

A Quantitative Gap Audit of LLM-era NLP for Education: Validity, Deployment, and Multilingual Equity

Farzana Zahid¹ Md. Rakibul Islam¹ Mayesha Kader¹ Israt Jahan Anika¹

¹ Ahsanullah University of Science and Technology

farzana.cse.20210204054@aust.edu, rakib.cse@aust.edu

mayeshakader13@gmail.com, isratjahan7908@email.com

Abstract

It remains unclear whether LLM capability gains translate into measurable student learning. We analyze 137 educational NLP papers (2022–2025) across four sectors: assessment and grammatical error correction, feedback and writing support, content and question generation, and tutoring and dialogue. We introduce three metrics: Validity Gap (VG, share of studies not evaluating learning outcomes), Deployment Readiness Index (DRI, real-world evaluation extent), and Multilingual Equity Index (MEI, language coverage). We find a high validity gap ($VG=0.927$), low deployment readiness ($DRI=0.126$), and limited multilingual coverage ($MEI=0.126$, an explicitly normative index; see Limitations); grammatical error correction shows no learning-improvement evidence at all ($VG=1.000$). DRI correlates with learning-outcome evaluation ($\rho = 0.425$, $p < 0.001$). Our taxonomy and metrics offer one of the first quantitative audits of educational NLP in the LLM era.

1 Introduction

Education is a high-impact NLP domain where systems such as grammatical error correction and tutoring dialogue can directly influence student learning. Recent years have seen a rapid shift toward large language models (LLMs) driven by instruction tuning (Ouyang et al., 2022) and advanced reasoning (OpenAI, 2023), alongside open-weight models enabling broader access (Touvron et al., 2023; Jiang et al., 2023).

Progress in capability has not been matched by evaluation. Educational NLP relies heavily on proxy metrics such as $F_{0.5}$ (Bryant et al., 2019) that do not measure actual learning (Ferris, 2004; Truscott, 1996). Learning science highlights deeper cognitive processes ignored in current evaluations (Kapur, 2016), and the field remains heavily English-centric with limited South Asian coverage (UNESCO, 2023).

This paper makes five contributions: (1) a systematic taxonomy of 137 papers (2022–2025) across four sectors and 14 sub-categories ($\kappa \geq 0.81$); (2) three formal metrics VG, DRI, MEI applied across all sectors; (3) a temporal analysis of the LLM transition; (4) one of the first quantitative analyses of South Asian underrepresentation; and (5) a five-gap research agenda for future work.

2 Related Work

Prior educational NLP work falls into three categories, none systematically analyzing the post-2022 transition.

Task-specific pre-LLM surveys. Surveys on AES (Yannakoudakis et al., 2011; Li and Ng, 2024a), GEC (Bryant et al., 2023) flag evaluation inconsistency but predate LLMs and skip deployment or learning outcomes.

Broad educational NLP reviews. General surveys and BEA overview papers are primarily descriptive and lack formal frameworks for deployment, learning impact, or equity.

LLM-era analyses. Dong et al. (2024) and Min et al. (2022) offer LLM evaluation insights; Kasneci et al. (2023) and odzikowski et al. (2024) discuss educational implications without corpus-level analysis or reproducible metrics.

3 Methodology

3.1 Search Protocol

We conducted a systematic search across four primary sources: the ACL Anthology (ACL, EMNLP, NAACL, EACL, COLING, and co-located workshops including BEA 2022–2025), the Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2022–2025), and arXiv (cs.CL, cs.AI) for preprints with ≥ 15 citations as of the search date, providing a community-

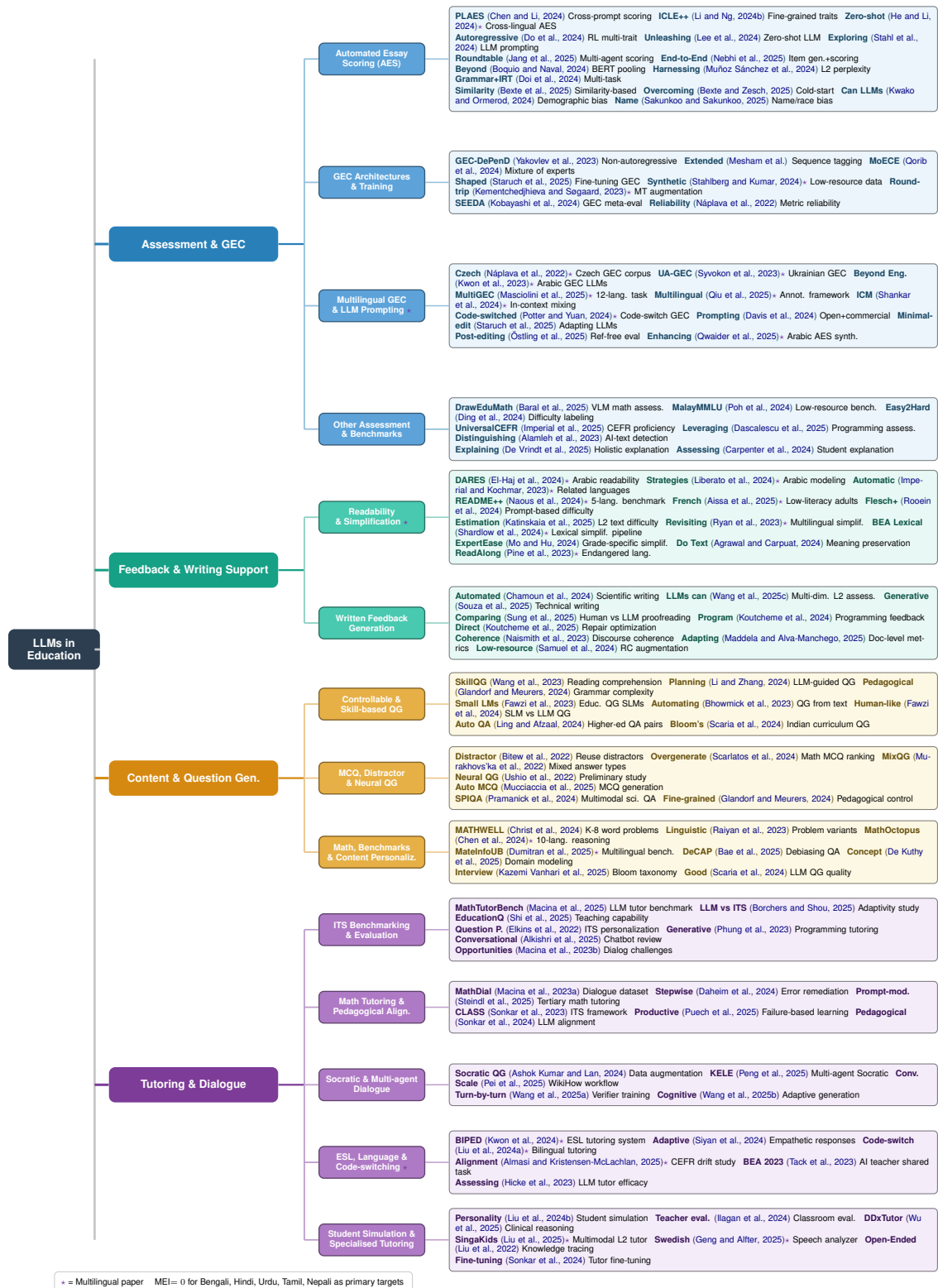


Figure 1: Taxonomy of NLP Education (2022–2025) across four sectors and subcategories.

recognition threshold consistent with prior NLP survey practice ¹. Search terms combined educational domain keywords (*education, learning, student, assessment, tutoring, grammatical error correction, readability, question generation, essay scoring, feedback*) with methodological terms (*NLP, LLM, large language model, transformer*). Papers were included if they (1) proposed or evaluated a computational system addressing an educational language task, and (2) were accepted at a peer-reviewed venue by December 2025, or, for arXiv preprints, met the citation threshold above as of search date.

Screening flowchart. The initial query returned 312 candidate papers. After deduplication across sources (43 removed), title-and-abstract screening (89 removed), and full-text review against the exclusion criteria below (43 removed), we obtained a final corpus of **137 papers**.

3.2 Inclusion and Exclusion Criteria

We excluded: (i) papers addressing educational administration without an NLP contribution; (ii) pure dataset papers without modelling; and (iii) papers where the educational framing was incidental to a broader NLP contribution (e.g., a general reading-comprehension benchmark that happened to use school-level passages).

3.3 Annotation Scheme

Each paper was annotated on five primary dimensions:

Sector four categories (see Section 4);

Model type LLM / BERT-era / Rule-based / Other;

Learning outcome YES: empirical learning measurement (pre/post-test, controlled experiment, or longitudinal study); NO: proxy metrics only;

Fairness YES: explicit fairness/equity analysis; NO: absent;

Deployment type Offline / Simulated / User study / Production.

Inter-annotator agreement. A second annotator independently labelled a stratified random sample of 30 papers (21.9% of the corpus, proportional to sector size) on all five dimensions. Co-

hen’s κ was: sector assignment $\kappa = 0.95$ (near-perfect); learning outcome $\kappa = 1.00$ (perfect). Fairness annotation initially yielded $\kappa = 0.56$ (moderate), revealing guideline ambiguity. We resolved this by tightening the criterion to require an *explicit* fairness analysis rather than incidental mention, and re-annotated the full corpus under the revised guideline. Post-resolution agreement on a fresh 20-paper validation set reached $\kappa = 0.81$ (substantial), which we adopt as the operative reliability estimate for the fairness dimension. All remaining disagreements were reconciled through discussion.

3.4 Three-Metric Gap Framework

Validity Gap (VG).

$$VG(s) = 1 - \frac{L(s)}{N(s)} \quad (1)$$

where $N(s)$ is the number of papers in sector s and $L(s)$ is the number reporting empirical student learning measurement. $VG = 1.0$ means no paper in the sector has demonstrated that students learn from the proposed system.

Deployment Readiness Index (DRI). We assign each paper a continuous deployment score: offline benchmark = 0.00; crowdworker or LLM-simulation = 0.25; small user study ($n < 20$) = 0.50; large user study ($n \geq 20$) = 0.75; production system = 1.00.

$$DRI(s) = \frac{1}{N(s)} \sum_{p \in s} d(p) \quad (2)$$

where $d(p)$ is the deployment score of paper p . Score boundaries were set by analogy to clinical-trial phase criteria (Friedman et al., 2010) and validated against annotator consensus on 10 pilot papers; boundary shifts of ± 0.10 changed no paper’s tier assignment stability rather than the sensitivity of downstream DRI averages and correlations to the weight values themselves.

Multilingual Equity Index (MEI). We assign each paper a language-equity score: English only = 0.0; high-resource non-English added (French, German, Chinese, Japanese) = 0.3; mid-resource language as primary (Arabic, Ukrainian, Czech) = 0.6; low-resource language as primary ($< 10M$ speakers) = 0.8; South Asian language as primary target = 1.0.

¹We set the arXiv threshold at 15 citations after piloting on 20 borderline preprints; all 15 papers above this threshold that overlapped with our domain criteria were independently judged relevant by both annotators.

$$\text{MEI}(s) = \frac{1}{N(s)} \sum_{p \in s} m(p) \quad (3)$$

The ceiling of 1.0 is reserved for South Asian primary targeting because these languages Bengali, Hindi, Urdu, Tamil, and Nepali collectively serve over 600 million school-age learners (UNESCO, 2023) yet have no dedicated educational NLP system in our corpus, making them the single most severe and most populous equity gap in the field. All other score tiers follow established resource-level classifications in the low-resource NLP literature (Joshi et al., 2020).

4 Taxonomy of NLP Education Research

Our corpus of 137 papers organises into four sectors according to primary pedagogical function: **Assessment & GEC** ($n=53$, 38.7%), **Feedback & Writing Support** ($n=26$, 19.0%), **Content & Question Generation** ($n=25$, 18.2%), and **Tutoring & Dialogue** ($n=33$, 24.1%). Table 1 summarises the three-metric gap profile for each sector and all sub-categories. Each sector is discussed in turn below.

4.1 Assessment & GEC ($n=53$, 38.7%)

This is the largest sector, comprising AES and GEC, both of which evaluate or improve student language.

Automated Essay Scoring. LLMs have displaced BERT-era regression as the dominant AES paradigm, with zero/few-shot scoring (He and Li, 2024; Stahl et al., 2024) and multi-agent debate improving scoring consistency (Jang et al., 2025).

Fairness is active in AES (39.1% of papers): LLMs infer demographic attributes from essay text (Kwako and Ormerod, 2024), and student names alone shift scores along race/gender lines (Sakunkoo and Sakunkoo, 2025). Cross-prompt generalisation remains open, addressed via zero-shot cross-lingual transfer (He and Li, 2024) and similarity-based scoring (Bexte et al., 2025).

Finding. $\text{VG} = 0.947$ across 19 papers; only one reports pre/post learning measurement, limited to a single session.

Grammatical Error Correction. GEC has shifted from seq2seq/tagging to LLM prompting (Davis et al., 2024; Staruch et al., 2025). Multilingual GEC has expanded: MultiGEC-2025 covers 12 languages (Masciolini et al., 2025), build-

ing on Czech (Náplava et al., 2022), Ukrainian (Syvokon et al., 2023), and Arabic (Kwon et al., 2023) corpora. Architectural work includes mixture-of-experts GEC (Qorib et al., 2024), non-autoregressive decoding (Yakovlev et al., 2023), and syntax-enhanced parsing (Zhang et al., 2022).

Finding. $\text{VG} = 1.000$ across all 23 papers: none measure whether learners’ grammatical accuracy improved.

Other Assessment. The remaining 11 papers cover short-answer scoring (Takano and Ichikawa, 2022), CEFR classification (Jalota et al., 2022; Imperial et al., 2025), AI-text detection (Alamleh et al., 2023), and programming assessment (Dascalescu et al., 2025). All 11: $\text{VG} = 1.000$, $\text{DRI} = 0.000$ none tested with real learners.

4.2 Feedback & Writing Support ($n=26$, 19.0%)

This sector supports writing via readability assessment, simplification, and feedback generation.

Readability and Simplification. Expanded multilingually: Naous et al. (2024) benchmark 5 languages; Liberato et al. (2024) address Arabic readability. Ryan et al. (2023) unify multilingual simplification benchmarks; BEA 2024 organised a multilingual lexical simplification task (Shardlow et al., 2024). Highest sub-sector MEI (0.377), yet $\text{VG} = 1.000$: no paper measures reader comprehension gains.

Writing Feedback and Programming Support. LLMs give feedback on scientific writing (Chamoun et al., 2024), technical documents (Souza et al., 2025), and L2 essays (Wang et al., 2025c; Sung et al., 2025), evaluated on fluency/relevance—never writer improvement. Programming feedback (Koutchme et al., 2024, 2025) covers error explanation/repair, evaluated only on test-set accuracy.

Finding. $\text{VG} = 1.000$ across both sub-sectors no paper shows writers improve. Writing Feedback’s $\text{DRI} = 0.308$ (highest sub-sector) still evaluates the feedback, not learner outcomes.

4.3 Content & Question Generation ($n=25$, 18.2%)

Question generation (QG) dominates. LLM-based QG has superseded neural seq2seq (Murakhovs’ka et al., 2022; Ushio et al., 2022); pedagogically

Sector / Sub-task	n	%	LLM%	Fair%	Learn%	DRI	MEI	VG↑
Assessment & GEC	53	38.7	62.3	39.6	1.9	0.1	0.2	98.1
Automated Essay Scoring	19		57.9	31.6	5.3	0.2	0.1	94.7
Grammatical Error Correction	23		60.9	39.1	0.0	0.1	0.2	100.0
Other Assessment	11		72.7	54.5	0.0	0.0	0.1	100.0
Feedback & Writing Support	26	19.0	69.2	30.8	0.0	0.2	0.2	100.0
Readability & Simplification	13		69.2	38.5	0.0	0.0	0.4	100.0
Writing Feedback & Programming	13		69.2	23.1	0.0	0.3	0.0	100.0
Content & Question Generation	25	18.2	88.0	16.0	4.0	0.1	0.0	96.0
Question Generation (QG/MCQ)	20		85.0	15.0	5.0	0.2	0.0	95.0
Math & Problem Generation	5		100.0	20.0	0.0	0.1	0.2	100.0
Tutoring & Dialogue	33	24.1	90.9	15.2	24.2	0.2	0.1	75.8
Dialogue Tutoring Systems	28		96.4	17.9	21.4	0.3	0.1	78.6
Student Simulation & ITS	5		60.0	0.0	40.0	0.2	0.0	60.0
Field-wide	137	100.0	75.2	27.7	7.3	0.1	0.1	92.7

Table 1: Cross-sectoral gap analysis ($n = 137$, 2022–2025).

grounded generation anchored to Bloom’s taxonomy (Scaria et al., 2024; Wang et al., 2023) is the most principled direction. Emerging work: controllable grammar-complexity generation (Glandorf and Meurers, 2024) and spaced-repetition sentence generation (Paddags et al., 2024).

Math and Problem Generation. Entirely LLM-driven (100% adoption): Christ et al. (2024) introduce MATHWELL for K–8; (Chen et al., 2024) address multilingual reasoning.

Finding. Lowest sector MEI (0.016) despite generating language-learning materials—every QG paper targets English only (MEI = 0.000). VG = 0.960; the sole learning-outcome paper measures task accuracy, not skill acquisition.

4.4 Tutoring & Dialogue ($n = 33$, 24.1%)

Most pedagogically mature sector: lowest VG, highest DRI. LLMs have reshaped intelligent tutoring (Borchers and Shou, 2025; Macina et al., 2025), Socratic dialogue (Macina et al., 2023a; Peng et al., 2025), ESL/code-switching (Kwon et al., 2024; Liu et al., 2024a; Siyan et al., 2024), and student simulation (Tack et al., 2023). Learning-gain studies cluster here (Liu et al., 2025; Wu et al., 2025; Wang et al., 2025b; Steindl et al., 2025; Puech et al., 2025), explaining the lower VG.

Finding. VG = 0.758, DRI = 0.189 are both best-in-field yet fairness engagement is lowest (15.2%), despite the greatest bias risk from di-

rect learner interaction.

5 Cross-Sectoral Analysis

Table 1 and Figure 4 present the three-metric gap profile across sectors.

5.1 The Validity Gap

Field-wide VG = 0.927: 127 of 137 papers never measure real student learning; only 10 report empirical measurement (8 in Tutoring & Dialogue)—a sign of metric misalignment, as the field relies on proxy metrics ($F_{0.5}$, Pearson r , BLEU) over student benefit. The 10 that do measure learning share three traits: real students, outcomes measured over time, and learning operationalised as skill change rather than task accuracy. DRI correlates with learning outcome at $\rho = 0.425$ ($p < 0.001$), our strongest paper-level correlation; part of this association is expected by construction, since our annotation scheme requires real-learner contact to qualify as deployed, which structurally limits how learning outcomes can be measured at low DRI.

5.2 Deployment Readiness

Field-wide DRI = 0.126. The binary “deployed” count overstates readiness: 77.4% of papers use offline benchmarks only ($d = 0.00$). Year vs. DRI correlates at $\rho = 0.326$ ($p < 0.001$), the strongest temporal correlation in our matrix, driven by 2025 recovery (DRI = 0.231). Assessment & GEC

Dimension	Total (n=137)	2022 (n=22)	2023 (n=22)	2024 (n=54)	2025 (n=39)	Trend	Spearman (n=137)
LLM adoption	78.8%	◦45%	86%	83%	87%	↑↑	$\rho = +0.251^{**}$
Learning outcome (VG)	7.3% (0.927)	◦9%	×0%	×2%	◦18%	↑	$\rho = +0.172^*$
DRI (mean)	0.126	0.045	0.125	0.083	0.231	↑	$\rho = +0.326^{***}$
Fairness discussed	27.7%	◦23%	◦27%	◦28%	◦31%	≈	$\rho = +0.054$ n.s.
MEI (mean)	0.126	0.082	0.136	0.148	0.113	≈	$\rho = +0.047$ n.s.
Low-resource focus	7.3%	◦5%	◦9%	◦11%	×5%	↓	$\rho = -0.200^*$

Table 2: Temporal gap matrix with paper-level Spearman correlations ($n = 137$). Significance: $***p < 0.001$; $**p < 0.01$; $*p < 0.05$; n.s. not significant. $>40\%$ ◦ 10–40% × $<10\%$. DRI and MEI reported as means (not proportions). Learning collapsed to 0% in 2023 coinciding with peak LLM adoption (86%). DRI has the strongest temporal correlation ($\rho = +0.326^{***}$). Low-resource focus is the only dimension in sustained decline.

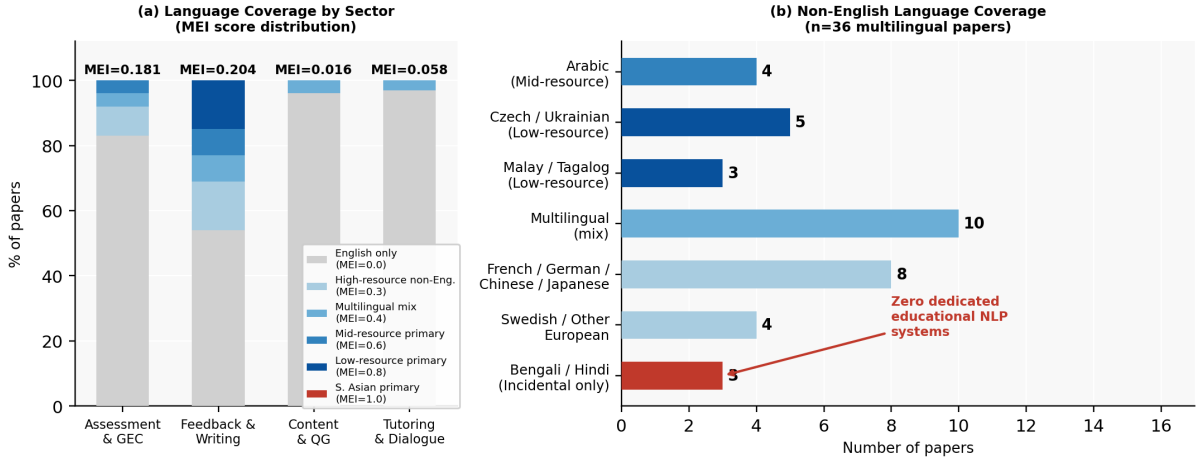


Figure 2: **Multilingual equity gap.** (a) MEI score distribution by sector; Feedback & Writing leads (MEI=0.204) while Content & QG lags (MEI=0.016). (b) Breakdown of the 36 non-English papers by language family. Bengali, Hindi, and Urdu appear only incidentally in multilingual benchmarks—no paper targets any South Asian language as its primary educational medium.

(DRI = 0.075) remains effectively untested with real learners.

5.3 Multilingual Equity

Field-wide MEI = 0.126 (a normative weighting, not a neutral benchmark; see Section 8). Of 137 papers, 101 (73.7%) address English only (MEI = 0.0). Zero papers achieve MEI = 1.0 in the sense of genuinely targeting a South Asian language as their primary educational medium. Bengali (230M native speakers), Hindi (600M), Urdu (170M), Tamil (80M), and Nepali (16M) have no dedicated educational NLP system in our entire corpus. Content & QG achieves MEI = 0.016 the lowest sector score despite being the sector most responsible for generating language learning materials.

6 Temporal Analysis

Table 2 and Figure 3 present the temporal gap matrix with paper-level Spearman correlations ($n =$

137).

6.1 The LLM Inflection (2022–2023)

LLM adoption increased sharply from 45% in 2022 to 86% in 2023, reflecting the emergence of ChatGPT (November 2022), GPT-4 (March 2023), and the growing availability of open-weight language models ($\rho = 0.251$, $p < 0.01$). The consequences were immediate: learning outcome measurement declined from 9% to 0%, while the Deployment Readiness Index (DRI) increased from 0.045 to 0.125 before falling to 0.083 in 2024.

6.2 Post-LLM Recovery (2024–2025)

Recovery is evident by 2025: learning outcome evaluation rose to 18%, and DRI reached its highest value (DRI = 0.231). Low-resource language support, however, moved the opposite way rising from 5% (2022) to 9% (2023) before falling back to 5% in 2025 ($\rho = -0.200$, $p = 0.019$): the field is deploying more while serving fewer low-

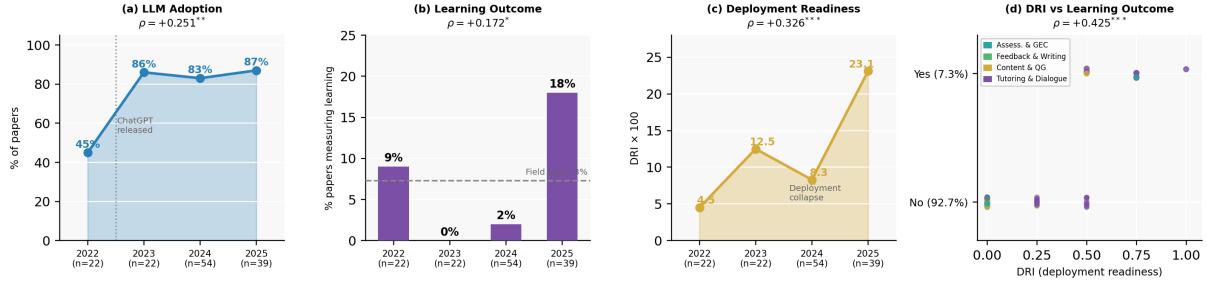


Figure 3: **Temporal trends in educational NLP (2022–2025).** (a) LLM adoption rose from 45% to 87% ($\rho = +0.251^{**}$). (b) Learning outcome measurement collapsed to 0% in 2023 and recovered to 18% by 2025. (c) DRI shows strongest temporal growth ($\rho = +0.326^{***}$), driven by 2025 deployment recovery. (d) Scatter of paper-level DRI vs learning outcome: $\rho = +0.425$, $p < 0.001$ —the structural link between deployment and learning accountability.

resource communities.

6.3 Fairness: Stable but Low

Fairness shows no significant temporal trend ($\rho = 0.054$, $p = 0.530$, n.s.), with only 27.7% of papers discussing it overall. Assessment & GEC leads at 39.6%, driven mainly by demographic-bias work

7 Research Gaps

7.1 Gap 1: The Validity Crisis (VG = 0.927)

VG = 1.000 in GEC and Feedback means zero evidence of learning benefit in either field. We propose that BEA and AIED reviewers require a *learning plausibility argument*: a brief, learning-science-grounded account of why a proposed system could improve student outcomes, not a demand for a clinical trial, but for intellectual accountability.

7.2 Gap 2: Deployment Readiness (DRI = 0.126)

DRI = 0.075 in Assessment & GEC leaves the largest sector essentially untested with real learners, due to *infrastructure* costs (LLM API access), *evaluation complexity* (IRB approval), and incentives that reward benchmark performance over classroom impact. The DRI–VG correlation ($\rho = 0.425$, $p < 0.001$) is consistent with deployment being a necessary condition for closing the validity gap, though part of this relationship follows from our coding scheme rather than being a purely empirical finding.

7.3 Gap 3: Multilingual Equity (MEI = 0.126)

MEI = 0.0 for Bengali, Hindi, Urdu, Tamil, and Nepali as primary targets reflecting our normative weighting choice is a measurable equity gap: these

languages serve over 600 million school-age learners including 30 million in Bangla-medium education alone—with no dedicated educational NLP system in our corpus (Figure 2).

7.4 Gap 4: Cross-Sector Fairness

Fairness (27.7% overall) is lowest in Tutoring & Dialogue (15.2%), the sector with the most direct learner contact. A chatbot deployed without demographic bias analysis is not merely incomplete, it is potentially harmful at scale.

7.5 Gap 5: Learning Science Integration

Educational NLP rarely engages learning science frameworks—deliberate practice (Ericsson et al., 1993), cognitive load theory (Sweller, 1988), productive failure (Kapur, 2016)—that predict when feedback and tutoring improve learning. Grounding systems in these frameworks would yield falsifiable hypotheses and easier-to-design validity studies.

8 Ethical and Societal Implications

This paper itself measures NLP’s social impact, the question this workshop poses: rather than proposing a new system, we audit whether 137 existing educational NLP systems demonstrate benefit to real learners, finding that 92.7% do not. To us, “positive impact” means closing this gap between reported performance and evidence of student benefit, not publishing another benchmark score. Our metrics enable field-level accountability rather than ranking individual papers a low VG or DRI does not imply a paper lacks value, since many contributions (e.g., architecture or metric-reliability studies) are not designed to measure

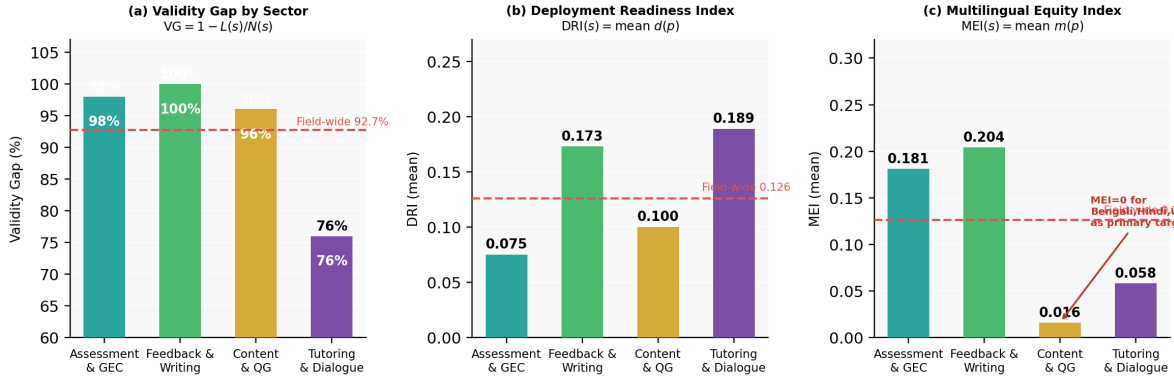


Figure 4: **Three-metric gap profile by sector.** (a) Validity Gap (VG) per sector; dashed line = field-wide 92.7%. (b) Deployment Readiness Index (DRI) per sector; dashed = 0.126. (c) Multilingual Equity Index (MEI) per sector; dashed = 0.126. Colors correspond to sectors throughout the paper. Content & QG achieves the lowest MEI (0.016) despite generating language learning materials. Tutoring & Dialogue achieves the best DRI (0.189) and lowest VG (75.8%).

classroom learning. Our MEI is explicitly normative: it assigns the highest weight to South Asian languages to foreground the disparity between their learner population and their absence from educational NLP, not to claim they are objectively least-resourced by every criterion, and should not be read as a neutral cross-language benchmark. Finally, treating learning-outcome evaluation as a required gate could discourage early-stage exploratory work; our learning-plausibility argument (Section 7.1) is intended as a lightweight accountability check, not a barrier to publication.

9 Conclusion

We present one of the first quantitative gap audit of educational NLP in the LLM era: 137 papers analyzed via Validity Gap ($VG = 0.927$), Deployment Readiness Index ($DRI = 0.126$), and Multilingual Equity Index ($MEI = 0.126$). GEC and Feedback sectors reach $VG = 1.000$ -zero evidence of learning benefit while DRI correlates with learning accountability at $\rho = 0.425$ ($p < 0.001$), and South Asian languages score $MEI = 0.0$ across all 137 papers. Until the field consistently asks “does this system help students learn?”, the validity gap will persist and the learners who need it most will remain unserved.

Limitations

Our study is scoped deliberately to enable a controlled, reproducible audit. First, we focus on ACL Anthology, BEA, AIED, and high-citation arXiv preprints as NLP-primary venues; this keeps annotation criteria consistent across the corpus,

though future work could extend the framework to education-primary venues such as JEDM, IEEE TLT, and CHI. Second, our three-metric framework (VG, DRI, MEI) is designed for cross-sector comparability rather than fine-grained per-study nuance; applying it to richer, study-level detail is a natural next step. Third, the survey window (January 2022–December 2025) captures the full LLM transition period central to our analysis, with earlier BERT-era context drawn from established prior surveys. We commit to publicly releasing the full annotated corpus (paper identifiers and all five dimensional labels), the annotation guidelines, and the analysis code on GitHub upon publication, to support full reproducibility of our metrics.

References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? a human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzevren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. 2025. [Assessing French readability for adults with low literacy: A global and local perspective](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20517–20539, Suzhou, China. Association for Computational Linguistics.
- Hosam Alamlah, Ali Alqahtani, and AbdElRahman ElSaid. 2023. [Distinguishing human-written and chatgpt-generated text using machine learning](#). pages 154–158.
- Wasin Alkishri, Jabar Yousif, Yousuf Al Husaini, and Al-Bahri Mahmood. 2025. [Conversational ai in ed-](#)

- education: A general review of chatbot technologies and challenges. *Journal of Logistics Informatics and Service Science*, 12:264–282.
- 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.
- 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.
- Suyoung Bae, YunSeok Choi, and Jee-Hyong Lee. 2025. [DeCAP: Context-adaptive prompt generation for debiasing zero-shot question answering in large language models](#). 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 12555–12574, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sami Baral, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. 2025. [Drawedumath: Evaluating vision language models with expert-annotated students hand-drawn math images](#). pages 6902–6920.
- Marie Bexthe, 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 Bexthe and Torsten Zesch. 2025. [Overcoming the cold-start problem in supervised content scoring using zero-shot LLM-generated training data](#). In *Proceedings of BEA 2025*.
- Ayan Kumar Bhowmick, Ashish Jagmohan, Aditya Vempaty, Prasenjit Dey, Leigh Hall, Jeremy Hartman, Ravi Kokku, and Hema Maheshwari. 2023. Automating question generation from educational text. In *Artificial Intelligence XL*, pages 437–450, Cham. Springer Nature Switzerland.
- Semere Bitew, Amir Hadifar, Lucas Sterckx, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. [Learning to reuse distractors to support multiple choice question generation in education](#). *IEEE Transactions on Learning Technologies*, PP:1–16.
- 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.
- Conrad Borchers and Tianze Shou. 2025. [Can large language models match tutoring system adaptivity? a benchmarking study](#). *Preprint*, arXiv:2504.05570.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. [The BEA-2019 shared task on grammatical error correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. [Grammatical error correction: A survey of the state of the art](#). *Computational Linguistics*, 49(3):643–701.
- 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.
- Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. [Automated focused feedback generation for scientific writing assistance](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. 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.
- Bryan R Christ, Jonathan Kropko, and Thomas Hartvigsen. 2024. [MATHWELL: Generating educational math word problems using teacher annotations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11914–11938, Miami, Florida, USA. Association for Computational Linguistics.
- Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. [Stepwise verification and remediation of student reasoning errors with large language model tutors](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.
- Stefan Dascalescu, Marius Dumitran, and Mihai Alexandru Vasiluta. 2025. [Leveraging gener-](#)

- ative 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 Gierbach, 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.
- Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. 2024. [Easy2hard-bench: Standardized difficulty labels for profiling llm performance and generalization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 44323–44365. Curran Associates, Inc.
- Heejin Do, Sangwon Ryu, and Gary Lee. 2024. [Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. 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.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). Preprint, arXiv:2301.00234.
- Marius Dumitran, Mihnea Buca, and Theodor Moroiu. 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.
- Mo El-Haj, Sultan Almujaivel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. [DARES: Dataset for Arabic readability estimation of school materials](#). In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.
- Sabina Elkins, Robert Belfer, Ekaterina Kochmar, Iulian Serban, and Jackie C. K. Cheung. 2022. [Question personalization in an intelligent tutoring system](#). Preprint, arXiv:2206.14145.
- Karl Ericsson, Ralf Krampe, and Clemens Tesch-Roemer. 1993. [The role of deliberate practice in the acquisition of expert performance](#). *Psychological Review*, 100:363–406.
- Fares Fawzi, Sadie Amini, Sahan Bulathwela, and Zhengxiang Shi. 2023. [Small generative language models for educational question generation](#). In *Proceedings of the NeurIPS 2023 Workshop on Generative AI for Education (GAIED)*, New Orleans, Louisiana, USA. Workshop paper at NeurIPS 2023.
- Fares Fawzi, Sarang Balan, Mutlu Cukurova, Emine Yilmaz, and Sahan Bulathwela. 2024. [Towards human-like educational question generation with small language models](#). In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 295–303, Cham. Springer Nature Switzerland.
- Dana R. Ferris. 2004. [The grammar correction debate in l2 writing: Where are we, and where do we go from here? \(and what do we do in the meantime ?\)](#). *Journal of Second Language Writing*, 13(1):49–62. Conceptualizing Discourse/Responding to Text.
- Lawrence M. Friedman, Curt D. Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. 2010. *Fundamentals of Clinical Trials*, 4th edition. Springer, New York.
- 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.
- 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.
- Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. [Assessing the efficacy of large language models in generating accurate teacher responses](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 745–755, Toronto, Canada. 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.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [UniversalCEFR: Enabling open multilingual research on language proficiency assessment](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.
- Joseph Marvin Imperial and Ekaterina Kochmar. 2023. [Automatic readability assessment for closely related languages](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.
- Richa Jalota, Peter Bourgonje, Jan Van Sas, and Huiyan Huang. 2022. [Mitigating learnerese effects for CEFR classification](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 14–21, Seattle, Washington. Association for Computational Linguistics.
- Jinhee Jang, Ayoung Moon, Minkyung Jung, Young-Bin Kim, and Seung Jin Lee. 2025. [LLM agents at the roundtable: A multi-perspective and dialectical reasoning framework for essay scoring](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19674–19687, Suzhou, China. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Manu Kapur. 2016. [Examining productive failure, productive success, unproductive failure, and unproductive success in learning](#). *Educational Psychologist*, 51(2):289–299.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. [Chatgpt for good? on opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- 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.
- Fatemeh Kazemi Vanhari, Christopher Anand, and Charles Welch. 2025. [Analyzing interview questions via bloom’s taxonomy to enhance the design thinking process](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 582–593, Vienna, Austria. Association for Computational Linguistics.
- Yova Kementchedjhieva and Anders Søgaard. 2023. [Grammatical error correction through round-trip machine translation](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2208–2215, Dubrovnik, Croatia. Association for Computational Linguistics.
- Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. [Revisiting meta-evaluation for grammatical error correction](#). *Transactions of the Association for Computational Linguistics*, 12:837–855.
- Charles Koutchme, 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 Koutchme, 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.

- 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.
- Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [Beyond English: Evaluating LLMs for Arabic grammatical error correction](#). In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.
- Soonwoo Kwon, Sojung Kim, Minju Park, Seunghyun Lee, and Kyuseok Kim. 2024. [BIPED: Pedagogically informed tutoring system for ESL education](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3389–3414, Bangkok, Thailand. Association for Computational Linguistics.
- Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. [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.
- Kunze Li and Yu Zhang. 2024. [Planning first, question second: An LLM-guided method for controllable question generation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand. Association for Computational Linguistics.
- Shengjie Li and Vincent Ng. 2024a. [Automated essay scoring: A reflection on the state of the art](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.
- Shengjie Li and Vincent Ng. 2024b. [ICLE++: Modeling fine-grained traits for holistic essay scoring](#). 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 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.
- Juan Liberato, Bashar Alhafni, Muhamed Khalil, and Nizar Habash. 2024. [Strategies for Arabic readability modeling](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 55–66, Bangkok, Thailand. Association for Computational Linguistics.
- Jintao Ling and Muhammad Afzaal. 2024. [Automatic question-answer pairs generation using pre-trained large language models in higher education](#). *Computers and Education: Artificial Intelligence*, 6:100252.
- Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. [Open-ended knowledge tracing for computer science education](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3849–3862, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhengyuan Liu, Geyu Lin, Hui Li Tan, Huayun Zhang, Yanfeng Lu, Xiaoxue Gao, Stella Xin Yin, Sun He, Hock Huan Goh, Lung Hsiang Wong, and Nancy F. Chen. 2025. [SingaKids: A multilingual multimodal dialogic tutor for language learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1244–1253, Vienna, Austria. Association for Computational Linguistics.
- Zhengyuan Liu, Stella Xin Yin, and Nancy Chen. 2024a. [Optimizing code-switching in conversational tutoring systems: A pedagogical framework and evaluation](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Kyoto, Japan. Association for Computational Linguistics.
- Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024b. [Personality-aware student simulation for conversational intelligent tutoring systems](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. [Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. [Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors](#). *Preprint*, arXiv:2502.18940.
- Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. [Opportunities and challenges in neural dialog tutoring](#). *Preprint*, arXiv:2301.09919.
- Mounica Maddela and Fernando Alva-Manchego. 2025. [Adapting sentence-level automatic metrics for document-level simplification evaluation](#). 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 6444–6459, Albuquerque, New Mexico. 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.

- Stuart Mesham, Christopher Bryant, Marek Rei, and Zheng Yuan. [An extended sequence tagging vocabulary for grammatical error correction](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1608–1619, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) *Preprint*, arXiv:2202.12837.
- Kaijie Mo and Renfen Hu. 2024. [ExpertEase: A multi-agent framework for grade-specific document simplification with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.
- Sérgio Silva Mucciaccia, Thiago Meireles Paixão, Filipe Wall Mutz, Claudine Santos Badue, Alberto Ferreira de Souza, and Thiago Oliveira-Santos. 2025. [Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, Abu Dhabi, UAE. Association for Computational Linguistics.
- Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024. [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.
- Lidiya Murakhovska, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. [MixQG: Neural question generation with mixed answer types](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. Association for Computational Linguistics.
- Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. [Automated evaluation of written discourse coherence using GPT-4](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.
- Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. [ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.
- Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. [Czech grammar error correction with a large and diverse corpus](#). *Transactions of the Association for Computational Linguistics*, 10:452–467.
- 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.
- OpenAI. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- 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.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*.
- 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.
- Jiahuan Pei, Fanghua Ye, Xin Sun, Wentao Deng, Koen Hindriks, and Junxiao Wang. 2025. [Conversational education at scale: A multi-LLM agent workflow for procedural learning and pedagogic quality assessment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2984–2997, Suzhou, China. Association for Computational Linguistics.
- Xian Peng, Pan Yuan, Dong Li, Junlong Cheng, Qin Fang, and Zhi Liu. 2025. [KELE: A multi-agent framework for structured socratic teaching with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16342–16362, Suzhou, China. Association for Computational Linguistics.
- Tung Phung, Victor-Alexandru Pdurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. [Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors](#). *Preprint*, arXiv:2306.17156.
- Aidan Pine, David Huggins-Daines, Eric Joanis, Patrick Littell, Marc Tessier, Delasie Torkornoo, Rebecca Knowles, Roland Kuhn, and Delaney Lothian. 2023. [ReadAlong studio web interface for digital interactive storytelling](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages

- 163–172, Toronto, Canada. Association for Computational Linguistics.
- Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Foong Chee Mun, and Chee Seng Chan. 2024. [MalayMMLU: A multitask benchmark for the low-resource Malay language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 650–669, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Potter and Zheng Yuan. 2024. [LLM-based code-switched text generation for grammatical error correction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16957–16965, Miami, Florida, USA. Association for Computational Linguistics.
- Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37. Publisher Copyright: © 2024 Neural information processing systems foundation. All rights reserved.; 38th Conference on Neural Information Processing Systems, NeurIPS 2024 ; Conference date: 09-12-2024 Through 15-12-2024.
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. [Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Syed Rifat Raiyan, Md Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. [Math word problem solving by generating linguistic variants of problem statements](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 362–378, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Michael J Ryan, Tarek Naous, and Wei Xu. 2023. [Revisiting non-English text simplification: A unified multilingual benchmark](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. [Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. [In-context mixing \(ICM\): Code-mixed prompts for multilingual LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

- 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 Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. [The BEA 2024 shared task on the multilingual lexical simplification pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Yao Shi, Rongkeng Liang, and Yong Xu. 2025. [EducationQ: Evaluating LLMs’ teaching capabilities through multi-agent dialogue framework](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828, 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.
- Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. [CLASS: A design framework for building intelligent tutoring systems based on learning science principles](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.
- Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. [Pedagogical alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA. Association for Computational Linguistics.
- Karen de Souza, Alexandre Nikolaev, and Maarit Koponen. 2025. [Generative AI for technical writing: Comparing human and LLM assessments of generated content](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 661–679, Tallinn, Estonia. University of Tartu Library.
- 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, 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.
- Sebastian Steindl, Fabian Brunner, Nada Sissouno, Dominik Schwagerl, Florian Schöler-Niewiera, and Ulrich Schäfer. 2025. [On the effectiveness of prompt-moderated LLMs for math tutoring at the tertiary level](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11310–11323, Suzhou, China. 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.
- John Sweller. 1988. [Cognitive load during problem solving: Effects on learning](#). *Cognitive Science*, 12(2):257–285.
- Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. [UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language](#). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. [The BEA 2023 shared task on generating AI teacher responses in educational dialogues](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.
- Shunya Takano and Osamu Ichikawa. 2022. [Automatic scoring of short answers using justification cues estimated by BERT](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 8–13, Seattle, Washington. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- John Truscott. 1996. The case against grammar correction in L2 writing classes. *Language Learning*, 46(2):327–369.

- UNESCO. 2023. [Global education monitoring report 2023: Technology in education – a tool on whose terms?](#) Technical report, UNESCO, Paris.
- Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. [Generative language models for paragraph-level question generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025a. [Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of llms as your coding tutors](#). pages 12416–12436.
- Qingsong Wang, Tao Wu, Wang Lin, Yueying Feng, Gongsheng Yuan, Chang Yao, and Jingyuan Chen. 2025b. [Cognitive-level adaptive generation via capability-aware retrieval and style adaptation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11054–11069, Suzhou, China. Association for Computational Linguistics.
- Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2023. [SkillQG: Learning to generate question for reading comprehension assessment](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13833–13850, Toronto, Canada. Association for Computational Linguistics.
- Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025c. [LLMs can perform multi-dimensional analytic writing assessments: A case study of L2 graduate-level academic English writing](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8637–8663, Vienna, Austria. Association for Computational Linguistics.
- Qian Wu, Zheyao Gao, Longfei Gou, and Qi Dou. 2025. [DDxTutor: Clinical reasoning tutoring system with differential diagnosis-based structured reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30934–30957, Vienna, Austria. Association for Computational Linguistics.
- Konstantin Yakovlev, Alexander Podolskiy, Andrey Bout, Sergey Nikolenko, and Irina Piontkovskaya. 2023. [GEC-DePenD: Non-autoregressive grammatical error correction with decoupled permutation and decoding](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1546–1558, Toronto, Canada. Association for Computational Linguistics.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.
- Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. [SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kacper Odziowski, Peter W. Foltz, and John T. Behrens. 2024. [Generative ai and its educational implications](#). *Preprint*, arXiv:2401.08659.