascu et al. (2016); Vajjala and Lucic (2019); Shubi et al. (2024)
Pre-LLM Approaches	Collins-Thompson and Callan (2004); Pitler and Nenkova (2008); Feng et al. (2010); Vajjala and Meurers (2012); Xia et al. (2016); Nadeem and Ostendorf (2018); Azpiazu and Pera (2019); Deutsch et al. (2020); Lee et al. (2021); Wilkens et al. (2024)
LLM Approaches	Lee and Lee (2023); Nohejl et al. (2024); Rooein et al. (2024); Wang et al. (2024f); Smădu et al. (2024)
Text Simplification	
Surveys	Siddharthan (2014); Alva-Manchego et al. (2020)
Datasets	Zhu et al. (2010); Coster and Kauchak (2011); Kauchak (2013); Hwang et al. (2015); Xu et al. (2015, 2016); Kajiwara and Komachi (2016); Zhang and Lapata (2017); Sulem et al. (2018b); Scarton et al. (2018); Vajjala and Lučić (2018); Saggion et al. (2022); Hayakawa et al. (2022); Ryan et al. (2023); Alhafni et al. (2024); Shardlow et al. (2024); Jiang and Xu (2024); Saggion et al. (2024); Qiu et al. (2024); Nagai et al. (2024)
Evaluation	Xu et al. (2016); Sulem et al. (2018a); Vásquez-Rodríguez et al. (2021); Alva-Manchego et al. (2021); Cardon et al. (2022); Huang and Kochmar (2024); Agrawal and Carpuat (2024)
Pre-LLM Approaches	Chandrasekar et al. (1996); Elhadad and Sutaria (2007); Zhu et al. (2010); Woodsend and Lapata (2011); Wubben et al. (2012); Kajiwara et al. (2013); Shardlow (2014); Xu et al. (2016); Paetzold and Specia (2016); Nisioi et al. (2017); Zhang and Lapata (2017); Alva-Manchego et al. (2017); De Hertog and Tack (2018); Štajner and Nisioi (2018); Guo et al. (2018); Maddela and Xu (2018); Zhao et al. (2018b); Gooding and Kochmar (2018); Vu et al. (2018); Gooding and Kochmar (2019); Surya et al. (2019); Qiang et al. (2020); Martin et al. (2020); Omelianchuk et al. (2021); Maddela et al. (2021); Qiang et al. (2021); Hazim et al. (2022); Martin et al. (2022); Sheang et al. (2022)
LLM Approaches	Chi et al. (2023); Nozza and Attanasio (2023b); Kew et al. (2023b); Trott and Rivière (2024); Scalercio et al. (2024); Mondal et al. (2024); Zečević et al. (2024); Tan et al. (2024); Hewett et al. (2024); Qiu and Zhang (2024); Zetsu et al. (2024); Asthana et al. (2024); Farajidizaji et al. (2024); Mo and Hu (2024); Barayan et al. (2025)
Question Generation	
Surveys	Kurdi et al. (2020); Al Faraby et al. (2024a); Flor (2025)
Datasets	Chen et al. (2018a)
Evaluation	Horbach et al. (2020); Xiao et al. (2023b); Gorgun and Bulut (2024); Deroy et al. (2025)
Pre-LLM Approaches	Flor (2025, ch4–ch9)
LLM Approaches	Flor (2025, ch10), Al Faraby et al. (2024b); Säuberli and Clematide (2024b); Scaria et al. (2024a); Kumar and Lan (2024)

Table 2: Additional references for reading tasks (assistive/assessment)

Spoken GEC & GED	
Evaluation	Lu et al. (2022); Qian et al. (2025b)
Datasets	Izumi et al. (2004); Caines et al. (2016); Kim et al. (2024a); Knill et al. (2024)
Pre-LLM Approaches	Izumi et al. (2003); Lee et al. (2011); Knill et al. (2019); Lu et al. (2019b,a, 2020, 2022); Bannò et al. (2023b, 2024a); Karanasou et al. (2025); Qian et al. (2025a,b)
LLM Approaches	Lu et al. (2025)
Spoken Language Assessment	
Surveys	Zechner and Evanini (2019)
Datasets	Menzel et al. (2000); Izumi et al. (2004); Yoon et al. (2009); Ishikawa (2014); Baur et al. (2017, 2018); Zhao et al. (2018a); Baur et al. (2019); Ishikawa (2019); Zhang et al. (2021); Coulange et al. (2024); Kim et al. (2024a); Knill et al. (2024)
Pre-LLM Approaches	Bernstein et al. (1990); Cucchiarini et al. (1997); Townshend et al. (1998); Franco et al. (2000); Xi et al. (2008); Qian et al. (2012); Malinin et al. (2017); Chen et al. (2018b); Evanini et al. (2018); Craighead et al. (2020); Raina et al. (2020); Peng et al. (2021); Wu et al. (2021a); Xu et al. (2021); Wang et al. (2021); Kim et al. (2022); Bannò et al. (2023a); McKnight et al. (2023); Park and Ubale (2023)
LLM Approaches	Fu et al. (2024); Phan et al. (2024); Bannò et al. (2025b); Ma et al. (2025); Voskoboinik et al. (2025)

Table 3: Additional references for speaking tasks (assistive/assessment)

Intelligent Tutoring Systems	
Surveys	Paladines and Ramirez (2020); Wollny et al. (2021); Wang et al. (2024e)
Datasets	Stasaski et al. (2020b); Caines et al. (2020); Suresh et al. (2022); Demszky and Hill (2023); Macina et al. (2023a)
Evaluation	Demszky et al. (2021); Vasselli et al. (2023); Jurenka et al. (2024); Maurya et al. (2025)
Pre-LLM Approaches	Evers and Nijholt (2000); Freedman (2000); Suraweera and Mitrovic (2002); Graesser et al. (2004, 2006); Weerasinghe and Mitrovic (2006); Dzikovska et al. (2010); D'Mello et al. (2012); Romero and Ventura (2013); Nye et al. (2014); Serban et al. (2020); Macina et al. (2023c)
LLM Approaches	Tack and Piech (2022); Tack et al. (2023); Vasselli et al. (2023); Wang and Demszky (2023); Sonkar et al. (2023); Lee et al. (2023); Markel et al. (2023); Daheim et al. (2024); Chowdhury et al. (2024); Wang et al. (2024b,c); Denny et al. (2024); Wang et al. (2024a); Nie et al. (2025); Srivatsa et al. (2025)

Table 4: Additional references for tutoring tasks (assistive/assessment)