Incentives Of EdTech: A Systematic Review Of EduNLP Research

Gabrielle Gaudeau[1], Aoife O'Driscoll[1], Jasper Degraeuwe[2], Andrew Caines[1], Donya Rooein[3], Zeerak Talat[4]

[1]ALTA Institute, Computer Laboratory, University of Cambridge (UK),
[2]Ghent University (Belgium), [3]Bocconi University (Italy), [4]University of Edinburgh (UK)

Correspondence: gjg34@cam.ac.uk, ao514@cam.ac.uk, jasper.degraeuwe@ugent.be, apc38@cam.ac.uk, donya.rooein@unibocconi.it, z@zeerak.org

Abstract

The global teacher shortage is pushing schools and institutions towards an ever-greater reliance on artificial intelligence. While the Natural Language Processing community has dedicated significant resources in developing educational technologies (EdTech) that support this shift, it remains unclear whose interests are being best served among the stakeholders of education.

In this paper, we present a systematic literature review of 204 papers published in venues of the Association for Computational Linguistics' Special Interest Group on Building Educational Applications in 2024 and 2025, and validate these against EdTech papers from the wider ACL Anthology. By examining stakeholder inclusion and the prioritisation of research tasks, our findings reveal a critical tension: a push and pull between private-sector incentives and the foundational needs of educational infrastructure. Our analysis reveals that teachers are systematically under-represented as beneficiaries of research (33.3%) despite being the most affected, that real-world deployment remains rare (9.8%), and that ethical engagement tends toward acknowledgement rather than action. Drawing on exemplary papers in our corpus, we offer concrete recommendations for more responsible EduNLP research practices.

1 Introduction

Education has long been a domain of inspiration for Artificial Intelligence (AI) and Natural Language Processing (NLP). From early feature-based auto-markers (e.g., *e-rater*®; Attali and Burstein, 2006) to large language model (LLM)-powered intelligent tutoring systems (ITS) (e.g., Khanmigo[1] by Khan Academy), the goals have remained constant: for technology to extend the reach of good teaching and to support learners who might otherwise go without. These are meaningful goals – socially urgent, technically challenging, and worthy of scientific investment – and their urgency has only grown in recent years with global teacher shortages (UNESCO, 2026), widening equity gaps (World Inequality Lab, 2026), and the rapid uptake of commercial AI products for education (Gomes, 2026). Held together, they have made the question of the role of technology in supporting education more pressing than ever.

There is a particular risk that comes with being deeply embedded in a fast-moving research area: the closer we are to the technical problems in front of us, the easier it is to lose sight of the overarching goal. As researchers, we are drawn towards the datasets we know, the metrics we trust, the tasks where progress is legible. Specialisation is necessary, but it can quietly narrow the frame of reference until the question, "Does this system work?", crowds out the most important question: "Does this actually serve the people we said we were building it for?" This paper is, in part, an attempt to step back from that narrowing and ask plainly: as a field, are we meeting our own aspirations?

To answer this question, we conduct a systematic literature review of EduNLP research. We survey 204 papers published in 2024 and 2025 at ACL SIGEDU venues (BEA[2] and NLP4CALL[3] workshops) and the main *ACL conferences. To the best of our knowledge, this is the first systematic review of EduNLP research that focuses on publications in the ACL Anthology. For each paper, we examine its tasks, motivations, stakeholder inclusion, incentives, and engagement with ethical risks to answer three research questions:

RQ 1 Which tasks are prioritised in EduNLP research, what motivates them, and in which contexts are the resulting systems deployed?

[1] https://www.khanmigo.ai

[2] Workshop on Innovative Use of NLP for Building Educational Applications

[3] Workshop on NLP for Computer-Assisted Language Learning

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 715–750

RQ 2 Who are the stakeholders of EduNLP research, how are they included, and whose interests does the research serve?

RQ 3 What risks, concerns, and limitations are raised, and to what extent does the research mitigate them?

Our findings show that teachers are systematically under-represented as beneficiaries in EduNLP research, real-world deployment is rare, and ethical engagement tends toward acknowledgement rather than action. We identify exemplary counter-examples and derive from them a set of concrete recommendations for the field.

2 Related Work

Education has been a domain for innovation dating back millennia. Digital technology is a modern feature of this long history: much of the early pioneering work on AI in the twentieth century was directed towards educational aims and applications in AIED (Newell et al., 1958; Minsky, 1974; Papert, 1980; Doroudi, 2023). In recent years the growth of interest in LLMs has also seen increasing application to education (Caines et al., 2023; Davis et al., 2024; Pack et al., 2024), further evidenced by the growing popularity of the annual Workshop on Innovative Use of NLP for Building Educational Applications (BEA), the foundation of the ACL SIGEDU in 2017,[4] and investment by large technology firms into products such as Google's LearnLM[5] and OpenAI's ChatGPT Edu.[6]

EdTech covers a wide-range of applications for educational purposes, often involving AI or NLP. There have been several surveys on EdTech and its use in various domains (Ahmad et al., 2024; Benedetto et al., 2023; Hidayat and Firmanti, 2024) spanning classroom support, virtual learning environments, websites, and tutoring chatbots. In this paper, we focus on ethical matters, which have received growing attention in AI and NLP more broadly, including the identification of different bias types throughout the "machine learning life cycle" (Suresh and Guttag, 2021).

Within EdTech, several surveys and position papers have addressed ethical issues. Yan et al. (2025) presents a systematic review of 34 publications involving EdTech with AI in schools or higher education from 2020-2024, reporting a "constellation of recurring ethical tensions" relating to algorithmic bias, data privacy, transparency, accountability, and academic integrity. They observe that these are known issues with AI applications, and recommend co-design with stakeholders, an emphasis on explainability, regulatory improvements, and AI literacy training for teachers. Alfredo et al. (2024) arrive at similar conclusions from a review of 108 papers relating to human-centred or participatory design and learning analytics.

Fu and Weng (2024) conduct a systematic review of empirical studies focused on EdTech and responsible AI, making similar conclusions to Yan et al. (2025) based on 40 selected papers. They present a vision for "responsible human-centered AIED" which includes core principles of Fairness and Equity, Transparency and Intelligibility, Agency and Autonomy, Privacy and Security, and Beneficence and Non-maleficence. Holmes et al. (2022) surveyed EdTech researchers, reporting high interest in but low confidence about ethical issues, attributed to a lack of ethics training in AI-related courses. They propose a framework for ethics in AIED aimed at ensuring "ethical by design" research, and emphasise the importance of cross-disciplinary engagement. Taken together, these reviews converge on a shared diagnosis: ethical considerations are widely recognised in principle but inconsistently integrated in practice.

This review extends these prior work by including research published throughout 2025, and by considering tasks, contexts, stakeholders, incentives, and risks across 204 EduNLP papers from *ACL main conference and workshop proceedings.

3 Methodology

Search Protocol. We collected all papers from the BEA and NLP4CALL workshops published in 2024 and 2025. We also conducted a search of the ACL Anthology using the Anthology API[7] for papers published in main *ACL and associated conferences whose title or abstract contained at least one of 38 EduNLP-relevant search terms (e.g., "student modeling"; see Appendix B for a complete list of venues and search terms).[8] This sampling approach affords an in-depth view into contemporary trends at the expense of longitudinal analyses. This search resulted in 191 papers from the two workshops, and 316 papers from *ACL conferences.

[4] https://sig-edu.org/

[5] https://cloud.google.com/solutions/learnlm

[6] https://openai.com/chatgpt/education/

[7] https://acl-anthology.readthedocs.io/py-v0.5.3/api/

[8] The search was conducted on January 21, 2026.

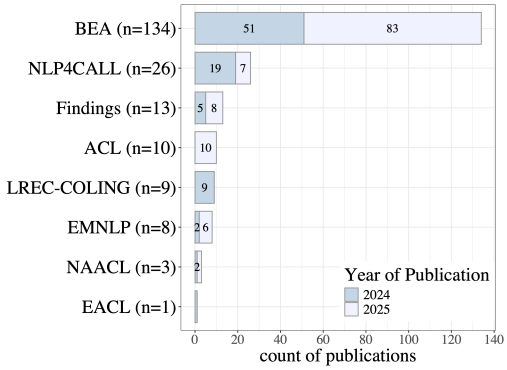

Figure 1: Number of papers per venue and year. We reviewed a total of 204 papers (160 BEA+NLP4CALL papers and 44 ACL Anthology main conference papers).

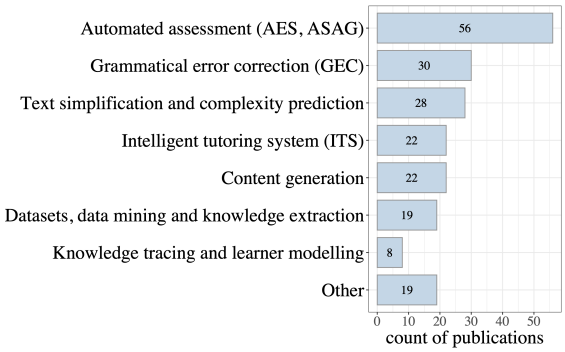

Figure 2: Number of papers per high-level task. See Table 8 for the detailed mapping.

For BEA and NLP4CALL, we randomly sample 25% of contributions for each shared task, with a minimum sampling threshold of 5 papers for each task. We further include all shared task overview papers, as these represent a qualitatively distinct type of contribution. For the *ACL main conference papers, we reviewed all abstracts for relevance to educational applications, excluding 214 papers as non-relevant. The remaining 102 papers were stratified by publication year, venue, and search term, yielding a sample of 44 papers. This resulted in a final sample of 160 papers from BEA and NLP4CALL workshops, and 44 papers from *ACL conferences, for a total of 204 papers (see Table 8 for paper details). Figure 1 shows the distribution of papers across venues and years.

Data extraction. Data extraction was conducted manually by three of the authors using a shared extraction schema (see Appendix C). The schema captures: the specific task addressed; datasets used and their availability; the explicit motivation for the research; stakeholders mentioned and included (with associated quotes); the level of stakeholder inclusion; the deployment context of any system; incentives (both explicit and implicit) that the research serves; ethical risks and concerns raised; measures taken to address those risks; and future directions pertaining to risk, ethics, or aspiration.

Extraction proceeded in three phases. In the first phase (1), a single paper was annotated collaboratively to develop and validate the schema. In the second phase (2), annotators independently reviewed a shared batch of 25 papers,[9] meeting to discuss schema revisions and resolve ambiguities. Note that phase (2) was conducted in an iterative manner: following phase (1), each time the schema was modified or extended, all annotators updated their previous phase (2) annotation to reflect the revised guidelines. In the third and final phase (3), the remaining papers were reviewed independently by three authors. Extracting data took an annotator on average 45 minutes per paper (ranging between 30–60 minutes); we estimate that the review took a combined total of about 190 hours to complete.

Agreement. Inter-annotator agreement (IAA) was measured on phase (2)'s independently reviewed shared batch. Table 1 in Appendix D shows the agreement for the free-text dimensions of our schema based on the Percentage Agreement (PA; Roaché, 2017) measure (see Tables 6 and 7 for illustrations of how free-text agreement was computed). For the four multi-label dimensions, we report both Krippendorff's α (Krippendorff, 2011) and PA in Tables 2, 3, 4 and 5.

For the free-text fields, PA ranges between 0.52 (for implicit incentives) and 1 (for deployment). For the multi-label dimensions, PA was consistently high (0.84–94 overall), while α was more variable. Agreement on the presence of stakeholders was generally moderate to strong (α = 0.49–0.7 overall, with agreement on teachers being particularly high at α = 0.79–0.84). Agreement on stakeholder inclusion level and risk engagement level was lower (α = 0.52–0.61 overall). Taking into account the qualitative and inherently interpretative nature of the annotation task (especially for dimensions such as risks/concerns), we consider these agreement values to be sufficiently high to

[9]The shared batch was a stratified sample from our corpus of 204 papers (12.3%) based on venue and year of publication; it included 6 BEA 2024, 10 BEA 2025, 2 NLP4CALL 2024, 1 NLP4CALL 2025, 1 EACL 2024, 1 LREC-COLING, 1 NAACL 2025, 1 ACL 2025, and 2 Findings 2025 papers.

justify the independent reviewing in phase (3).

4 Tasks, Motivations, Deployment

Tasks. Figure 2 shows the distribution of high-level tasks across our corpus of papers. Automated assessment – i.e., automated essay scoring (AES) and automated short-answer scoring (ASAG) – is by far the most common task (56 papers), followed by grammatical error correction (GEC, 30 papers) and text simplification and complexity prediction (28). Content generation (22), intelligent tutoring systems (ITS, 22 papers), dataset creation and knowledge extraction (19), and knowledge tracing and learner modelling (8) are also represented. The "Other" task type includes a variety of research, most often relating to the novel capabilities of LLMs (e.g., multimodal assessment, alignment with human eye-tracking data, and discourse evaluation) and detecting LLM-generated texts.

The dominance of language assessment and feedback tasks is striking: taken together, AES/ASAG and GEC account for almost half of the corpus. This reflects a longstanding priority in EduNLP: indeed, automated assessment has been an active area of research for decades, benefitting from well-established datasets (e.g., ASAP; Hamner et al., 2012). However, this prevalence also raises questions about whose priorities are being served: automated assessment and feedback tools are of direct commercial value to large-scale testing organisations and EdTech companies.

Shared tasks. The NLP4CALL 2025 shared task introduced multilingual GEC (Masciolini et al., 2025), a direction of particular importance given that GEC, while already the second most represented task in our corpus, has historically been dominated by English-language systems. Broadening GEC to multilingual settings introduces non-trivial challenges around low-resource languages, cross-lingual transfer, and the availability of annotated learner corpora, and a shared task framing is well-suited to mobilising community effort around these barriers. On the other hand, the BEA 2024 shared tasks addressed automated prediction of item difficulty and response time (Yaneva et al., 2024a), and multilingual lexical simplification (Shardlow et al., 2024); the 2025 shared task addressed pedagogical ability assessment of AI-powered tutors (Kochmar et al., 2025). We note that all three of these problems receive less attention in the non-shared-task literature.

This suggests that shared tasks are playing a valuable role in broadening the community's agenda, including towards less commercially obvious but educationally important problems such as pedagogical quality assessment, and towards underserved languages in otherwise established tasks. Beyond their immediate proceedings, shared tasks also exert a longer-lasting influence through the datasets they produce; resources like the W&I+LOCNESS dataset which was introduced for the BEA 2019 Shared Task on GEC (Bryant et al., 2019) tend to attract sustained reuse by the community (as illustrated by Figure 3), and thus continue to shape which problems remain visible and tractable long after the shared task itself has concluded.

Datasets. Papers in the corpus reported using 284 distinct datasets used a combined total of 460 times (373 for public datasets, 33 for those available upon-request and 54 for private datasets). Figure 11 shows that 73.9% of datasets used are publicly available, 7.4% are only available upon-request or through paid licences, and 18.7% are private. While the high proportion of public datasets is a positive indicator for reproducibility, Figure 3 reveals a high concentration of usage around a small number of datasets: the top three – W&I+LOCNESS (Bryant et al., 2019), ASAP (Hamner et al., 2012), and CoNLL-2014 (Ng et al., 2014) – together account for 12.9% of total public dataset usage (373), with a long tail of datasets used only once. This concentration partially reflects the task distribution noted previously: namely that AES and GEC are both well-established. However, this also raises questions about whether research findings generalise beyond the narrow slice of learner populations, languages, and educational contexts that these datasets represent. We return to this concern in Section 7.

Motivations. During extraction, we took note of the explicit motivation presented by papers for their presented research, and later classified each into one or more of seven high-level categories. Figure 12 shows that the most common motivation type across our corpus is to "help a stakeholder" (110 papers), followed by addressing a pedagogical or ethical concern (82), and assuming the role of a stakeholder (53). Technical motivation alone, with no stated stakeholder benefit, accounts for 43 papers, which is a non-trivial proportion (21.1%). Figure 4 reveals the stakeholder composition underlying papers' motivations: learners and students are

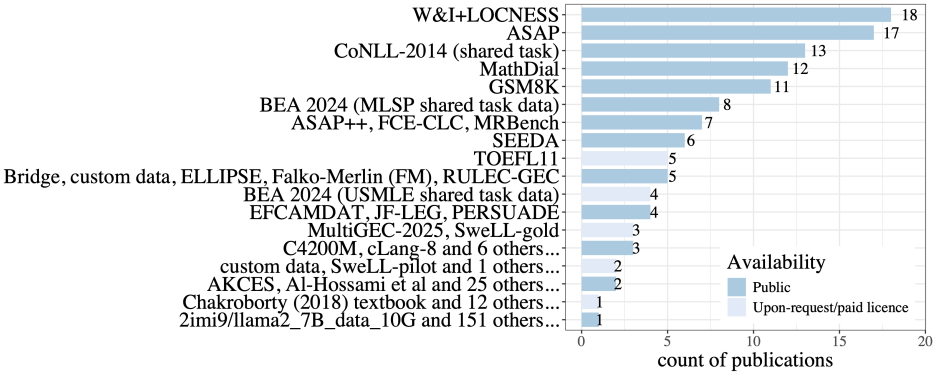

Figure 3: Dataset popularity (i.e., the number of times a dataset was used, and not only mentioned). We do not report private datasets given their absence of references.

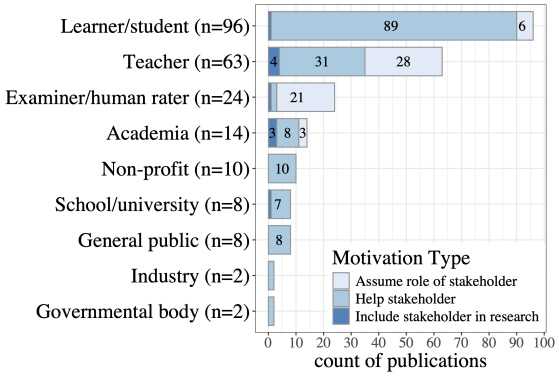

Figure 4: Distribution of different stakeholders for the three stakeholder-based motivations in Figure 12.

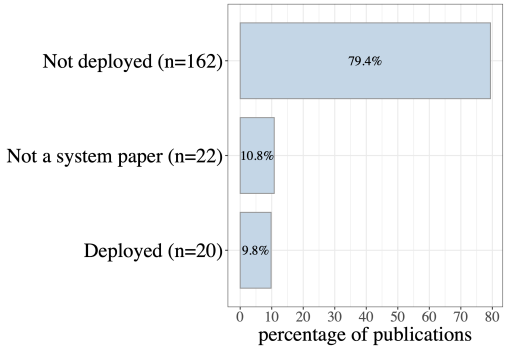

Figure 5: Papers that deployed their method to real-world users or tested it on pre-existing real-world data.

invoked in 61.1% of papers with stakeholder-based motivation (96 of 157 papers), making them by far the most frequently cited intended beneficiary. Teachers appear in 40.1% of such papers (63 of 157 papers), though they are most commonly invoked as a pressure points, referenced in terms of the cost, time, or burden associated with their labour, and implicitly positioned as a bottleneck that automation should relieve. This framing matters. A motivation to reduce teacher burden through automation is meaningfully different from one that seeks to augment teacher capability or support teacher agency. In a number of papers in our corpus, teachers appear in the motivation but then disappear from the research design entirely: they are not consulted, included in evaluation, or named as beneficiaries of the results. We discuss this pattern and its implications further in Section 6.

Context deployment. Figure 5 shows that 79.4% of papers (162 papers) present systems or models that are never deployed to real-world users. Only 9.8% of papers report genuine deployment. We label resource and survey papers as "Not a system paper." Non-deployment is not itself a failing: fundamental research that develops methods, datasets, or evaluation frameworks may legitimately precede any deployment. More concerning is that papers describing non-deployed systems rarely discuss the pathway to deployment: the educational contexts in which the system might operate, the stakeholders who would need to be involved, or the risks real-world deployment would introduce. This creates a body of research that is optimised for benchmark performance in conditions that may bear little resemblance to the classrooms, tutoring sessions, and assessment environments it nominally serves.

5 The Roles of Stakeholders

Author affiliations and acknowledged entities. Figure 15 shows that the paper author affiliations in our corpus are geographically concentrated: the United States accounts for the largest single-country share of author affiliations (58 papers), followed by Germany (29 papers), China (23 papers),

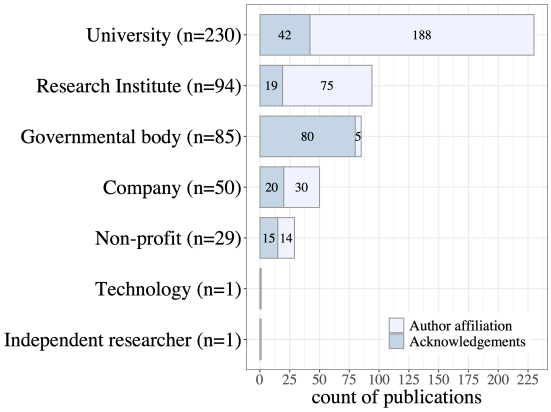

Figure 6: Number of papers per type of author affiliation and type of entity mentioned in acknowledgements.

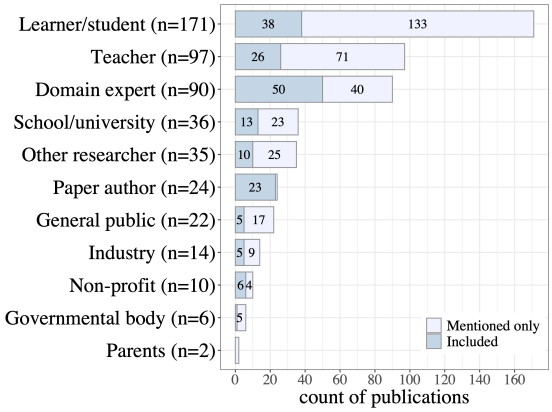

Figure 7: Number of papers per type of stakeholder *included* or *mentioned only* in the research.

and other European countries (similar observations can be made on the origin of the acknowledged entities in Figure 14). Figure 6 shows that universities dominate author affiliations (188 papers), followed by research institutes (75) and companies (30). Funding acknowledgements are concentrated within governmental bodies (80), with national science foundations of China and the US appearing the most frequently (Figure 16). Industry acknowledgements (e.g., Microsoft) appear in a small but non-trivial number of papers (20). While industry involvement in research funding is not inherently problematic, it creates potential conflicts of interest that deserve explicit discussion, particularly in a field where commercial EdTech products are directly shaped by research agendas. Notably, few papers in our corpus explicitly disclose or discuss potential conflicts of interest arising from their funding sources; a gap that mirrors findings in adjacent fields (Garrett et al., 2020).

Stakeholders mentioned or included. Figure 7 shows that learners and students are mentioned in the most papers overall (170 papers), followed by teachers (97), and domain experts (88). However, mention does not equate to inclusion: the proportion of mentioned stakeholders who are also actively included in the research is substantially lower across all groups. Among teachers, 26.8% of papers that mention them also include them in the research (26). For learners, 22.4% of mentioning papers include them (38). Domain experts show a much higher inclusion rate (56.8%), in part because they are frequently recruited as annotators or raters. Most strikingly, parents were only mentioned in two papers, despite their having such an important

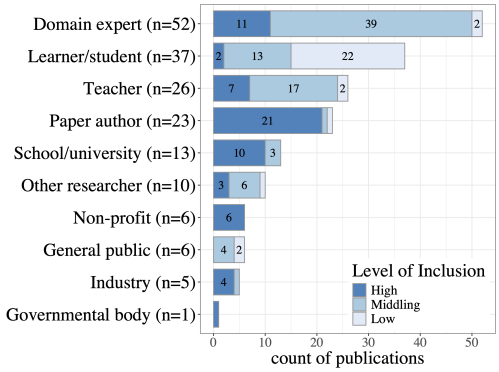

Figure 8: Level of inclusion of included stakeholders by stakeholder type; we distinguish 3 levels: *High* (integral to research design & completion), *Middling* (involved in data evaluation or annotation, without input on research design), and *Low* (test subjects in data collection only).

role in children education (Kostov, 2026).

Figure 13 reveals the overall distribution of inclusion levels across all included stakeholders: 47.0% of inclusions are classified as *Middling* (involved in data evaluation or annotation, but with no input on research design), 32.1% as *High* (integral to research design and completion), and 20.9% as *Low* (test subjects in data collection only). Figure 8 shows that this breakdown varies substantially by stakeholder type. Other than paper authors themselves, schools and universities are most likely to be included at a *High* level (76.9%), while teachers, when included at all, are predominantly included at a *Middling* level (65.5%), most often as annotators. Learners are most often included as test subjects (59.5%). The implication is that even when stakeholders are formally included, they are rarely positioned as agents who shape the research, they are more often positioned as instruments of it.

Incentives. Figure 9 shows the distribution of stakeholders explicitly mentioned as benefiting from the research alongside those we identified as implicit beneficiaries. We note that the identification of implicit beneficiaries is the most subjective dimension of our annotation: it required annotators to infer who stands to gain from a piece of research beyond what authors themselves state, based on the nature of the task, the deployment context, and the funding sources involved. For instance, a paper developing an AES system for standardised testing, funded by a testing organisation, was coded as implicitly benefiting industry, even if no such benefit was named. Due to the subjective nature of this dimension, inter-annotator agreement was accordingly lower (0.53; Table 1), and these findings should be read as indicative rather than definitive.

Learners and students are the most frequently named explicit beneficiary (125 papers). Teachers stand out starkly here: 80.9% of their appearances are explicit (55 papers). Stated differently, teachers are almost never the unstated but evident beneficiary of research; when they benefit, papers say so. However, the vast majority of papers do not position them as benefiting at all. On the other hand, non-profit organisations, industry and governmental bodies appear prominently as implicit beneficiaries. That is, while they are not named in the paper as intended beneficiaries, the research clearly serves their interests. This is most visible in the task-level breakdown in Figure 17: automated assessment research (the largest task category in the corpus) consistently benefits learners and industry, while teachers and examiners are sparsely represented. The commercial relationship here is direct: automated scoring tools reduce the need for human markers and are of clear value to large-scale testing organisations. For ITS, learners dominate, with limited acknowledgement of teachers. GEC research shows the broadest stakeholder spread, in part because GEC tools serve not only learners and teachers but also the general public who use writing assistance tools in everyday tasks.

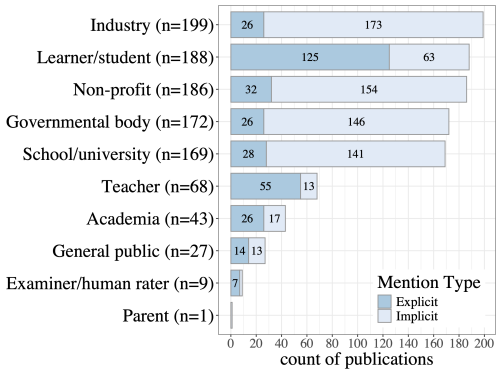

Figure 9: Stakeholders explicitly stated as benefitting from the research, as well as those that we could see benefitting that were not explicitly mentioned (*Implicit*). Note that a stakeholder may be both explicitly mentioned to benefit in some way and implicitly in another.

6 Risks, Concerns, Limitations, and Measures Taken

Risks, concerns and limitations raised. Figure 18 shows the distribution of risks, concerns, and limitations explicitly raised by paper authors, organised into six high-level categories. We note that inter-annotator agreement was lower for this dimension than others (0.57; Table 1), owing to the need to assess coverage across a large and varied set of concerns; these results should therefore be read as indicative trends rather than precise counts. The most commonly noted concerns are methodology limitations (69 papers), dataset limitations (60), followed by lack of generalisability and language-specificity (56), risk of bias (46) and and task/domain-specific limitations (44), reflecting the tendency of research to develop systems for specific languages or educational contexts that may not transfer. Several important risk categories are raised much less frequently. Risk of hallucination appears in only 12 papers, risk of dual-use in 6, and safety concerns in 26. Within the contextualising research category, the gap between research and real-world application is noted in 32 papers and the need for human-in-the-loop in 19, suggesting some awareness of deployment limitations, this rarely translates into direct mitigation (Figure 19). Data protection and anonymisation concerns are raised in 37 papers, while informed consent and fair compensation for included stakeholders, critical ethical requirements for human-subjects research, appear in only 11 and 10 papers respectively. That human-subjects protections remain among the least commonly raised concerns in a corpus that routinely collects learner data and recruits human annotators is itself a notable finding.

Engagement with risks. Figure 19 distinguishes three levels of engagement with stated risks: *High* (directly mitigated or discussed in substantial depth), *Middling* (discussed as part of future work), and *Low* (briefly mentioned only). Across most risk

categories, the majority of engagement is at a *Low* or *Middling* level. *High* engagement is most consistently found in the participant and data concern category: fair compensation for included stakeholders (100.0%) and informed consent (72.7%) are the most actively addressed concerns, though both are raised by relatively few papers to begin with. By contrast, the largest categories show the weakest engagement: methodology limitations are 98.6% *Middling* or *Low*, and dataset limitations 90.0% *Middling* or *Low*. Risk of bias, one of the most frequently raised concerns at 46 papers, is engaged at a *High* level in only 15.2% of cases. The gap between research and real-world application and the need for human-in-the-loop, two concerns with clear implications for responsible deployment, are predominantly *Middling* or *Low*. This pattern suggests a community that is aware of the ethical dimensions of its work but has not yet developed consistent norms for acting on them within the scope of individual papers.

Future work. Figure 20 shows a distribution of the areas of future work explicitly mentioned in the papers. We report future work specifically related to any risks, concerns or higher aspirations rather than any purely technical work; of our data sample, 21 papers do not discuss any such future work. Four high-level themes emerge within discussed future work: stakeholder inclusion, technical development, expanding the scope of the research, and engaging with issues emerging from the research. Of these, the most frequently mentioned category is expanding the scope of the research, with expanding the data (42), language selection (36), and subject domain (35) the most common fine-grained directions. EduNLP research is often performed at language- or task-specific levels, resulting in common limitations which translate to clear future directions. The least common high-level category is engaging with issues emerging from research, with fine-grained categories including interpretability and bias mitigation (16 and 14 papers respectively), exploring performance-cost trade-offs (12), and initiating broader discussions in the EduNLP space. Within the fine-grained categories overall, the most frequently referenced future direction is general technical improvements related to the paper's risks and concerns (91 papers). Despite controlling for purely technical work in our analysis, the primary focus for EduNLP researchers remains within this domain. Within the stakeholder inclusion category, user study and user inclusion (37) and integration into real-world systems (32) are the most common directions, suggesting some awareness that current work falls short. Analysis of the future directions in our sample therefore reveals a tendency towards prioritising empirically-motivated fine-grained technical work rather than ethically-driven broader work. In part, this may be due to an imbalance in available resources for conducting such research.

7 Discussion: Opportunities, Recommendations, Aspirations

Opportunities. Our findings reveal some structural gaps in EduNLP research that constitute genuine opportunities for the field. First, teachers are under-represented both as beneficiaries and active participants, despite their central role in education. This represents a significant misalignment between stated purpose and actual design. Research that nominally aims to support education but systematically excludes the professional educators who mediate it risks building tools that are technically sophisticated but pedagogically ill-fitting, or that automate away precisely the human judgement that makes good teaching effective. Second, the gap between research development and real-world deployment is striking: only 9.8% of system papers are deployed in live educational settings. This reflects a missing discourse about what responsible deployment looks like: which stakeholders need to be involved, what evaluation is appropriate for real students and teachers, and what accountability mechanisms should be in place. This gap is further sharpened by the concentration of datasets around high-stakes standardised testing contexts, and by the dominance of assessment tasks which, given their direct commercial value to the testing industry, risk pulling the research agenda toward institutional efficiency over the full range of educational stakeholders.

Recommendations. Drawing on exemplary papers in our corpus, we offer three concrete recommendations for the EduNLP community:

1. **Co-design with teachers and learners from the outset.** Research that positions stakeholders as genuine co-designers, rather than test subjects or future-work items, produces better-grounded systems and more honest evaluation. Galletti and Cesaroni (2025) offer a replicable model for this: conducting focus

groups and questionnaires with teachers at different stages of system development surfaced concerns around transparency, autonomy, and pedagogical alignment that would not have emerged from technical evaluation alone. See also Huovinen and Hämäläinen (2025). Their work echoes principles of design justice (Costanza-Chock, 2020) which seek to decentre technical expertise in favour of lived experience and domain expertise – in all regards save technical implementation – as a mechanism for ensuring that those affected by a system retain meaningful agency in shaping it. As it stands, a true expression of design justice was not found in any of the reviewed papers of this corpus, however Wang et al. (2025c) embodies some aspects of it. Though not a system, the paper demonstrates that design justice principles can be embedded even at the resource creation stage: their math world problem benchmark was developed through structured interviews with primary school math teachers, whose pedagogical expertise directly shaped what counts as a meaningful visual, ensuring that future systems trained or evaluated on this benchmark will be held to a standard defined by them.

2. **Make deployment contexts and costs explicit.** Authors should describe the educational context in which their system could or has been deployed, the stakeholder roles involved, and provide an honest account of computational, financial, and human costs alongside claimed benefits (Akter et al., 2025; Gupta et al., 2025; Li and Ng, 2024).

3. **Adopt structured ethical reflection and act on it.** Our data show that named concerns rarely translate into mitigation within the same paper. Venues should normalise the expectation that ethical risks raised are addressed in the current work, not deferred to future work. Checklists like the ARR Responsible NLP Checklist already support this: they prompt authors to interrogate their own design choices (e.g., Goto et al., 2025a, who voluntarily engage in detailing a number of the checklist items) and give reviewers a structured basis for evaluating ethical engagement.

Aspirations. The potential for AI in education is genuine: it could improve access to education, helping reduce inequalities related to geography, language, resources, and infrastructure, for learners who might otherwise go without. It could also help free educators of repetitive and time-consuming tasks so they can concentrate on the relational aspect of education that systems cannot and should not replace. Automated tools can also mitigate some human weaknesses that threaten fairness in assessment: fatigue, inconsistency, and unconscious biases. The question is not whether AI belongs in education, but whether we are stewarding its development responsibly. The exemplary papers in our corpus demonstrate that it is possible. Our aspiration for the field is a research community that treats educational infrastructure as a site of social responsibility, not merely technical opportunity. The trajectory the field takes will depend on choices that are made now about which tasks to prioritise, whose voices to include, and what counts as success. Harding (2025), reflecting on AI in language assessment, frames this as a choice between utopian and dystopian futures: one in which assessment technology is context-sensitive, transparent, connected with learning, and deeply oriented toward justice; as opposed to one driven by expediency, opacity, and the logic of scale.

8 Conclusion

This paper has presented a systematic review of 204 EduNLP papers published at ACL SIGEDU venues and main *ACL conferences in 2024 and 2025, examining tasks, motivations, stakeholder inclusion, incentive structures, and ethical engagement. Our analysis reveals a field that is technically productive but structurally misaligned with key educational stakeholders, particularly teachers who are rarely included in research and almost never positioned as implicit beneficiaries. At the same time, our corpus contains exemplary work that demonstrates what responsible, stakeholder-grounded EduNLP research looks like in practice. The norms and practices embedded in these papers are neither technically burdensome nor novel in principle. What is needed is for the community to adopt them consistently, and for publication venues to create the conditions in which doing so is expected rather than exceptional. We hope this review serves as both a diagnostic and a resource: a map of where the field currently stands, and a set of orientations for where it should go.

9 Limitations

This review has several limitations that should be noted. First, while our corpus of 204 papers is broad in scope, it is not exhaustive, meaning that some relevant papers will have been missed. Our focus on ACL Anthology venues also means that work published in AIED journals, learning analytics conferences, and EdTech-specific venues falls outside our scope: the picture we paint is of the NLP community specifically, not the broader field. Second, annotation of inherently interpretive dimensions, particularly stakeholder inclusion level and risk engagement level, carries subjectivity that agreement scores can only partially represent. We report these as indicative trends rather than precise counts, but readers should bear this in mind when interpreting figures. Third, our corpus covers 2024–2025 only; while this captures the most recent work, it is a short window and trends may not generalise to earlier or future periods. An interesting direction for future work would be to extend the temporal scope to publications published before the release of ChatGPT (OpenAI, 2026) in November 2022, which would allow for a direct comparison of research priorities, stakeholder inclusion, and ethical engagement before and after the widespread availability of generative AI. Finally, as researchers embedded in the EduNLP community ourselves, we are not neutral observers, our framing of what constitutes meaningful stakeholder inclusion or adequate ethical engagement reflects our own values, which we have tried to make explicit throughout.

On financial disclosures. While complete financial disclosures of which entities have funded research is a desirable trait in papers due to the transparency it affords, disclosure can be structurally limited. For example, some grant funders – particularly military funding – may require non-disclosure, nation-wide regulation may limit disclosure, research can be funded across multiple grants, work may be conducted on an entirely voluntary basis, among many other reasons. We therefore see financial disclosure as a spectrum between complete opacity and complete transparency. We advocate for researchers to approach the question of financial disclosure according to a maximalist approach, i.e., we argue that researchers should share as much information as is possible to them in a given situation. A transparency maximalist approach will afford greater insight into how research into educational technologies is being, forcefully and subtly, shifted by the interests of different entities.

10 Ethical considerations

All papers surveyed in this review are publicly available through the ACL Anthology; no private or unpublished materials were used. No human subjects were involved in the review itself. The annotation process involved researchers reading and characterising the work of others, which carries a risk of misrepresentation; we have sought to mitigate this through iterative schema development, inter-annotator agreement measurement, and the use of direct quotes to ground our characterisations. Our normative claims – that teachers are under-served, that ethical engagement is insufficient, that commercial incentives distort research agendas – are recommendations and observations, not accusations about individual papers or authors. We acknowledge that we are ourselves part of the community we critique, and that future reviews may find similar gaps in our own work. This paper has been pre-registered on OSF.[10]

Acknowledgements

Gabrielle Gaudeau, Aoife O'Driscoll and Andrew Caines are supported by Cambridge University Press & Assessment. Donya Rooein is a member of the MilaNLP group and the Data & Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Her research is supported through the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). We thank the anonymous reviewers for their time and valuable feedback. Finally, we note that Claude Sonnet 4.6[11] was used to improve the language and readability of the manuscript.

References

Kashif Ahmad, Waleed Iqbal, Ammar El-Hassan, Junaid Qadir, Driss Benhaddou, Moussa Ayyash, and Ala Al-Fuqaha. 2024. Data-driven artificial intelligence in education: A comprehensive review. *IEEE Transactions on Learning Technologies*, 17:12–31.

Soroosh Akef, Detmar Meurers, Amália Mendes, and Patrick Rebuschat. 2025. Interpretable machine learning for societal language identification: Modeling English and German influences on Portuguese

[10] https://osf.io/nhb2q/overview?view_only=533ca23658d644a4abaf0bbd7e63087c

[11] https://www.anthropic.com/claude/sonnet

heritage language. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 50–62, Tallinn, Estonia. University of Tartu Library.

Syeda Sabrina Akter, Seth Hunter, David Woo, and Antonios Anastasopoulos. 2025. Costs and benefits of AI-enabled topic modeling in P-20 research: The case of school improvement plans. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 460–476, Vienna, Austria. Association for Computational Linguistics.

Riordan Alfredo, Vanessa Echeverria, Yueqiao Jin, Lixiang Yan, Zachari Swiecki, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6:100215.

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

David Alfter. 2025. The need for truly graded lexical complexity prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.

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.

Mina Almasi and Ross Deans Kristensen-McLachlan. 2025. Alignment drift in CEFR-prompted LLMs for interactive Spanish tutoring. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 70–88, Vienna, Austria. Association for Computational Linguistics.

Jiyuan An, Xiang Fu, Bo Liu, Xuquan Zong, Cunliang Kong, Shuliang Liu, Shuo Wang, Zhenghao Liu, Liner Yang, Hanghang Fan, and Erhong Yang. 2025. BLCU-ICALL at BEA 2025 shared task: Multi-strategy evaluation of AI tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1084–1097, Vienna, Austria. Association for Computational Linguistics.

Aitor Arronte Alvarez and Naiyi Xie Fincham. 2025. Automated L2 proficiency scoring: Weak supervision, large language models, and statistical guarantees. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 384–397, Vienna, Austria. Association for Computational Linguistics.

Yuya Asano, Beata Beigman Klebanov, and Jamie Mikeska. 2025. Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 716–736, Vienna, Austria. Association for Computational Linguistics.

Nischal Ashok Kumar and Andrew Lan. 2024. Improving socratic question generation using data augmentation and preference optimization. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 108–118, Mexico City, Mexico. Association for Computational Linguistics.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Sarra El Ayari and Zhongjie Li. 2024. Potential of ASR for the study of L2 learner corpora. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 49–58, Rennes, France. LiU Electronic Press.

Nicolas Ballier and Adrien Méli. 2024. Investigating acoustic correlates of whisper scoring for L2 speech using forced alignment with the Italian component of the ISLE corpus. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 20–32, Rennes, France. LiU Electronic Press.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Stefano Bannò, Hari K. Vydana, Kate M. Knill, and Mark J. F. 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.

Beata Beigman Klebanov, Michael Suhan, Tenaha O'Reilly, and Zuowei Wang. 2024. From miscue to evidence of difficulty: Analysis of automatically detected miscues in oral reading for feedback potential. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 459–469, Mexico City, Mexico. Association for Computational Linguistics.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Luca Benedetto, Shiva Taslimipoor, and Paula Buttery. 2025. A survey on automated distractor evaluation in multiple-choice tasks. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 55–69, Vienna, Austria. Association for Computational Linguistics.

Santiago Berruti, Arturo Collazo, Diego Sellanes, Aiala Rosá, and Luis Chiruzzo. 2024. Automatic crossword clues extraction for language learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 381–390, Mexico City, Mexico. Association for Computational Linguistics.

Marie Bexte, Yuning Ding, and Andrea Horbach. 2025. Increasing the generalizability of similarity-based essay scoring through cross-prompt training. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 225–236, Vienna, Austria. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence? – exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124, Mexico City, Mexico. Association for Computational Linguistics.

Marie Bexte and Torsten Zesch. 2025. Is lunch free yet? overcoming the cold-start problem in supervised content scoring using zero-shot LLM-generated training data. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 144–159, Vienna, Austria. Association for Computational Linguistics.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. Leveraging LLMs for Bangla grammar error correction: Error categorization, synthetic data, and model evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8220–8239, Vienna, Austria. Association for Computational Linguistics.

Louise Bloch, Johannes Rückert, and Christoph Friedrich. 2025. Towards automatic formal feedback on scientific documents. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 334–344, Vienna, Austria. Association for Computational Linguistics.

Stephen Bodnar. 2025. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 63–71, Tallinn, Estonia. University of Tartu Library.

Eujene Nikka V. Boquio and Prospero C. Naval, Jr. 2024. Beyond canonical fine-tuning: Leveraging hybrid multi-layer pooled representations of BERT for automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2285–2295, Torino, Italia. ELRA and ICCL.

Allison Bradford, Kenneth Steimel, Brian Riordan, and Marcia Linn. 2024. Building robust content scoring models for student explanations of social justice science issues. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 450–458, Mexico City, Mexico. Association for Computational Linguistics.

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.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of USMLE items. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 522–527, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology. In *Proceedings of the Empowering Education with LLMs – the Next-Gen Interface and Content Generation Workshop at AIED*.

Yayu Cao, Tianxiang Wang, Lvxiaowei Xu, Zhenyao Wang, and Ming Cai. 2025. CxGGEC: Construction-guided grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6156, Vienna, Austria. Association for Computational Linguistics.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.

Abhirup Chakravarty, Mark Brenchley, Trevor Breakspear, Ian Lewin, and Yan Huang. 2025. Enhancing marker scoring accuracy through ordinal confidence modelling in educational assessments. In *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 6: Industry Track), pages 1498–1507, Vienna, Austria. 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.

Po-Kai Chen, Bo-Wei Tsai, Shao Kuan Wei, Chien-Yao Wang, Jia-Ching Wang, and Yi-Ting Huang. 2025. Mixture of ordered scoring experts for cross-prompt essay trait scoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18071–18084, Vienna, Austria. Association for Computational Linguistics.

Ruishi Chen and Yiling Zhao. 2025. EduCSW: Building a Mandarin-English code-switched generation pipeline for computer science learning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 908–919, Vienna, Austria. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12775–12786, Torino, Italia. ELRA and ICCL.

Mihail Chifligarov, Jammila Laâguidi, Max Schellenberg, Alexander Dill, Anna Timukova, Anastasia Drackert, and Ronja Laarmann-Quante. 2025. Automated scoring of a German written elicited imitation test. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 237–247, Vienna, Austria. Association for Computational Linguistics.

Aymeric de Chillaz, Anna Sotnikova, Patrick Jermann, and Antoine Bosselut. 2025. Challenges for AI in multimodal STEM assessments: a human-AI comparison. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 279–293, Vienna, Austria. Association for Computational Linguistics.

Madalina Chitez, Liviu Dinu, Marius Micluta-Campeanu, Ana-Maria Bucur, and Roxana Rogobete. 2025. Assessing critical thinking components in Romanian secondary school textbooks: A data mining approach to the ROTEX corpus. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 780–793, Vienna, Austria. Association for Computational Linguistics.

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025a. Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5811–5829, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhendong Chu, Jian Xie, Shen Wang, Zichao Wang, and Qingsong Wen. 2025b. UniEDU: Toward unified and efficient large multimodal models for educational tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1007–1016, Suzhou (China). Association for Computational Linguistics.

Sofía Correa Busquets, Valentina Córdova Véliz, and Jorge Baier. 2025. IALab UC at BEA 2025 shared task: LLM-powered expert pedagogical feature extraction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1187–1193, Vienna, Austria. Association for Computational Linguistics.

Sasha Costanza-Chock. 2020. *Design justice*. The MIT Press.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at mlsp 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Scott Crossley, Perpetual Baffour, Mihai Dascalu, and Stefan Ruseti. 2024. A world CLASSE student summary corpus. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 99–107, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Dascalescu, Marius Dumitran, and Mihai Alexandru Vasiluta. 2025. Leveraging generative AI for enhancing automated assessment in programming education contests. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 89–99, Vienna, Austria. 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.

Kordula De Kuthy, Leander Girrbach, and Detmar Meurers. 2025. Automatic concept extraction for learning domain modeling: A weakly supervised approach using contextualized word embeddings. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 175–185, Vienna, Austria. Association for Computational Linguistics.

Michiel De Vrindt, Renske Bouwer, Wim Van Den Noortgate, Marije Lesterhuis, and Anaïs Tack. 2025. Explaining holistic essay scores in comparative judgment assessments by predicting scores on rubrics. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 535–548, Vienna, Austria. Association for Computational Linguistics.

Michiel De Vrindt, Anaïs Tack, Renske Bouwer, Wim Van Den Noortgate, and Marije Lesterhuis. 2024. Predicting initial essay quality scores to increase the efficiency of comparative judgment assessments. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 125–136, Mexico City, Mexico. Association for Computational Linguistics.

Jasper Degraeuwe. 2025. You shall know a word's difficulty by the family it keeps: Word family features in personalised word difficulty classifiers for L2 Spanish. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 312–325, Vienna, Austria. Association for Computational Linguistics.

Jasper Degraeuwe and Patrick Goethals. 2024. Leading by example: The use of generative artificial intelligence to create pedagogically suitable example sentences. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 33–48, Rennes, France. LiU Electronic Press.

Yuning Ding, Julian Lohmann, Nils-Jonathan Schaller, Thorben Jansen, and Andrea Horbach. 2024. Transfer learning of argument mining in student essays. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 439–449, Mexico City, Mexico. Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

Shayan Doroudi. 2023. The intertwined histories of artificial intelligence and education. *International Journal of Artificial Intelligence in Education*, 33:885–928.

Marius Dumitran, Mihnea Buca, and Theodor Moroianu. 2025. MateInfoUB: A real-world benchmark for testing LLMs in competitive, multilingual, and multimodal educational tasks. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 24–37, Vienna, Austria. Association for Computational Linguistics.

Matthew Durward and Christopher Thomson. 2024. Evaluating vocabulary usage in LLMs. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 266–282, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Dutilleul, Mathis Debaillon, and Sandeep Mathias. 2024. ISEP_Presidency_University at MLSP 2024 shared task: Using GPT-3.5 to generate substitutes for lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 605–609, Mexico City, Mexico. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2025. Lessons learned in assessing student reflections with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 672–686, Vienna, Austria. Association for Computational Linguistics.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. TRATES: Trait-specific rubric-assisted cross-prompt essay scoring. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20528–20543, Vienna, Austria. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Annie Foret, Erwan Hupel, and Pêr Morvan. 2024. Enhancing a multi-faceted verb-centered resource to help a language learner: the case of breton. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 59–66, Rennes, France. LiU Electronic Press.

Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt, and Gerardo Castañeda-Garza. 2025. Advances in auto-grading with large language models: A cross-disciplinary survey. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 477–498, Vienna, Austria. Association for Computational Linguistics.

Yao Fu and Zhenjie Weng. 2024. Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence*, 7:100306.

Martina Galletti and Valeria Cesaroni. 2025. From end-users to co-designers: Lessons from teachers. In

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 505–516, Vienna, Austria. Association for Computational Linguistics.

Natalie Garrett, Nathan Beard, and Casey Fiesler. 2020. More than "if time allows": The role of ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, page 272–278. Association for Computing Machinery.

Tianyi Geng and David Alfter. 2025. Towards a real-time Swedish speech analyzer for language learning games: A hybrid AI approach to language assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 186–201, Vienna, Austria. Association for Computational Linguistics.

Dominik Glandorf and Detmar Meurers. 2024. Towards fine-grained pedagogical control over English grammar complexity in educational text generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.

Ben Gomes. 2026. Learners and educators are AI's new "super users". *Google Blog: The Keyword*. [Accessed 31-03-2026].

Dhiman Goswami, Kai North, and Marcos Zampieri. 2024. GMU at MLSP 2024: Multilingual lexical simplification with transformer models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 627–634, Mexico City, Mexico. Association for Computational Linguistics.

Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025a. gec-metrics: A unified library for grammatical error correction evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–534, Vienna, Austria. Association for Computational Linguistics.

Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025b. Reliability crisis of reference-free metrics for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24913–24926, Suzhou, China. Association for Computational Linguistics.

Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025c. Rethinking evaluation metrics for grammatical error correction: Why use a different evaluation process than human? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1165–1172, Vienna, Austria. Association for Computational Linguistics.

Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. Are large language models for education reliable across languages? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 612–631, Vienna, Austria. Association for Computational Linguistics.

Abigail Gurin Schleifer, Beata Beigman Klebanov, Moriah Ariely, and Giora Alexandron. 2024. Anna karenina strikes again: Pre-trained LLM embeddings may favor high-performing learners. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 391–402, Mexico City, Mexico. Association for Computational Linguistics.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring. `https://kaggle.com/competitions/asap-aes`. Kaggle.

Junzhi Han and Jinho D. Choi. 2025. Beyond linear digital reading: An LLM-powered concept mapping approach for reducing cognitive load. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 805–817, Vienna, Austria. Association for Computational Linguistics.

Luke Harding. 2025. Utopian and dystopian visions: Steering a course for the responsible use of artificial intelligence (ai) in language testing and assessment. *Language Testing*, 42(4):561–575.

Ahatsham Hayat, Bilal Khan, and Mohammad Hasan. 2024. Improving transfer learning for early forecasting of academic performance by contextualizing language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 137–148, Mexico City, Mexico. Association for Computational Linguistics.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Angga Hidayat and Pipit Firmanti. 2024. Navigating the tech frontier: a systematic review of technology integration in mathematics education. *Cogent Education*, 11(1):2373559.

Nils Hjortnaes, Daniel Dakota, Sandra Kübler, and Francis Tyers. 2024. Evaluating automatic pronunciation scoring with crowd-sourced speech corpus annotations. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 67–77, Rennes, France. LiU Electronic Press.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Shum, Olga C. Santos, Mercedes Rodrigo, Mutlu Cukurova, Ig Bittencourt, and Kenneth Koedinger. 2022. Ethics of

AI in Education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32:504–526.

Anna Hülsing and Andrea Horbach. 2024. Opinions are buildings: Metaphors in secondary education foreign language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 78–95, Rennes, France. LiU Electronic Press.

Leo Huovinen and Mika Hämäläinen. 2025. LLM-assisted, iterative curriculum writing: A human-centered AI approach in Finnish higher education. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1002–1010, Vienna, Austria. Association for Computational Linguistics.

Fareya Ikram, Alexander Scarlatos, and Andrew Lan. 2025. Exploring LLMs for predicting tutor strategy and student outcomes in dialogues. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 765–779, Vienna, Austria. Association for Computational Linguistics.

Michael Ilagan, Beata Beigman Klebanov, and Jamie Mikeska. 2024. Automated evaluation of teacher encouragement of student-to-student interactions in a simulated classroom discussion. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 182–198, Mexico City, Mexico. Association for Computational Linguistics.

Zhuoxuan Jiang, Tianyang Zhang, Peiyan Peng, Jing Chen, Yinong Xun, Haotian Zhang, Lichi Li, Yong Li, and Shaohua Zhang. 2025. Towards generating controllable and solvable geometry problem by leveraging symbolic deduction engine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1378–1398, Vienna, Austria. Association for Computational Linguistics.

Ahmed Karim, Qiao Wang, and Zheng Yuan. 2025. Beyond the score: Uncertainty-calibrated LLMs for automated essay assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19631–19636, Suzhou, China. Association for Computational Linguistics.

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 594–611, Vienna, Austria. Association for Computational Linguistics.

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.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024a. Complex word identification: A comparative study between ChatGPT and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024b. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 96–114, Rennes, France. LiU Electronic Press.

Deokgi Kim, Joonyoung Jo, Byung-Won On, and Ingyu Lee. 2025a. Representation-to-creativity (R2C): Automated holistic scoring model for essay creativity. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5272–5290, Albuquerque, New Mexico. Association for Computational Linguistics.

Euigyum Kim, Seewoo Li, Salah Khalil, and Hyo Jeong Shin. 2025b. STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 920–930, Vienna, Austria. Association for Computational Linguistics.

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, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Zahra Kolagar, Frank Zalkow, and Alessandra Zarcone. 2025. Investigating methods for mapping learning objectives to bloom's revised taxonomy in course descriptions for higher education. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 415–445, Vienna, Austria. Association for Computational Linguistics.

Seonmin Koo, Jinsung Kim, Chanjun Park, and Heuiseok Lim. 2024. Search if you don't know! knowledge-augmented Korean grammatical error correction with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 96–125, Miami, Florida, USA. Association for Computational Linguistics.

Georgi Kostov. 2026. The role of the parents in modern education: Educational policy and interaction with the family environment. *International Journal of Didactical Studies*, 7.

Charles Koutcheme, Nicola Dainese, and Arto Hellas. 2024. Using program repair as a proxy for language models' feedback ability in programming education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 165–181, Mexico City, Mexico. Association for Computational Linguistics.

Charles Koutcheme, Nicola Dainese, and Arto Hellas. 2025. Direct repair optimization: Training small language models for educational program repair improves feedback. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 564–581, Vienna, Austria. Association for Computational Linguistics.

Aomi Koyama, Masato Mita, Su-Youn Yoon, Yasufumi Takama, and Mamoru Komachi. 2025. Targeted syntactic evaluation for grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21108–21125, Vienna, Austria. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Andrei Kucharavy, Cyril Vallez, and Dimitri Percia David. 2025. LLMs protégés: Tutoring LLMs with knowledge gaps improves student learning outcome. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 248–257, Vienna, Austria. Association for Computational Linguistics.

Aayush Kucheria, Nitin Sawhney, and Arto Hellas. 2025. Comparing behavioral patterns of LLM and human tutors: A population-level analysis with the CIMA dataset. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 873–881, Vienna, Austria. Association for Computational Linguistics.

Alexander Kwako and Christopher Ormerod. 2024. Can language models guess your identity? analyzing demographic biases in AI essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 78–86, Mexico City, Mexico. Association for Computational Linguistics.

Mihwa Lee, Björn Rudzewitz, and Xiaobin Chen. 2024a. Developing a pedagogically oriented interactive reading tool with teachers in the loops. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 115–125, Rennes, France. LiU Electronic Press.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024b. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Bernardo Leite and Henrique Lopes Cardoso. 2025. Advancing question generation with joint narrative and difficulty control. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 647–659, Vienna, Austria. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. 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. 2025. Graph-based multi-trait essay scoring. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33325–33351, Suzhou, China. Association for Computational Linguistics.

Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. 2024. Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 365–380, Mexico City, Mexico. Association for Computational Linguistics.

Wei Li, Wen Luo, Guangyue Peng, and Houfeng Wang. 2025. Explanation based in-context demonstrations retrieval for multilingual grammatical error correction. 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 4881–4897, Albuquerque, New Mexico. Association for Computational Linguistics.

Ho Hung Lim and John Lee. 2024. Improving readability assessment with ordinal log-loss. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 343–350, Mexico City, Mexico. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Dion Hoe-Lian Goh, and Nancy Chen. 2025. COGENT: A curriculum-oriented framework for generating grade-appropriate

educational content. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 129–143, Vienna, Austria. Association for Computational Linguistics.

Sarah Löber, Björn Rudzewitz, Daniela Verratti Souto, Luisa Ribeiro-Flucht, and Xiaobin Chen. 2024. Developing a web-based intelligent language assessment platform powered by natural language processing technologies. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 126–136, Rennes, France. LiU Electronic Press.

Agnes Luhtaru, Elizaveta Korotkova, and Mark Fishel. 2024. No error left behind: Multilingual grammatical error correction with pre-trained translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1222, St. Julian's, Malta. Association for Computational Linguistics.

Wanjing (Anya) Ma, Michael Flor, and Zuowei Wang. 2025. Automatic generation of inference making questions for reading comprehension assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 398–414, Vienna, Austria. Association for Computational Linguistics.

Zhenjiang Mao, Artem Bisliouk, Rohith Nama, and Ivan Ruchkin. 2025. Temporalizing confidence: Evaluation of chain-of-thought reasoning with signal temporal logic. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 882–890, Vienna, Austria. Association for Computational Linguistics.

Jacek Marciniak, Marek Kubis, Michał Gulczyński, Adam Szpilkowski, Adam Wieczarek, and Marcin Szczepański. 2025. Improving AI assistants embedded in short e-learning courses with limited textual content. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 794–804, Vienna, Austria. Association for Computational Linguistics.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. Can LLMs effectively simulate human learners? teachers' insights from tutoring LLM students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117, Vienna, Austria. Association for Computational Linguistics.

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.

Noah-Manuel Michael and Andrea Horbach. 2025. GermDetect: Verb placement error detection datasets for learners of Germanic languages. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 818–829, Vienna, Austria. Association for Computational Linguistics.

Marvin Minsky. 1974. A framework for representing knowledge. *MIT Artificial Intelligence Laboratory Memo*, 306.

Adriana Mirabella and Dominique Brunato. 2025. Exploring LLM-based assessment of Italian middle school writing: A pilot study. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 708–715, Vienna, Austria. 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.

Rina Miyata, Toru Urakawa, Hideaki Tamori, and Tomoyuki Kajiwara. 2025. Unsupervised sentence readability estimation based on parallel corpora for text simplification. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 499–504, Vienna, Austria. Association for Computational Linguistics.

Phoebe Mulcaire and Nitin Madnani. 2025. Span labeling with large language models: Shell vs. meat. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 850–859, Vienna, Austria. Association for Computational Linguistics.

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, and Elena Volodina. 2024a. Jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 137–152, Rennes, France. LiU Electronic Press.

Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024b. Harnessing GPT to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 414–427, Mexico City, Mexico. Association for Computational Linguistics.

Karthika N J, Krishnakant Bhatt, Ganesh Ramakrishnan, and Preethi Jyothi. 2025. LEVOS: Leveraging vocabulary overlap with Sanskrit to generate technical lexicons in Indian languages. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 258–265, Vienna, Austria. Association for Computational Linguistics.

Kamel Nebhi, Amrita Panesar, and Hans Bantilan. 2025. End-to-end automated item generation and scoring for adaptive English writing assessment with large language models. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 968–977, Vienna, Austria. Association for Computational Linguistics.

Allen Newell, J. C. Shaw, and Herbert Simon. 1958. Elements of a theory of human problem solving. *Psychological Review*, 65:151–166.

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.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024. Generating contexts for ESP vocabulary exercises with LLMs. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 153–175, Rennes, France. LiU Electronic Press.

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.

OpenAI. 2026. ChatGPT. `https://chat.openai.com`. Accessed: May 13, 2026.

Robert Östling, Murathan Kurfali, and Andrew Caines. 2025. LLM-based post-editing as reference-free GEC evaluation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 213–224, Vienna, Austria. Association for Computational Linguistics.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Benjamin Paddags, Daniel Hershcovich, and Valkyrie Savage. 2024. Automated sentence generation for a spaced repetition software. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 351–364, Mexico City, Mexico. Association for Computational Linguistics.

Francesca Padovani, Caterina Marchesi, Eleonora Pasqua, Martina Galletti, and Daniele Nardi. 2024. Automatic text simplification: A comparative study in Italian for children with language disorders. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 176–186, Rennes, France. LiU Electronic Press.

Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan, and Shashank Sonkar. 2025a. Large language models for education: Understanding the needs of stakeholders, current capabilities and the path forward. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1–10, Vienna, Austria. Association for Computational Linguistics.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser, and Mrinmaya Sachan. 2025b. Educators' perceptions of large language models as tutors: Comparing human and AI tutors in a blind text-only setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Seymour Papert. 1980. *Mindstorms: Children, Computers, and Powerful Ideas*. Harvester Press.

Nisarg Parikh, Alexander Scarlatos, Nigel Fernandez, Simon Woodhead, and Andrew Lan. 2025. LookAlike: Consistent distractor generation in math MCQs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 294–311, Vienna, Austria. Association for Computational Linguistics.

Geon Park, Jiwoo Song, Gihyeon Choi, Juoh Sun, and Harksoo Kim. 2025. K-NLPers at BEA 2025 shared task: Evaluating the quality of AI tutor responses with GPT-4.1. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1145–1163, Vienna, Austria. Association for Computational Linguistics.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez, and Lev Berezhnoy. 2024. A conversational intelligent tutoring system for improving English proficiency of non-native speakers via debriefing of online meeting transcriptions. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 187–198, Rennes, France. LiU Electronic Press.

Kseniia Petukhova and Ekaterina Kochmar. 2025. Intent matters: Enhancing AI tutoring with fine-grained pedagogical intent annotation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 860–872, Vienna, Austria. Association for Computational Linguistics.

Yin Poon, Qiong Wang, John S. Y. Lee, Yu Yan Lam, and Samuel Kai Wah Chu. 2025. PIRLS category-specific question generation for reading comprehension. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 72–80, Tallinn, Estonia. University of Tartu Library.

Mengyang Qiu, Tran Minh Nguyen, Zihao Huang, Zelong Li, Yang Gu, Qingyu Gao, Siliang Liu, and Jungyeul Park. 2025. Multilingual grammatical error annotation: Combining language-agnostic framework with language-specific flexibility. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 202–212, Vienna, Austria. Association for Computational Linguistics.

Muhammad Reza Qorib, Alham Fikri Aji, and Hwee Tou Ng. 2024. Efficient and interpretable grammatical error correction with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17127–17138, Miami, Florida, USA. Association for Computational Linguistics.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing Arabic automated essay scoring with synthetic data and error injection. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 549–563, Vienna, Austria. Association for Computational Linguistics.

Saed Rezayi, Le An Ha, Yiyun Zhou, Andrew Houriet, Angelo D'Addario, Peter Baldwin, Polina Harik, Ann King, and Victoria Yaneva. 2025. Automated scoring of communication skills in physician-patient interaction: Balancing performance and scalability. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 891–897, Vienna, Austria. Association for Computational Linguistics.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable AI in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of Portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209, Mexico City, Mexico. Association for Computational Linguistics.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2025. A framework for proficiency-aligned grammar practice in LLM-based dialogue systems. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 978–987, Vienna, Austria. Association for Computational Linguistics.

David J. Roaché. 2017. *Intercoder Reliability Techniques: Percent Agreement*. SAGE Publications, Inc, California.

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.

Alla Rozovskaya. 2024. Universal Dependencies for learner Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17112–17119, Torino, Italia. ELRA and ICCL.

Annabella Sakunkoo and Jonathan Sakunkoo. 2025. Name of thrones: How do LLMs rank student names in status hierarchies based on race and gender? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 697–707, Vienna, Austria. Association for Computational Linguistics.

Ignacio Sastre, Leandro Alfonso, Facundo Fleitas, Federico Gil, Andrés Lucas, Tomás Spoturno, Santiago Góngora, Aiala Rosá, and Luis Chiruzzo. 2024. RETUYT-INCO at MLSP 2024: Experiments on language simplification using embeddings, classifiers and large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 618–626, Mexico City, Mexico. Association for Computational Linguistics.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. 2025. Do LLMs give psychometrically plausible responses in educational assessments? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 266–278, Vienna, Austria. Association for Computational Linguistics.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. 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.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221, Mexico City, Mexico. Association for Computational Linguistics.

Nils-Jonathan Schaller, Yuning Ding, Thorben Jansen, and Andrea Horbach. 2025. Don't score too early! evaluating argument mining models on incomplete essays. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 345–355, Vienna, Austria. Association for Computational Linguistics.

Veronica Schmalz and Anaïs Tack. 2025. Can GPTZero's AI vocabulary distinguish between LLM-generated and student-written essays? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 937–952, Vienna, Austria. Association for Computational Linguistics.

Olga Seminck, Yoann Dupont, Mathieu Dehouck, Qi Wang, Noé Durandard, and Margo Novikov. 2025. Lattice @MultiGEC-2025: A spitful multilingual language error correction system using LLaMA. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 34–41, Tallinn, Estonia. University of Tartu Library.

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, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 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.

Mayank Sharma and Jason Zhang. 2025. Decoding actionability: A computational analysis of teacher observation feedback. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 898–907, Vienna, Austria. Association for Computational Linguistics.

Kevin Shi and Karttikeya Mangalam. 2025. UPSC2M: Benchmarking adaptive learning from two million MCQ attempts. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 931–936, Vienna, Austria. Association for Computational Linguistics.

Takumi Shibata and Yuichi Miyamura. 2025. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29988–30001, Suzhou, China. Association for Computational Linguistics.

Mariana Shimabukuro, Deval Panchal, and Christopher Collins. 2025. LangEye: Toward 'anytime' learner-driven vocabulary learning from real-world objects. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 446–459, Vienna, Austria. Association for Computational Linguistics.

Astha Singh, Mark Torrance, and Evgeny Chukharev. 2025. EyeLLM: Using lookback fixations to enhance human-LLM alignment for text completion. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 841–849, Vienna, Austria. Association for Computational Linguistics.

Li Siyan, Teresa Shao, Julia Hirschberg, and Zhou Yu. 2024. Using adaptive empathetic responses for teaching English. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 34–53, Mexico City, Mexico. Association for Computational Linguistics.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Alexey Sorokin and Regina Nasyrova. 2025. LLMs in alliance with edit-based models: advancing in-context learning for grammatical error correction by specific example selection. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 517–534, Vienna, Austria. Association for Computational Linguistics.

KV Aditya Srivatsa, Kaushal Maurya, and Ekaterina Kochmar. 2025a. Can LLMs reliably simulate real students' abilities in mathematics and reading comprehension? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 988–1001, Vienna, Austria. Association for Computational Linguistics.

Kv Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025b. LLMs cannot spot math errors, even when allowed to peek into the solution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages

10914–10928, Suzhou, China. 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. 2024. Synthetic data generation for low-resource grammatical error correction with tagged corruption models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 11–16, Mexico City, Mexico. Association for Computational Linguistics.

Ryszard Staruch. 2025. UAM-CSI at MultiGEC-2025: Parameter-efficient LLM fine-tuning for multilingual grammatical error correction. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 42–49, Tallinn, Estonia. University of Tartu Library.

Ryszard Staruch, Filip Gralinski, and Daniel Dzienisiewicz. 2025. Adapting LLMs for minimal-edit grammatical error correction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 118–128, Vienna, Austria. Association for Computational Linguistics.

Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, and John P. McCrae. 2024. Evaluating the generalisation of an artificial learner. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 199–208, Rennes, France. LiU Electronic Press.

Kevin Stowe, Benny Longwill, Alyssa Francis, Tatsuya Aoyama, Debanjan Ghosh, and Swapna Somasundaran. 2024. Identifying fairness issues in automatically generated testing content. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 232–250, Mexico City, Mexico. Association for Computational Linguistics.

David Strohmaier and Paula Buttery. 2024. Semantic error prediction: Estimating word production complexity. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 209–225, Rennes, France. LiU Electronic Press.

Jiamin Su, Yibo Yan, Fangteng Fu, Zhang Han, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. EssayJudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6363–6389, Vienna, Austria. Association for Computational Linguistics.

Hakyung Sung, Karla Csuros, and Min-Chang Sung. 2025. Comparing human and LLM proofreading in L2 writing: Impact on lexical and syntactic features. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 11–23, Vienna, Austria. Association for Computational Linguistics.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*.

Anaïs Tack. 2024. ITEC at MLSP 2024: Transferring predictions of lexical difficulty from non-native readers. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 635–639, Mexico City, Mexico. Association for Computational Linguistics.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. ITEC at BEA 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, Mexico City, Mexico. Association for Computational Linguistics.

Chenming Tang, Fanyi Qu, and Yunfang Wu. 2024. Ungrammatical-syntax-based in-context example selection for grammatical error correction. 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 1758–1770, Mexico City, Mexico. Association for Computational Linguistics.

Rajneesh Tiwari and Pranshu Rastogi. 2025. Phaedrus at BEA 2025 shared task: Assessment of mathematical tutoring dialogues through tutor identity classification and actionability evaluation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1098–1107, Vienna, Austria. Association for Computational Linguistics.

Guillaume Toussaint, Yannick Parmentier, and Claire Gardent. 2024. GRAMEX: Generating controlled grammar exercises from various sources. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 226–234, Rennes, France. LiU Electronic Press.

Nhat Tran, Diane Litman, Benjamin Pierce, Richard Correnti, and Lindsay Clare Matsumura. 2025. Improving in-context learning example retrieval for classroom discussion assessment with re-ranking and

label ratio regulation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 752–764, Vienna, Austria. Association for Computational Linguistics.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. LLM chatbots as a language practice tool: a user study. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 235–247, Rennes, France. LiU Electronic Press.

Till Überrück-Fries, Agata Savary, and Agnieszka Dryjańska. 2024. Sailing through multiword expression identification with Wiktionary and linguse: A case study of language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 248–262, Rennes, France. LiU Electronic Press.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88, Tallinn, Estonia. University of Tartu Library.

UNESCO. 2026. Global report on teachers: Addressing teacher shortages and transforming the profession.

Felipe Urrutia, Cristian Buc, Roberto Araya, and Valentin Barriere. 2025. Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 38–54, Vienna, Austria. Association for Computational Linguistics.

Martin Vainikko, Taavi Kamarik, Karina Kert, Krista Liin, Silvia Maine, Kais Allkivi, Annekatrin Kaivapalu, and Mark Fishel. 2025. Paragraph-level error correction and explanation generation: Case study for Estonian. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 953–967, Vienna, Austria. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 561–566, Mexico City, Mexico. Association for Computational Linguistics.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Anh-Duc Vu, Jue Hou, Anisia Katinskaia, Ching-Fan Sheu, and Roman Yangarber. 2025. A Bayesian approach to inferring prerequisite structures and topic difficulty in language learning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 737–751, Vienna, Austria. Association for Computational Linguistics.

Deliang Wang, Chao Yang, and Gaowei Chen. 2025a. Wonderland_EDU@HKU at BEA 2025 shared task: Fine-tuning large language models to evaluate the pedagogical ability of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1040–1048, Vienna, Austria. Association for Computational Linguistics.

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025b. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12416–12436, Vienna, Austria. Association for Computational Linguistics.

Junling Wang, Anna Rutkiewicz, April Wang, and Mrinmaya Sachan. 2025c. Generating pedagogically meaningful visuals for math word problems: A new benchmark and analysis of text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11229–11257, Vienna, Austria. Association for Computational Linguistics.

Xiaoman Wang, Dan Yuan, Xin Liu, Yike Zhao, Xiaoxiao Zhang, Xizhi Chen, and Yunshi Lan. 2025d. VisCGEC: Benchmarking the visual Chinese grammatical error correction. 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 5054–5068, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoying Wang, Lingling Mu, Jingyi Zhang, and Hongfei Xu. 2024a. Multi-pass decoding for grammatical error correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9916, Miami, Florida, USA. Association for Computational Linguistics.

Yixuan Wang, Baoxin Wang, Yijun Liu, Dayong Wu, and Wanxiang Che. 2024b. LM-combiner: A contextual rewriting model for Chinese grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10675–10685, Torino, Italia. ELRA and ICCL.

Yupei Wang, Renfen Hu, and Zhe Zhao. 2024c. Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on

linguistically-informed counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8906–8925, Miami, Florida, USA. Association for Computational Linguistics.

World Inequality Lab. 2026. World inequality report 2026. [Accessed 31-03-2026].

Tianxiang Wu, Han Chen, Luozheng Qin, Ziqiang Cao, and Chunhui Ai. 2024. Improving copy-oriented text generation via EDU copy mechanism. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8768–8780, Torino, Italia. ELRA and ICCL.

Yuyang Yan, Hui Liu, and Toby Chau. 2025. A systematic review of AI ethics in education: Challenges, policy gaps, and future directions. *Journal of Global Information Management*, 33(1):1–50.

Kevin P. Yancey, Andrew Runge, Geoffrey LaFlair, and Phoebe Mulcaire. 2024. BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024a. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Yaneva, King Yiu Suen, Le An Ha, Janet Mee, Milton Quranda, and Polina Harik. 2024b. Automated scoring of clinical patient notes: Findings from the Kaggle competition and their translation into practice. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 87–98, Mexico City, Mexico. Association for Computational Linguistics.

Haihui Yang and Xiaojun Quan. 2024. Alirector: Alignment-enhanced Chinese grammatical error corrector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2531–2546, Bangkok, Thailand. Association for Computational Linguistics.

Haiyin Yang, Zoey Liu, and Stefanie Wulff. 2025. Using NLI to identify potential collocation transfer in L2 English. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 687–696, Vienna, Austria. Association for Computational Linguistics.

Sahar Yarmohammadtoosky, Yiyun Zhou, Victoria Yaneva, Peter Baldwin, Saed Rezayi, Brian Clauser, and Polina Harik. 2025. Enhancing security and strengthening defenses in automated short-answer grading systems. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 830–840, Vienna, Austria. Association for Computational Linguistics.

Mazen Yasser, Mariam Saeed, Hossam Elkordi, and Ayman Khalafallah. 2025. Averroes at BEA 2025 shared task: Verifying mistake identification in tutor, student dialogue. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1121–1126, Vienna, Austria. Association for Computational Linguistics.

Jingheng Ye, Zishan Xu, Yinghui Li, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, Wenhao Jiang, Hong-Gee Kim, Ruitong Liu, Xin Su, and Zifei Shan. 2025. CLEME2.0: Towards interpretable evaluation by disentangling edits for grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 204–222, Vienna, Austria. Association for Computational Linguistics.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for rubric-based essay scoring on EFL writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454, Vienna, Austria. Association for Computational Linguistics.

Mehrdad Yousefpoori-Naeim, Shayan Zargari, and Zahra Hatami. 2024. Using machine learning to predict item difficulty and response time in medical tests. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 551–560, Mexico City, Mexico. Association for Computational Linguistics.

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch, and Nico Andersen. 2025. Down the cascades of omethi: Hierarchical automatic scoring in large-scale assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 660–671, Vienna, Austria. Association for Computational Linguistics.

Torsten Zesch, Dominic Gardner, and Marie Bexte. 2025. Transformer-based real-word spelling error feedback with configurable confusion sets. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 375–383, Vienna, Austria. Association for Computational Linguistics.

Yike Zhao, Simin Guo, Ziqing Yang, Shifan Han, Dahua Lin, and Fei Tan. 2025. More data or better data? a critical analysis of data selection and synthesis for mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 618–629,

Suzhou (China). Association for Computational Linguistics.

A A Structured Taxonomy for Ethical and Stakeholder Review of EduNLP Research

Figure 10[12] presents the complete taxonomy developed in this review, offered here as a standalone contribution. The taxonomy is organised around the three research questions and a concluding recommendations dimension. Beyond its role in this review, the taxonomy can be reused for future surveys of EdTech research. It can also serve as a practical self-audit tool: researchers can use it to situate themselves within the EduNLP space, and assess the rigour and inclusivity of their work before submission.

B ACL Anthology Main Conference Search

We retrieve all papers with at least one of the following search terms in the title or abstract. The venues included are: ACL, EACL, NAACL, EMNLP, LREC-COLING, and Findings, for the years 2024 and 2025. The search terms were developed through internal discussion and discussion with other researchers in the EduNLP field. The search terms are as follows:

- "automated essay scoring",
- "automated writing evaluation",
- "short answer grading",
- "automatic short answer grading",
- "open-ended response assessment",
- "automated assessment of spoken responses",
- "spoken response scoring",
- "speech-based assessment",
- "automatic speech scoring",
- "dialogue-based tutoring",
- "spoken dialogue system education",
- "intelligent tutoring systems NLP",
- "student modeling",
- "learner modeling",
- "knowledge tracing",
- "learner cognition modeling",
- "educational data mining NLP",
- "learning analytics text",
- "game-based learning assessment",
- "stealth assessment",
- "peer assessment NLP",
- "peer review automated feedback",
- "automated feedback generation",
- "formative feedback writing",
- "grammatical error correction",
- "grammar error detection",
- "lexical complexity prediction",
- "text simplification for learners",
- "multimodal learning analytics",
- "generative AI in education",
- "mathematic education",
- "math education",
- "math word problems",
- "mathematical reasoning",
- "student error in mathematics",
- "intelligent tutoring system math",
- "knowledge tracing mathematics",
- "misconception detection mathematics".

C Extraction Schema

For extracting the entities relevant to our research questions, we used the following schema:

- **[RQ2]** Author affiliations
- **[RQ1]** Specific task worked on
- **[RQ1]** Datasets used and availability
- **[RQ1]** Explicit motivation for the paper and associated quotes
- **[RQ2]** Stakeholders mentioned (multi-label; the options being *Learner/student*, *Teacher*, *School/university*, *Paper author*, *Other researcher*, *Domain expert*, *Parent*, *Governmental body*, *Inudustry*, *Non-profit* which includes large-standardised testing providers, and *None / N.A.*), and associated quotes
- **[RQ2]** Stakeholders included in the research (multi-label; same list as above), as well as their respective level of inclusion (multi-label with *High*, *Middling* and *Low*) and associated quotes for how they are included
- **[RQ1]** Context in which the system deployed (if any) and relevant quotes
- **[RQ2]** Explicit stakeholder incentives
- **[RQ2]** Implicit stakeholder incentives
- **[RQ3]** Risk, concerns and limitations raised, associated quotes, level of engagement (multi-label with *High*, *Middling* and *Low*) and measures taken to address risk

[12] This figure was created in app.xmind.com.

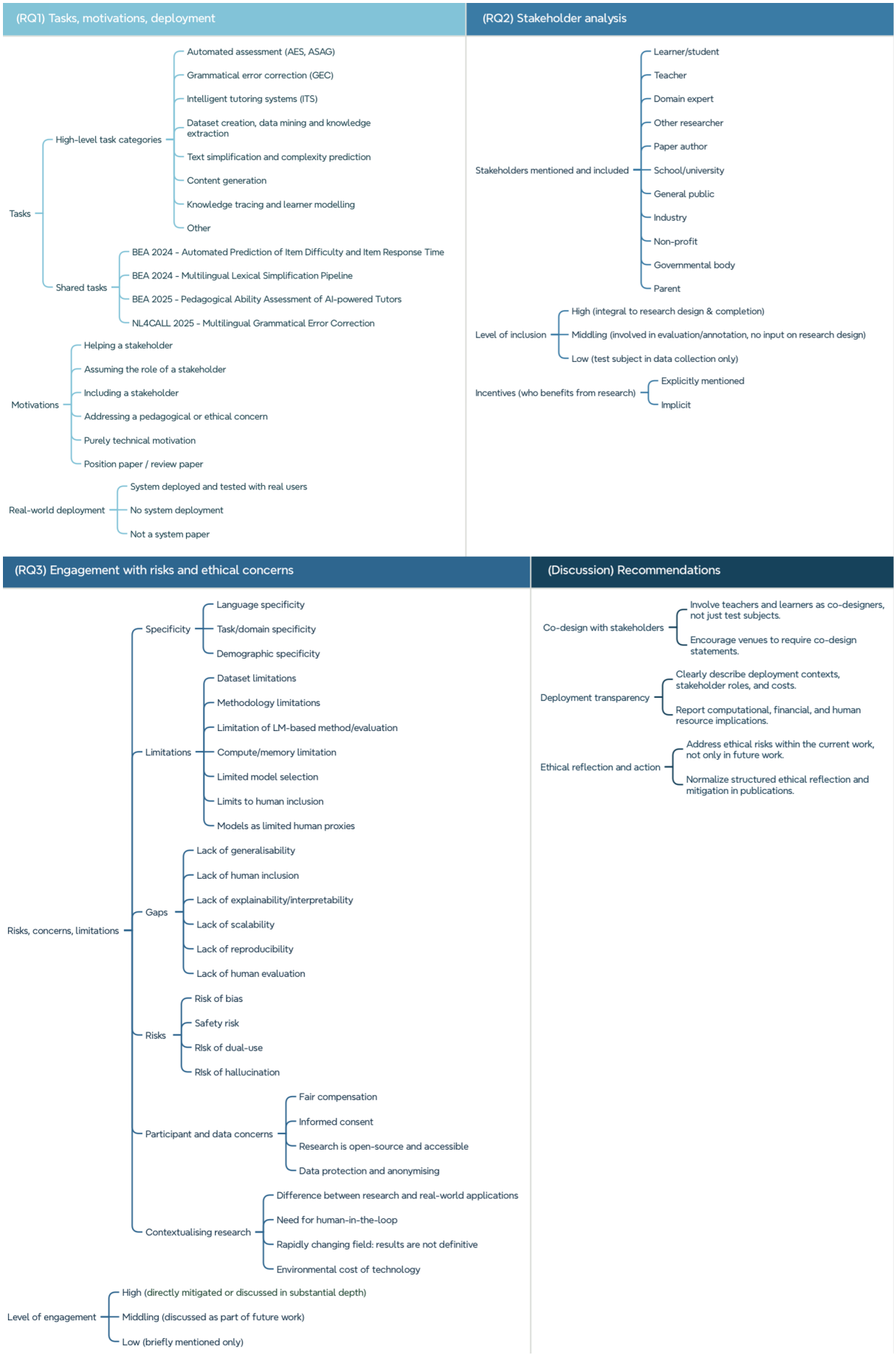

Figure 10: Detailed taxonomy for reviewing EduNLP research.

- **[RQ3]** Future directions/aspirations mentioned and relevant quotes
- **[RQ1]** Entities acknowledged (including funding)

Once phase (3) was completed, the three annotators independently extracted high-level categories and labels from the data:

- **[RQ1]** Mapping specific tasks to high-level tasks (as reported in Figure 2)
- **[RQ2]** Mapping free-text dataset names to unique labels (as reported in Figure 3) and their availability (as reported in Figure 11)
- **[RQ1]** Categorising free-text motivations to high-level labels (as reported in Figure 12) and the mentioned stakeholders (as reported in Figure 4)
- **[RQ1]** Mapping context deployment to high-level labels (as reported in Figure 5)
- **[RQ2]** Extracting standardised stakeholder types from explicit and implicit incentives (as reported in Figure 9)
- **[RQ3]** Mapping risks, concerns and limitations to high-level categories (as reported in Figure 18) and the authors' level of engagement associated to each (as reported in Figure 19)
- **[RQ3]** Mapping future directions and aspirations high-level categories (as reported in Figure 20)

For each dimension, the mapping was made by one of the three annotators independently. This was considered sufficient given that the high-level categories were derived directly from the extracted free-text data rather than applied to raw papers: the iterative reconciliation process in phases (1) and (2) had already established shared interpretive norms among annotators, and the categorisation task at this stage involved consolidating labels that were already grounded in agreed extractions rather than making independent judgements about unseen material.

D Agreement Computation

Table 1 reports agreement for the free-text dimensions. Table 6 includes examples of how percentage agreement (PA) was calculated for the following free-text dimensions: "datasets used", "risks/concerns", "measures taken to address risks/concerns", and "future directions". As these dimensions often included many different elements (e.g., dataset names for the "datasets used" dimension and specific suggestions for future research for "future directions"), PA was calculated in the form of pairwise agreements (at the level of the paper, with the score reported in Table 1 corre-

Dimension	Agreement
Task	0.95
Datasets used*	0.87
Dataset availability	0.93
Methods	0.93
Evaluation	0.92
Motivation	0.96
Deployment	1.0
Explicit incentives	0.72
Implicit incentives	0.53
Risks/concerns*	0.57
Measures taken to address risks/concerns*	0.91
Future directions*	0.69
Future deployment	0.92

Table 1: Agreement for free-text annotation dimensions. For dimensions with an asterisk, we computed per-paper percentage agreement. For the other dimensions, we computed majority percentages (i.e. 0 if none of the annotators agree, 0.67 if 2/3 annotators agree, and 1 if 3/3 annotators agree). The values reported correspond to the averages across all 25 papers in the shared batch. Computation examples are included in Table 6 and Table 7.

Label	PA	α
Overall	0.91	0.49
Domain experts	0.84	0.64
General user	0.97	0.74
Industry/company	0.97	0.49
Learners	0.92	-0.03
None / N.A.	0.95	-0.01
Other Researchers	0.79	0.44
Paper authors	0.92	0.46
Policy makers/governments/ministries	0.95	0.31
Schools/universities	0.79	0.27
Special needs/disability user	0.97	0.0
Teachers	0.89	0.79

Table 2: Average percentage agreement (PA) and inter-annotator agreement (Krippendorff's α) for "stakeholders mentioned".

Label	PA	α
Overall	0.94	0.7
Domain experts	0.89	0.61
General user	1.0	1.0
Industry/company	0.97	0.49
Learners	0.92	0.71
None / N.A.	0.92	0.84
Other Researchers	0.92	0.53
Paper authors	0.92	0.46
Policy makers/governments/ministries	0.97	0.0
Schools/universities	0.97	0.0
Teachers	0.97	0.84

Table 3: Average percentage agreement (PA) and inter-annotator agreement (Krippendorff's α) for "stakeholders included".

Label	PA	α
Overall	0.88	0.61
High	0.84	0.51
Middling	0.84	0.41
Minimal	0.92	0.53
None / N.A.	0.92	0.84

Table 4: Average percentage agreement (PA) and inter-annotator agreement (Krippendorff's alpha) for "level of inclusion stakeholders included".

Label	PA	α
Overall	0.84	0.52
High	0.87	0.66
Middling	0.79	0.56
Minimal	0.79	0.57
None / N.A.	0.92	0.37

Table 5: Average percentage agreement (PA) and inter-annotator agreement (Krippendorff's alpha) for "level of engagement risks/concerns".

sponding to the mean of these per-paper pairwise agreement values).

Table 7 contains examples of how PA was calculated for the following free-text dimensions: "task", "dataset availability", "methods", "evaluation", "motivation", "deployment", "explicit incentives", "implicit incentives", and "future deployment". As these dimensions virtually always included only a very limited number of core elements, PA was calculated in the form of majority percentages (at the level of the paper, with the score reported in Table 1 corresponding to the mean of these per-paper majority percentage values).

Tables 2 to 5 further present the agreement for the multilabel dimensions (stakeholders mentioned, stakeholders included, level of inclusion stakeholders included, and level of engagement risks/concerns).

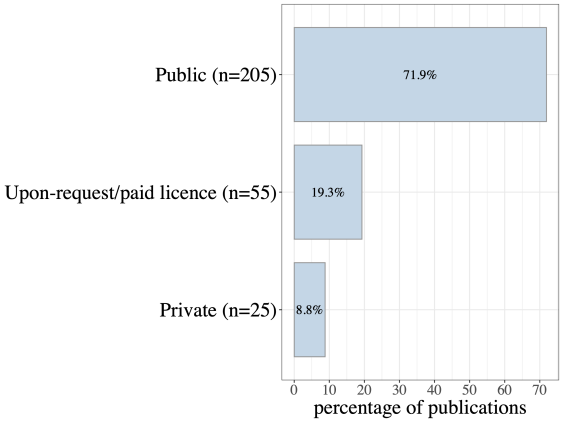

Figure 11: Availability of datasets used in the surveyed papers.

Figure 12: Distribution of papers' explicit motivations for the task; papers may belong to more than one category.

E Mapping Papers to High-Level Tasks

The mapping of the analysed papers to the eight high-level tasks is presented in Table 8.

F Availability of Datasets

The availability of the datasets used in the analysed papers is presented in Figure 11.

G Explicit Motivations

The distribution of the papers' explicit motivations for the research conducted is visualised in Figure 12.

Paper ID	Annotator 1	Annotator 2	Annotator 3	Computation
Datasets				
2025.bea-1.38	LORuGEC RULEC-GEC RU-LAng8 GERA	LORuGEC RULEC-GEC RU-LAng8 GERA	LORuGEC RULEC-GEC RU-LAng8 GERA English_BEA	• LORuGEC: 3/3 PWA • RULEC-GEC: 3/3 PWA • RU-LAng8: 3/3 PWA • GERA: 3/3 PWA • English_BEA: 1/3 PWA → **PA paper = 13/15 = 86.87%**
2025.bea-1.72	custom_dataset EKI-L2 Estonian National Corpus EstGEC-L2	custom_dataset EKI-L2 EstGEC-L2	EKI-L2	• custom_dataset: 1/3 PWA • EKI-L2: 3/3 PWA • Estonian National Corpus: 1/3 PWA • EstGEC-L2: 1/3 PWA → **PA paper = 6/12 = 50%**
Risks/concerns				
2025.acl-long.1026	Limited dataset size; Error Operation Ratio Limitation; limited language coverage; fair pay of annotators	Limited to English; ensured fair compensation for the annotators; limited number of minimal pairs due to manual creation	Manual annotation; dataset size; ratio of error operations is not controlled for; limited focus (on English)	• (Remunerate) annotators: 3/3 PWA • (Limited) language/resource coverage: 3/3 PWA • Error operation ratio: 1/3 PWA → **PA paper = 7/9 = 77.78%**

Table 6: Examples of percentage agreement (PA) computation for free-text annotation dimensions. "PWA" stands for pairwise agreement among the annotators (i.e. Annotator 1 compared to 2, Annotator 1 compared to 3, and Annotator 2 compared to 3). Note that *absence* of annotation also counts as agreement (e.g., between Annotator 1 and 2 for the "English_BEA" dataset).

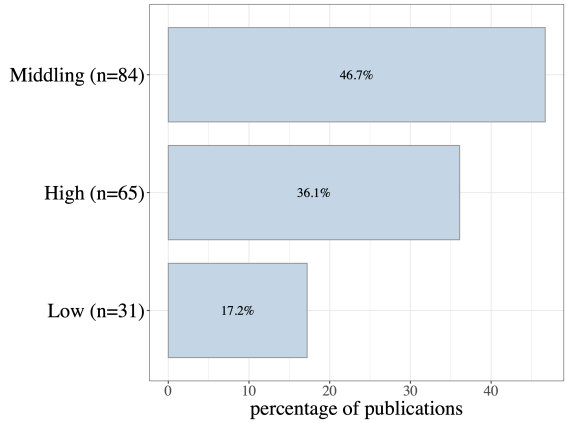

Figure 13: Overall level of inclusion of included stakeholders; we distinguish 3 levels: *High* (integral to research design & completion), *Middling* (involved in data evaluation or annotation, but have no input on research design), and *Low* (test subjects in data collection only).

H Stakeholders Included and Mentioned

As the authors are the primary stakeholders of the research, we present demographic information in terms of country of author affiliation in Figure 15. Secondly, Figure 13 visualises the level of inclusion of the stakeholders included in the research.

I Acknowledged Entities

Figure 14 depicts the "acknowledged countries" (i.e., the the number of papers per country of affiliation of entities acknowledged), while Figure 16 provides more details on the number of times entities were acknowledged.

J Relation between Tasks and Implicitly Benefitting Stakeholders

Figure 17 presents a heat-map that links the high-level tasks to the combined explicitly and implicitly benefitting stakeholders.

K Risks, Concerns and Limitations Breakdown

Figure 19 distinguishes three levels of engagement with stated risks: High (directly mitigated or discussed in substantial depth), Middling (discussed as part of future work), and Low (briefly mentioned only). Across most risk categories, the majority of engagement is at a Low or Middling level. Methodology limitations show 90% Middling engagement – they are widely acknowledged but rarely addressed in the current work. Dataset limitations are 56%

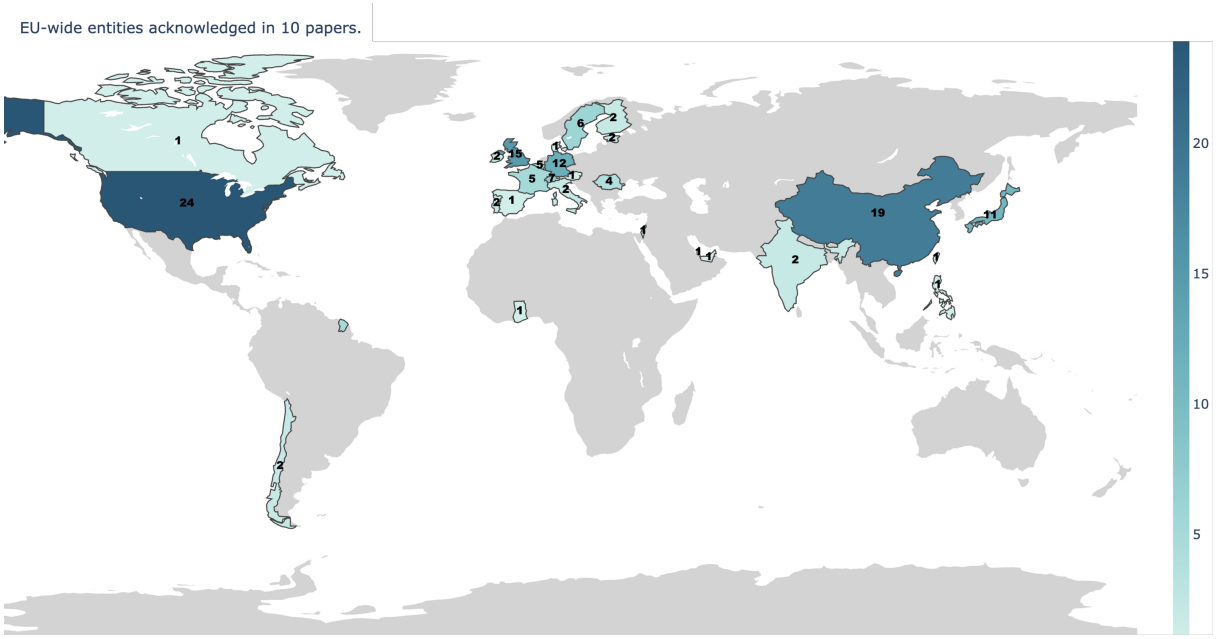

Figure 14: Acknowledged countries (i.e., the number of papers per country of affiliation of entities acknowledged in the surveyed papers).

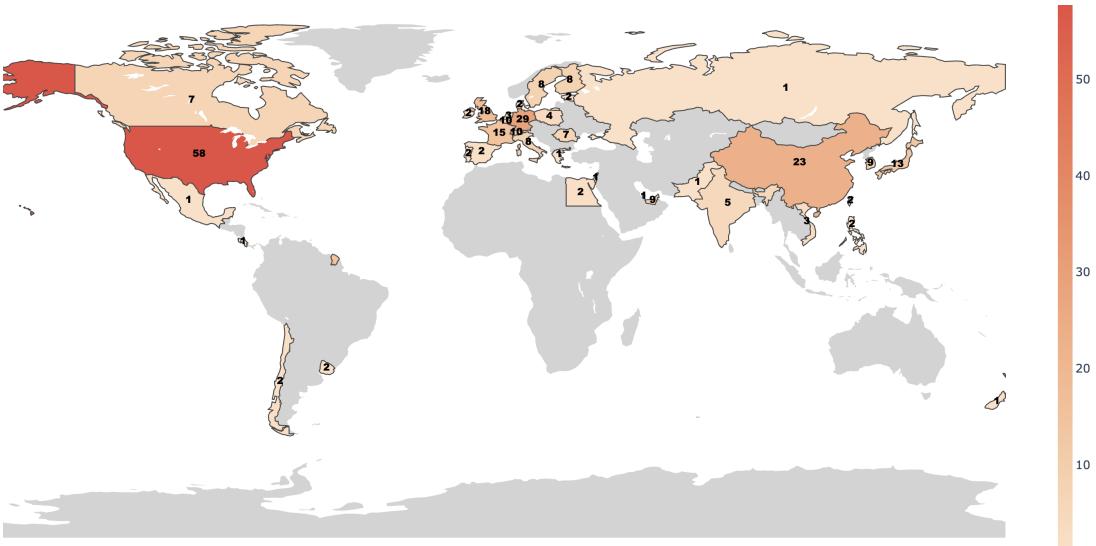

Figure 15: Author countries (i.e., the number of papers per country of author affiliation).

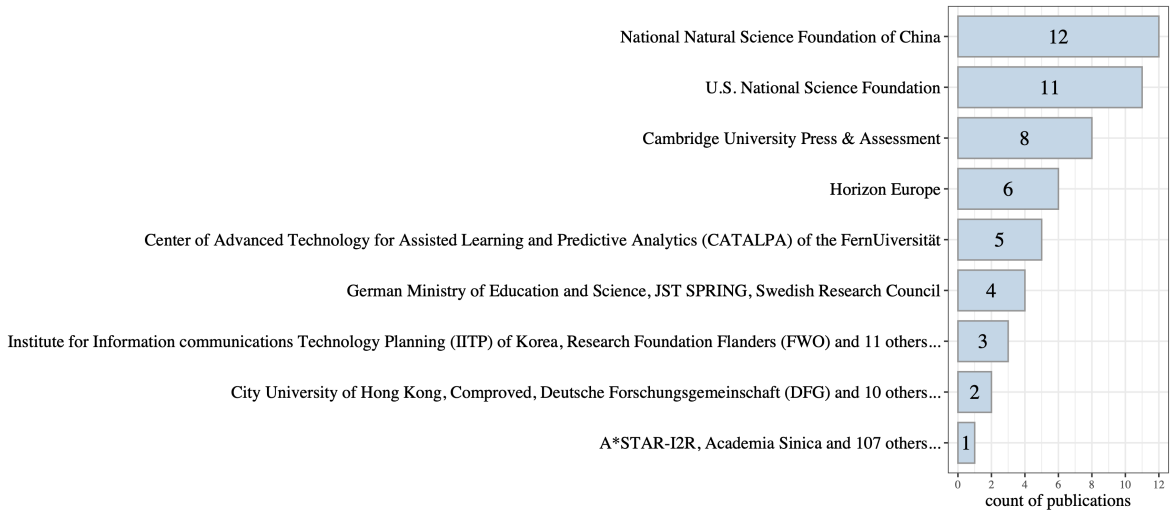

Figure 16: Acknowledged entities (i.e., the number of times an entity was acknowledged in the surveyed papers).

Figure 17: Heat-map relating the high-level tasks to the combined explicitly and implicitly benefitting stakeholders.

Paper ID	Annotator 1	Annotator 2	Annotator 3	Required elements	Computation
Task					
2024.nlp4call-1.14	Conversational intelligent tutoring system (ITS) for EFL speakers	Intelligent tutoring systems	Conversational intelligent tutoring system (ITS) for L2 English	• ITS	ITS mentioned in all three annotations → **agreement = 100%**
2025.bea-1.2	LLM vs. human proofreading of L2 writing	Proofreading in L2 writing (automated error correction)	Analysis of lexical and syntactic interventions of human and LLM proofreading aimed at improving intelligibility in identical L2 writings	• Human vs. LLM proofreading • L2 writing	Element of "human vs. LLM" not mentioned by Annotator 2 → **agreement = 66.67%**
Methods					
2024.bea-1.58	Compare zero-short and few-shot prompting with LLMs vs. fine-tuned models for assessing short answers.	Zero-shot and few-shot prompting (with some fine-tuning) for automated scoring	Use of LLMs (GPT and LLaMA) for automated scoring of short answer responses	• Comparison between zero-shot and few-shot • Use of fine-tuned models	Element of "zero-shot vs. few-shot" not mentioned by Annotator 3 → **agreement = 66.67%**
Future deployment					
2025.emnlp-main.992	In teacher-in-the-loop real world applications	In high-stakes assessment scoring	AES systems for English	• Teacher in the loop • High-stakes assessments • AES systems	None of the annotators fully overlap → **agreement = 0%**

Table 7: Examples of majority agreement computation for free-text annotation dimensions. Possible values: 0% (none of the annotations fully overlap), 66.67% (full overlap for two of the three annotations), or 100% (full overlap for all three annotations).

Middling. Risk of bias, one of the most commonly cited concerns, is engaged at a High level in only 17% of papers that raise it; in 45% of cases it is Middling, and in 38% it is Low.

High engagement is most consistently found in a small set of categories: fair compensation for stakeholders (100% High, though only 6 papers raise this at all), and to a lesser extent data protection (65% High). Risk of hallucination, lack of human evaluation, and the gap between research and real-world application are predominantly engaged at Low or Middling levels – noted as future work, but rarely designed around. This pattern suggests a community that is aware of the ethical dimensions of its work but has not yet developed consistent norms for acting on them within the scope of individual papers.

L Areas of Future Work

Figure 20 shows the areas of future work explicitly mentioned in the papers, split across five high-level categories ("stakeholder inclusion", "technical development", "expand scope", "engage with issues", and "none/not applicable).

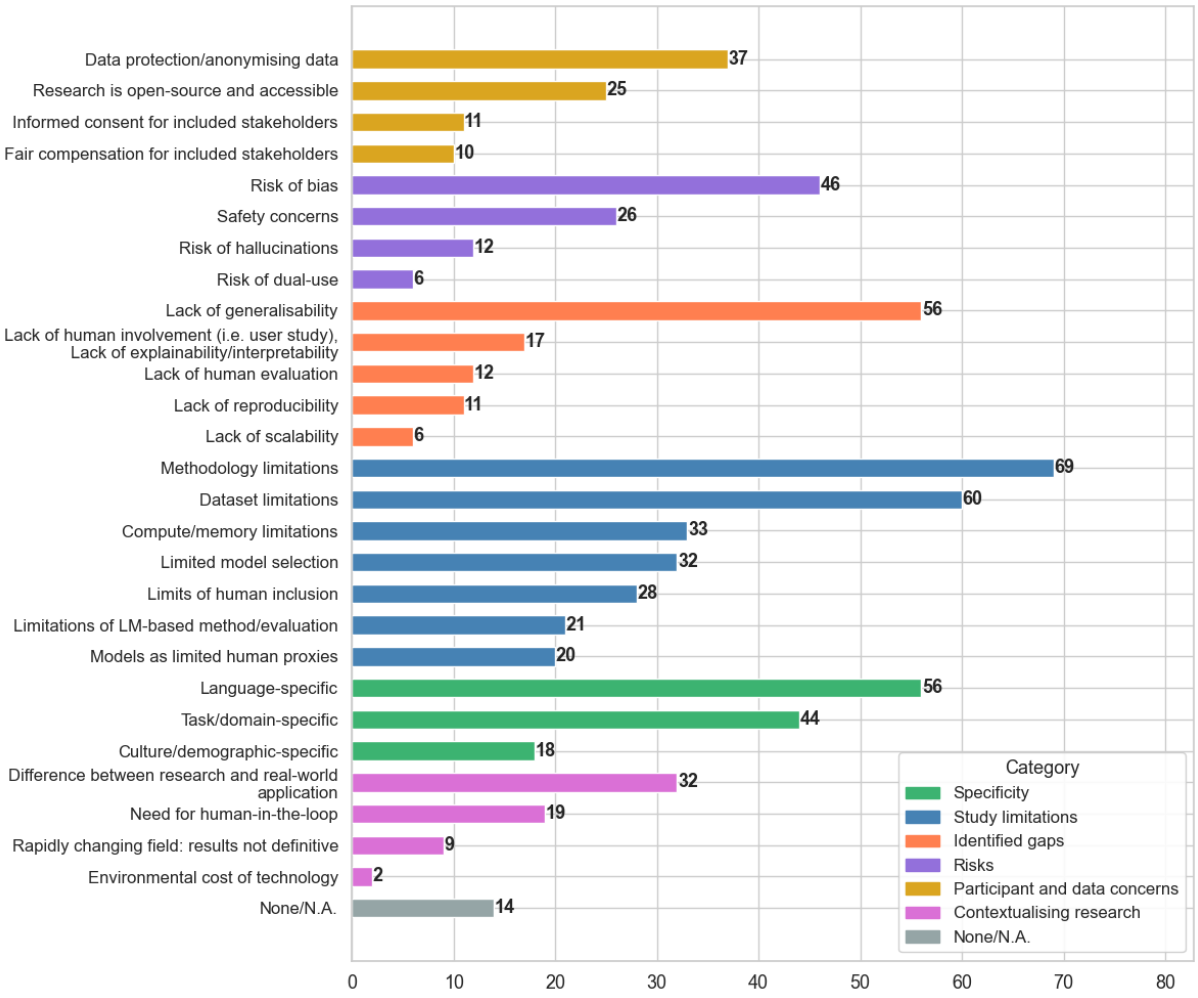

Figure 18: Risks, concerns and limitations explicitly raised by paper authors split across six high-level categories (showing the number of papers; note that a paper may report more than one area of risk).

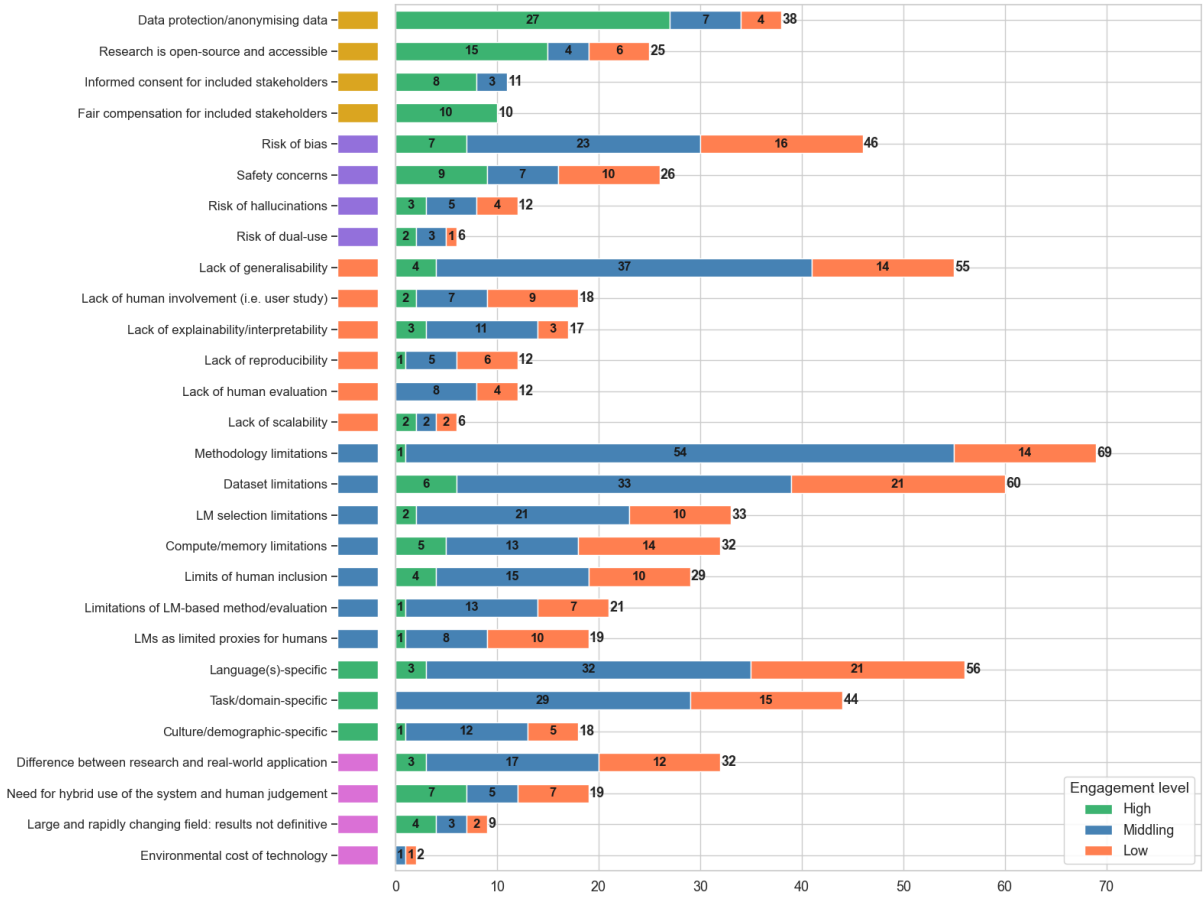

Figure 19: Engagement levels for the risks, concerns and limitations explicitly raised by paper authors; we distinguish 3 levels of risk engagement: *High* (a risk, concern or limitation that is directly mitigated in the paper or discussed in great depth), *Middling* (discussed as part of future work), and *Low* (briefly mentioned only).

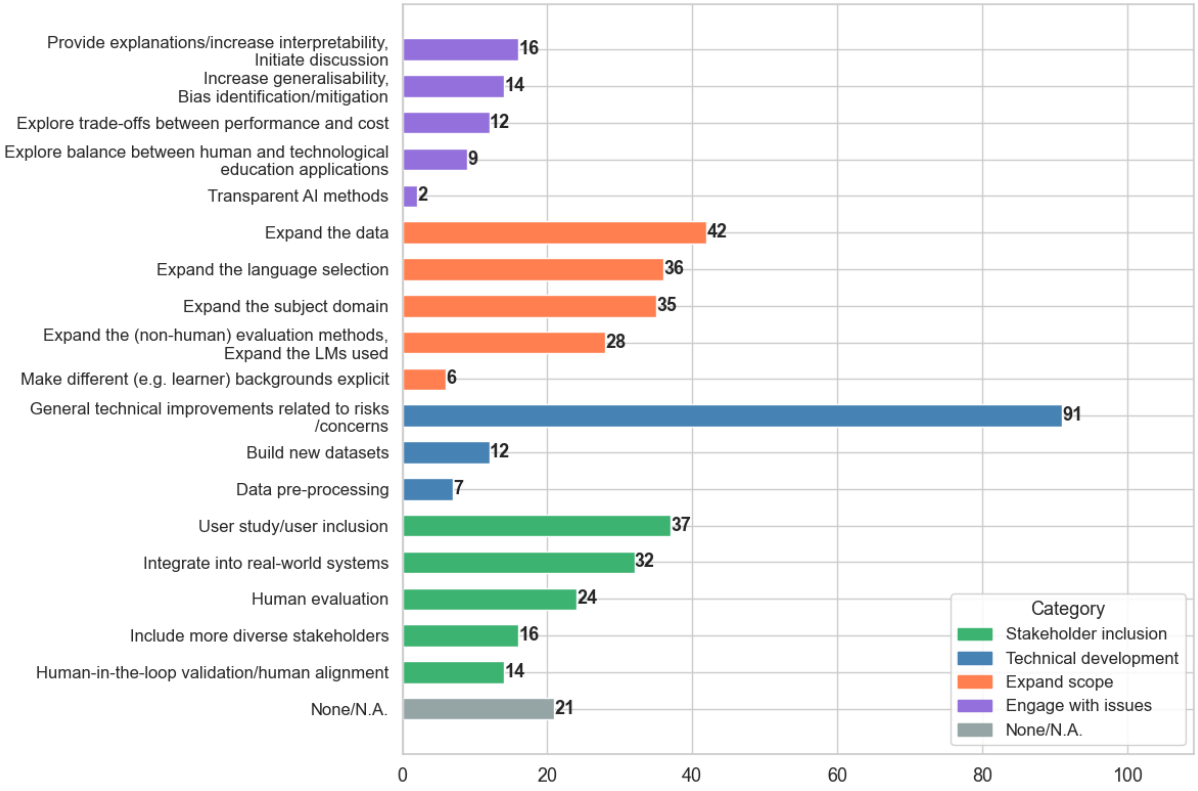

Figure 20: Areas of future work explicitly mentioned in the papers split across four high-level categories (showing the number of papers; note that a paper may report more than one area of future work).

High-level task	Papers
Automated assessment (AES, ASAG)	Yaneva et al. (2024b); Yarmohammadtoosky et al. (2025); Chamieh et al. (2024); De Vrindt et al. (2024); Crossley et al. (2024); Frederick Eneye et al. (2025); Carpenter et al. (2024); Doi et al. (2024); Arronte Alvarez and Xie Fincham (2025); Chifligarov et al. (2025); Rezayi et al. (2025); Yancey et al. (2024); Bradford et al. (2024); Bannò et al. (2024); Kwako and Ormerod (2024); Löber et al. (2024); Nebhi et al. (2025); Qwaider et al. (2025); Hjortnaes et al. (2024); De Vrindt et al. (2025); Bannò et al. (2025); Stahl et al. (2024); Mirabella and Brunato (2025); Asano et al. (2025); Li et al. (2024); Li and Ng (2024); Wang et al. (2024c); Chen and Li (2024); Chen et al. (2025); Shibata and Miyamura (2025); Srivatsa et al. (2025b); Eltanbouly et al. (2025); Kim et al. (2025a); Lee et al. (2024b); He and Li (2024); Boquio and Naval (2024); Chakravarty et al. (2025); Yoo et al. (2025); Karim et al. (2025); Li and Ng (2025); Su et al. (2025); Chu et al. (2025a); Bexte et al. (2024); Koutcheme et al. (2024); Schaller et al. (2024); Muñoz Sánchez et al. (2024a); Bexte and Zesch (2025); Geng and Alfter (2025); Bexte et al. (2025); Bloch et al. (2025); Zesch et al. (2025); Urrutia et al. (2025); Zehner et al. (2025); Elaraby and Litman (2025); Tran et al. (2025); Dascalescu et al. (2025)
Grammatical error correction (GEC)	Staruch et al. (2025); Sorokin and Nasyrova (2025); Vainikko et al. (2025); Luhtaru et al. (2024); Yang and Quan (2024); Wang et al. (2024b); Tang et al. (2024); Goto et al. (2025c); Ye et al. (2025); Goto et al. (2025b); Bhattacharyya and Bhattacharya (2025); Li et al. (2025); Qorib et al. (2024); Koo et al. (2024); Katinskaia and Yangarber (2024); Wu et al. (2024); Alhafni and Habash (2025); Cao et al. (2025); Koyama et al. (2025); Goto et al. (2025a); Wang et al. (2025d); Stahlberg and Kumar (2024); Omelianchuk et al. (2024); Kobayashi et al. (2024); Qiu et al. (2025); Östling et al. (2025); Michael and Horbach (2025); Masciolini et al. (2025); Seminck et al. (2025); Staruch (2025); Tyen et al. (2024); Kucharavy et al. (2025); Galletti and Cesaroni (2025); Koutcheme et al. (2025); Marciniak et al. (2025); Petukhova and Kochmar (2025); Tiwari and Rastogi (2025); Correa Busquets et al. (2025)
Text simplification and complexity prediction	Tack (2024); Veeramani et al. (2024); Alfter (2024); Sastre et al. (2024); Padovani et al. (2024); Rooein et al. (2024); Katinskaia et al. (2025); Ribeiro-Flucht et al. (2024); Yousefpoori-Naeim et al. (2024); Wang et al. (2024a); Kelious et al. (2024a); Cristea and Nisioi (2024); Lim and Lee (2024); Yaneva et al. (2024a); Tack et al. (2024); Bulut et al. (2024); Cristea and Nisioi (2024); Shardlow et al. (2024); Enomoto et al. (2024); Dutilleul et al. (2024); Goswami et al. (2024); Strohmaier and Buttery (2024); Überrück-Fries et al. (2024); Kelious et al. (2024b); Alfter (2025); Miyata et al. (2025); Vu et al. (2025); Degraeuwe (2025)
Intelligent tutoring system (ITS)	Pérez-Ortiz et al. (2024); Yasser et al. (2025); An et al. (2025); Kochmar et al. (2025); Park et al. (2025); Wang et al. (2025a); Ribeiro-Flucht et al. (2025); Almasi and Kristensen-McLachlan (2025); Kucheria et al. (2025); Lee et al. (2024a); Pal Chowdhury et al. (2025b); Ikram et al. (2025); Siyan et al. (2024); Wang et al. (2025b)
Content generation	Bodnar (2025); Benedetto et al. (2025); Leite and Lopes Cardoso (2025); Paddags et al. (2024); Berruti et al. (2024); Ma et al. (2025); Liu et al. (2025); Durward and Thomson (2024); Jiang et al. (2025); Wang et al. (2025c); Scaria et al. (2024); Ashok Kumar and Lan (2024); Scarlatos et al. (2024); Stowe et al. (2024); Glandorf and Meurers (2024); Nikolova-Stoupak et al. (2024); Toussaint et al. (2024); Säuberli et al. (2025); Parikh et al. (2025); Kim et al. (2025b); Huovinen and Hämäläinen (2025); Poon et al. (2025)
Datasets, data mining and knowledge extraction	Sung et al. (2025); Akef et al. (2025); Chitez et al. (2025); De Kuthy et al. (2025); Han and Choi (2025); Akter et al. (2025); Sharma and Zhang (2025); Chen and Zhao (2025); Foret et al. (2024); Yang et al. (2025); Velentzas et al. (2024); Rozovskaya (2024); Zhao et al. (2025); Mita et al. (2024); Ding et al. (2024); Beigman Klebanov et al. (2024); Schaller et al. (2025); Dumitran et al. (2025); Mulcaire and Madnani (2025)
Knowledge tracing and learner modelling	Gurin Schleifer et al. (2024); Cristea and Nisioi (2024); Martynova et al. (2025); Srivatsa et al. (2025a); Stearns et al. (2024); Chu et al. (2025b); Hayat et al. (2024); Shi and Mangalam (2025)
Other	Gupta et al. (2025); Ilagan et al. (2024); Schmalz and Tack (2025); de Chillaz et al. (2025); Singh et al. (2025); Muñoz Sánchez et al. (2024b); Ballier and Méli (2024); Degraeuwe and Goethals (2024); Ayari and Li (2024); Hülsing and Horbach (2024); Pal Chowdhury et al. (2025a); Skidmore et al. (2025); N J et al. (2025); Kolagar et al. (2025); Shimabukuro et al. (2025); Sakunkoo and Sakunkoo (2025); Mao et al. (2025); Uluslu and Schneider (2025)

Table 8: Table showing the mapping from high-level tasks to individual papers and their specific tasks.