Rewrites Are Not Lessons: Teachable Edits as an Output Specification for Multimodal Language Learning

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Multimodal large language models are increasingly described as tutors for language learning: they take a handwritten page, a spoken utterance, or a slide and return fluent target-language text. We argue that this description confuses a *correction engine* with a *teaching intervention*. For systems claiming to teach through corrective feedback, we propose a *teachable edit*: localized on the learner's production, typed by error category, conservative with respect to the learner's interlanguage, and accompanied either by a faithful, level-appropriate explanation or by a recorded, calibrated abstention. Fluency of a rewritten string is the wrong success criterion. Existing explainable and pedagogically targeted grammatical error correction (GEC) work already supplies evidence words, error types, and grammar-item evaluation on text. What remains underspecified is a single record that also localizes into a multimodal artifact, scores conservatism and explanation-to-edit faithfulness, and requires calibrated abstention. We illustrate the specification on Chinese and Arabic examples, using the latter to show how language-specific error categories and right-to-left script affect localization while leaving the functional structure of the record unchanged. We state that record, compare it with EXPECT, EXCGEC, and JELV, and separate correction correctness from feedback quality and from learning outcomes. The correction, localization, and reference-based estimates of conservatism can be measured on existing benchmarks; the explanation, level, and abstention fields need a small annotation pass, which we specify and do not run.

1 Introduction

Surveys of multimodal LLMs in education already describe these models as supporting "personalized learning paths" and language acquisition (Xing et al., 2024; Yin et al., 2024). The typical demonstration is straightforward. The user uploads a handwritten sentence; the model returns a fluent target-language string. This output is then presented as tutoring. It is not.

Consider a schematic page from a Chinese as Foreign Language (CFL) course whose intended meaning is "I have a teacher," written 我有一个老师 *Wǒ yǒu yí gè lǎoshī* where the task asked for the polite form 我有一位老师 *Wǒ yǒu yí wèi lǎoshī*. A correction engine emits 我有一位老师 *Wǒ yǒu yí wèi lǎoshī* and stops. A teaching move would instead point at the classifier slot on the page, name the register contrast, leave the rest of the sentence alone, and either explain that contrast or abstain. The first output can look like a tutor in a product screenshot. It is a rewrite, not a lesson.

The distinction is not specific to Chinese. A learner of Modern Standard Arabic writes الطالبة مجتهد (*al-ṭāliba mujtahid*), "the female student is diligent," where the target is الطالبة مجتهدة (*al-ṭāliba mujtahida*). A correction engine can return the corrected sentence and stop. A teachable edit points to the adjective مجتهد *mujtahid*, types the problem as gender agreement, adds the feminine ending, and explains that the predicate adjective must agree with the feminine subject الطالبة *al-ṭāliba*, leaving the rest of the sentence untouched. If the handwritten final character is visually ambiguous, the system coarsens the pointer to the whole adjective or abstains rather than presenting a recognition error as learner feedback. This Arabic example is constructed for illustration and is not drawn from a visual Arabic GEC benchmark.

Second-language writing research distinguishes direct correction, indirect marking, and metalinguistic explanation (Ellis, 2009; Ferris, 2002; Bitchener and Storch, 2016). Educational NLP has largely optimized the cleaned draft or the reference-matching edit set (Bryant et al., 2023; Ng et al., 2014; Bryant et al., 2017). Visual CGEC

Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning, pages 56–64
November 9, 2026 ©2026 Association for Computational Linguistics

makes the same move on images (Wang et al., 2025).

This paper takes three linked positions.

Claim A (capability is not instruction). Showing that a model can read a page and emit a correct sentence does not show that it can teach. Generation quality and instructional validity come apart (Caines et al., 2023).

Claim B (rewrites are not lessons). Full-sentence generation obscures what the learner already knows (Selinker, 1972). A system that cannot show selective feedback fails low-proficiency learners: they cannot see which forms were already usable and which form was the lesson.

Claim C (explanation is a first-class output). The rationale is part of the intervention. A fluent wrong rule is worse than marking a span and stopping (Ellis, 2009; Fei et al., 2023).

We take VisCGEC as the main public visual CGEC benchmark built on CFL handwriting and explicitly framed for language learners (Wang et al., 2025). If that setting adopts a teachable-edit record, downstream visual GEC papers are likely to follow. The argument is stated so it can transfer to speech and slides; we do not treat those modalities as already solved.

If VisCGEC-style shared tasks were to report a teachable-edit record, downstream systems would have a reason to emit these fields; we have no evidence of uptake and claim none.

Two clarifications follow. First, standard M2/ERRANT $F_{0.5}$ is an edit-level metric against reference edits, not a raw string match (Bryant et al., 2017, 2019). Agreement with reference edits does not measure pedagogical appropriateness, explanation faithfulness, proficiency alignment, localization to a multimodal artifact, or learner uptake. Second, text-only explainable GEC already exists. Section 2 states what we add.

2 Related work: what is already specified

Explainable GEC. EXPECT annotates evidence words and grammatical error types, and evaluates whether explanations help second-language learners accept suggestions and understand rules (Fei et al., 2023). EXGEC/EXCGEC extend this for Chinese: edit-wise explanations pair each correction with an error type and free-text description. Their findings show that explaining *after* committing the correction outperforms explaining beforehand, and explicit edit extraction improves explanation faithfulness (Ye et al., 2025). Both systems implement parts of Claims B and C on tokenized text. Neither requires a pointer into a handwritten page, speech span, or slide region; neither scores conservatism as a first-class property; and neither treats calibrated abstention as a required field of one output record. The natural integration is to keep EXCGEC's edit-wise explanation framework and add localization plus abstention.

Arabic provides a complementary, text-based case. The QALB shared task established large-scale correction evaluation for Modern Standard Arabic (Mohit et al., 2014). ARETA performs automatic Arabic error type annotation over a taxonomy adapted from the Arabic Learner Corpus, designed around morphological richness and orthographic ambiguity (Belkebir and Habash, 2021). Nahw evaluates Arabic grammatical error detection, correction, and explanation jointly (Mubarak et al., 2026). These resources establish correction, error typing, and explanation as active Arabic targets. None of them provides pointers into handwritten artifacts, recognition-aware pseudo-error reporting, or the joint record proposed here. Arabic therefore tests whether the teachable-edit specification can hold a shared functional layer fixed while accommodating language-specific morphology, orthography, and script direction.

Proficiency-aligned evaluation. CTSEG evaluates GEC with CEFR-linked grammar-item minimal pairs (Koyama et al., 2025). Current models are much stronger on beginner items than on intermediate and advanced ones. Aggregate correction scores therefore hide instructional mismatch. CTSEG remains text-only and does not evaluate explanations; extending it to multimodal or explanation-aware settings would require additional annotation and design work.

Minimal edits, fluency edits, and LLM instructions. Annotation schemes already split minimal correction from fluency rewriting. JFLEG is fluency-oriented (Napoles et al., 2017); CoNLL-style and many BEA references aim at more local repairs (Ng et al., 2014; Bryant et al., 2019). MultiGEC-2025 separates a minimal-edit track from a fluency track (Masciolini et al., 2025). Recent work analyzes how instruction-following LLMs change behavior across editing

	EXPECT	EXCGEC	JELV	Ours
Edit-wise type	yes	yes	n/a	yes
Free-text expl.	limited	yes	no	yes
Edit-validity test	no	no	yes	yes
Pointer to artifact	no	no	no	yes
Scored abstention	no	no	no	yes
Split eval. targets	no	no	no	yes

Table 1: Closest prior instruments are text-only or validity-only. The specification is the joint record. "Limited" for EXPECT means evidence words plus an error type rather than a natural-language rationale. "Split eval. targets" means separate measurement of four items: correction correctness, feedback quality (pointer, type, conservatism, explanation, abstention), immediate uptake, and learning outcomes; EXPECT already includes correction and learner-facing evaluation, while EXCGEC reports separate edit and explanation measures that map to two of the four.

modes (Türker and Eryigit, 2026). Sequence-to-edit taggers already expose an edit inventory (Omelianchuk et al., 2020). JELV judges whether an individual edit is valid on grammaticality, faithfulness, and fluency, and can reclassify over-corrections that $F_{0.5}$ treats as false positives (Zhan et al., 2026). That is a ready instrument for the conservatism field. The remaining gap is that multimodal learning papers still present a fluent hypothesis as the educational object.

What we claim is new. The proposed specification adds five jointly required properties: (i) localization into the original multimodal artifact; (ii) conservatism as an evaluated property, not only a prompting style; (iii) faithfulness of the explanation to the proposed edit; (iv) calibrated abstention; (v) reporting these fields in one record, so feedback quality can be scored separately from correction correctness and from uptake. Table 1 places those properties against the closest prior systems.

3 Multimodal models are not tutors by default

A tutor locates a problem in the learner's production; chooses a feedback strategy matched to proficiency and goal (Ellis, 2009; Lyster and Ranta, 1997); leaves intact what is already successful; and supports uptake. None of these is implied by image-to-text or speech-to-text generation.

Surveys of multimodal LLMs emphasize architecture, alignment, and benchmarks (Yin et al., 2024). Surveys of MLLMs in education emphasize engagement and content generation (Xing et al., 2024). Work on LLMs for language assessment is more cautious: generation can improve while grading and error-correction validity do not automatically follow (Caines et al., 2023). Unless a paper specifies a learning construct (accuracy on a treatable form, uptake of a feedback move, reduction of a recurring L1-transfer error), it should not claim tutoring capability.

Pretraining, supervised instruction tuning, and preference optimization all change model behavior. None of those stages is a substitute for an explicit instructional objective.

VisCGEC contains 2,451 images of CFL handwriting with expert correction texts (Wang et al., 2025). It is the main public visual CGEC benchmark built on CFL handwriting, which is why we treat it as the place to set a reporting standard. Reported two-stage systems combine OCR or CLIP-based recognition with GECToR-style editors, sequence-to-sequence correctors, or fine-tuned LLMs. OCR followed by fine-tuned Qwen achieves 28.59 character-level correction $F_{0.5}$ and the highest sentence-level correction $F_{0.5}$, 15.85, among the reported baselines. That is a capability result on a hard visual task, not a learning result. A stronger future system that emitted only polished strings would still be a correction engine unless its output implemented a teaching move.

4 Rewrites are not lessons

Written corrective feedback research distinguishes direct correction, indirect marking, and metalinguistic comment (Ellis, 2009; Bitchener and Storch, 2016), and focused versus comprehensive scope (Ferris, 2002; Lee, 2019). Findings on which strategy wins are mixed (Bitchener and Knoch, 2008). The underspecified prior is that the form of the feedback is part of the treatment.

Full-sentence generation obscures what the learner already knows. The convenient object is a fluent hypothesis; the pedagogical cost is that successful interlanguage disappears into the rewrite (Selinker, 1972). Low-proficiency learners need selective feedback: one treatable form marked, the rest left standing so it can be reused (Lee, 2019; Ferris, 2002). A system that only returns a polished sentence cannot offer that view. Edit extraction can be applied after the fact (Bryant et al., 2017); the default object exposed to the learner is still the rewrite.

Over-correction and reference design. Some GEC references are local repairs; JFLEG authorizes fluency rewriting toward native-like text (Napoles et al., 2017; Bryant et al., 2023). LLM GEC often drifts toward the fluency pole unless constrained (Davis et al., 2024; Staruch et al., 2025; Türker and Eryigit, 2026). A model rewarded for matching a fluency reference will change grammatical but non-native wording. Interlanguage is evidence (Selinker, 1972), not noise.

Unfocused feedback. A CFL page may contain a lookalike character, an L1-calqued particle, and a discourse inconsistency. Comprehensive correction of every deviation in one pass is a strategy writing researchers often advise against for lower-proficiency learners (Lee, 2019). Focus is a teaching decision. Standard benchmark outputs and official metrics do not require a system to declare that decision.

Loss of the page. In visual GEC the learner's object is the image. Recognition-first pipelines reduce the page to a token sequence, so layout, visible revisions, or character formation never reach the correction stage. Wang et al. (2025) document recognition error as a major pipeline failure; they do not isolate those visual factors. An edit that cannot be found on the page is not feedback the learner can apply.

The default research target for multimodal learning tools should therefore be conservative, typed, localized edits. Conservative: change only what is erroneous relative to a stated goal. Typed: attach a category a learner or teacher can act on. Localized: point at a span, timestamp, or box on the original production. Direct correction of a treatable form, with the rest left alone, still satisfies this target.

5 Explanation is a first-class output

Ellis treats metalinguistic feedback as a distinct option (Ellis, 2009). EXPECT shows that evidence words plus error type can help learners decide whether to accept a suggestion (Fei et al., 2023). Generated rationales introduce a further failure: models invent grammatical stories that sound expert (Fei et al., 2023; Caines et al., 2023).

Consider the classifier contrast in 一位老师 *yí wèi lǎoshī* versus 一个老师 *yí gè lǎoshī*, both meaning "a teacher" in standard Chinese. Either classifier can count people, but they differ in register: the first is the polite form, while the second is grammatical in many contexts and less formal. A useful explanation names that register choice and points at the classifier slot. A harmful explanation invents a rule such as "一个老师 *yí gè lǎoshī* is impossible, because that classifier cannot be used for people." Faithfulness fails when the explanation describes a different phenomenon than the edit.

Explanation should therefore be constrained: it must describe the edit actually proposed; it must match the intended proficiency level (Koyama et al., 2025); and the system must be allowed to withhold a rule rather than fabricate one.

Field	Example	Evaluation
Pointer	box over the classifier slot on the page	box IoU sweep / span F_1; coarsening rate
Edit	个 → 位 (*context:* 我有一个老师 *I have a teacher* → 我有一位老师 *polite*)	edit P/R; pseudo-error rate
Error type	classifier / register	type accuracy (κ)
Target form	一位老师 *yí wèi lǎoshī*	match to references
Explanation	the polite classifier 位 matches the formal task requirement	see faithfulness row
Faithfulness	covers this edit / related / unrelated	3-way rating; span grounding
Factuality	correct rule / invented / abstained	3-way rating; entailment
Level	elementary band	lexical band; teacher rating
Abstention	no rule if type unknown	coverage-risk; teacher-anchored
Conservatism	no other tokens changed	unnecessary-edit rate; budget

Table 2: Teachable-edit record. The Chinese example is illustrative, not an item from VisCGEC.

6 An output specification and four evaluation targets

Table 2 states a teachable-edit record. The example is schematic: a handwritten CFL sentence intended as "I have a teacher," written with the general classifier where the task asked for the polite form.

The specification is compatible with ERRANT-style types (Bryant et al., 2017), EXPECT evidence (Fei et al., 2023), EXCGEC edit-wise descriptions (Ye et al., 2025), sequence-to-edit tags (Omelianchuk et al., 2020), and CEFR item labels

Field	Arabic instantiation
Pointer	adjective span مجتهد and its final-character boundary
Source	الطالبة مجتهد
Edit	insert feminine suffix: مجتهد → مجتهدة
Surface type	Arabic gender agreement
Functional type	form agreement
Explanation	the predicate adjective agrees with the feminine subject الطالبة
Abstention	withhold if the final handwritten character is uncertain
Conservatism	no other source word changed

Table 3: Arabic instantiation of the teachable-edit record. The example is constructed to illustrate Modern Standard Arabic gender agreement; it is not drawn from a visual Arabic GEC benchmark.

(Koyama et al., 2025). The missing piece is refusing to collapse the record into one fluent hypothesis before scoring.

6.1 Operational measures

The record can be started without new data, but not finished without it: the correction and localization fields can be retrofitted onto existing benchmarks. Conservatism can be evaluated using reference and validity judgments available now; teacher-validated pedagogical conservatism (determining whether the edit aligns with lesson goals for that learner) requires additional annotation. The explanation, level, and abstention fields need a small annotation pass, which Section 8 specifies and does not run.

Conservatism. Report three quantities because they fail differently. Edit rate is descriptive: a high rate on a heavily erroneous page is correct behavior, not overcorrection. Unnecessary-edit rate is the share of edits that neither match a reference nor survive validity checking; only genuine overcorrections count against conservatism. Budget compliance is the share of records respecting a declared feedback budget: the number of edits the system will surface for this learner at this level.

Faithfulness and factuality. Three checks replace proxy overlap with gold rationales, which correlates poorly with human usefulness (Schuff et al., 2022). *Span grounding*: every form the explanation cites must lie inside the pointer span or declared evidence tokens. *Entailment*: the explanation must justify the proposed edit using the source context, target form, and declared task goal. *Counterfactual sensitivity*: changing the target to instantiate a different linguistic phenomenon should change the corresponding explanation.

Abstention. Score abstention as selective prediction: a coverage-risk curve over edit-level confidence, the area under it, and calibration error (Guo et al., 2017). Two reporting points matter pedagogically. *Teacher-anchored* selective risk takes the error indicator from teacher judgment rather than reference match. *Abstention appropriateness* is the share of withheld cases a teacher agrees should have been withheld.

Types and levels. The type field carries two layers because no single scheme is both language-faithful and portable. The surface layer is language-specific and adopts whatever is established (ERRANT for space-delimited orthographies, character-level annotation for Chinese). The functional layer is a small closed set: form agreement, argument structure, selection and collocation, orthography and character formation, cohesion, and pragmatic appropriateness. Only the functional layer is claimed to transfer. Proficiency labels stay inside one framework per item (HSK or CEFR), not forced mappings, and bind the explanation on two computable axes: the share of explanation tokens inside the level's vocabulary, and the count of grammatical terms outside its declared metalanguage. Inter-rater agreement on types is measured with κ (Cohen, 1960).

6.2 Localization under recognition noise

The pointer is the specification's load-bearing multimodal commitment and the field most exposed to recognition error. Where a transcript exists and is aligned to the artifact, report span F_1; where the pointer is a region, report intersection over union against annotated boxes at several thresholds.

Uncertainty has to propagate rather than vanish. Distinguish recognition confidence (the recognizer's confidence over the spanned units) from geometric localization confidence (reliability of the region coordinates themselves). The specification imposes a two-stage gate applied in order: if recognition confidence falls below a declared threshold, the system abstains entirely. Otherwise, if fine-grained geometric localization confidence falls below its threshold, the system coarsens the pointer to the enclosing unit and records that change.

Arabic exposes a localization problem that Chinese hides. Arabic text is stored in Unicode logical order but displayed right to left, while an artifact

pointer is geometric image coordinates. A record must therefore keep text-span order and image-coordinate order distinct, and state the script direction explicitly. For Arabic handwriting a word-level region is often more reliable than independent character boxes.

Recognition failure also manufactures errors the learner did not make. These belong in a distinct pseudo-error rate rather than absorbed into precision. Spoken GEC already treats removal of transcription-induced edits as a precondition for usable feedback (Qian et al., 2025); the visual case has the same structure.

6.3 Retrofitting the specification

Nothing in the record requires new architecture. A team with a published visual CGEC pipeline can emit teachable-edit records by building on existing components: retain recognizer confidences and box coordinates; keep the corrector, but align source and hypothesis afterward to recover an explicit edit set; filter that set with a validity judge; explain each surviving edit; finally, project each edit back onto coordinates and apply the gate. Existing components support several steps, but their integration, multimodal projection, and calibration require additional engineering. The barrier to teachable edits is partly engineering, partly a reporting convention.

7 The same specification on speech

Writing and speech are both classroom artifacts. The specification does not change across modalities; the pointer does.

Speech. A learner says an English past-tense form as a bare stem. A correction engine returns a fluent transcript with the *-ed* supplied and no timestamp. A teachable edit marks the span, types the error as tense, leaves the rest of the utterance, and either names the regular past or abstains when ASR confidence is low. Spoken GEC already treats removal of transcription-induced edits as a precondition for usable feedback, which is the pseudo-error requirement under another name.

Cross-script transfer. The Chinese and Arabic examples share the record structure but not the surface analysis. Chinese permits fairly direct character-box localization; connected Arabic handwriting may require word-level regions and boundary anchors. These differences are not noise for a universal taxonomy to remove. They live in the language-specific surface layer while the record continues to require a pointer, an edit, a type, an explanation or an abstention, a declared level, and a conservatism decision.

Shared failure. In both modalities the attractive demo is end-to-end fluency. In both the missing object is a pointer the learner can use without rereading a polished string.

Four evaluation targets should be reported separately.

Correction correctness. Does the proposed edit match a reference repair? M2/ERRANT $F_{0.5}$ is appropriate for this target. It is not a feedback metric.

Feedback quality. Are the pointer, type, conservatism, explanation, and abstention appropriate for the stated learner and goal? EXPECT's learner study is a text-only instance of this target. Multimodal localization and explanation-to-edit faithfulness still need protocols.

Immediate uptake. Does the learner use the information on a later attempt in the same session (Lyster and Ranta, 1997)?

Learning outcomes. Does the learner retain or transfer the form to new contexts? Immediate uptake is not a proxy for delayed learning or transfer.

We do not claim that $F_{0.5}$ should disappear, or that every deployment must be indirect. Direct correction of a treatable form, localized and explained, satisfies the specification. VisCGEC already annotates bounding boxes, correction texts, and error types, so three fields of Table 2 can be scored on existing data. That is why we frame the record as an extension of current practice rather than a departure from it.

8 What M2/ERRANT $F_{0.5}$ cannot see

M2/ERRANT $F_{0.5}$ rewards agreement with reference edits and weights precision over recall. That design is right for a correction engine. It is the wrong primary metric for a claim about teaching.

The reference defines the lesson. Fluency-oriented datasets such as JFLEG normally use GLEU; they authorize rewording unrelated to any local error. Our argument applies if fluency references are evaluated using edit-based $F_{0.5}$. A system can raise $F_{0.5}$ by matching those rewrites and still hide the learner's interlanguage. Minimal-edit references constrain this less, but they still do

not record whether the edit was the right teaching move.

$F_{0.5}$ is also silent on explanation. An invented rule attached to a correct edit can score perfectly and still be harmful. Faithfulness and factuality are invisible to the metric. So is the pointer into the page: in visual CGEC the learner must find the intervention on an image. A correct string with a wrong box is a failed lesson and a successful $F_{0.5}$ item.

The metric is not misspecified for correction correctness. It cannot see pedagogical appropriateness, explanation quality, proficiency alignment, multimodal localization, or uptake. Papers that report only $F_{0.5}$ and then claim educational use are changing the target after the measurement.

VisCGEC is the main public visual CGEC benchmark built on CFL handwriting. Its published metric is character- and sentence-level $F_{0.5}$, but its annotation already carries boxes, corrections, and types, so localization can be scored without new data. The remaining distance to the full specification is explanation, level, conservatism, and abstention.

A checklist for workshop chairs. A workshop that wants this standard in the next cycle could (i) add feedback-quality and, where possible, uptake columns to the leaderboard without changing $F_{0.5}$ rankings; (ii) encourage system papers to report the record when they claim educational use; and (iii) reserve a short session for systems that score high $F_{0.5}$ but low explanation faithfulness.

A protocol we are not running here. A feasibility study needs only a few dozen VisCGEC-style pages, double-annotated by trained instructors: inter-rater agreement for error type, Krippendorff's α for ordinal faithfulness ratings (Krippendorff, 2018), and mean pairwise IoU for the pointer. We state in advance what would count against us: $\alpha < 0.67$ on faithfulness should be read as evidence that the rubric needs revision, not the annotators. In the proposed annotation pass, instructors rate factuality (correct rule, invented, abstained), level appropriateness (matches declared proficiency band), and abstention appropriateness (teacher agrees withheld cases should have been withheld).

9 Conclusion

Multimodal models can see a learner's page and hear a learner's voice. That is not the same as teaching. The Chinese and Arabic cases show why the record holds its pedagogical function fixed while leaving the linguistic analysis language-specific. For systems claiming to teach through corrective feedback, their output should be a teachable edit: localized into the original artifact, typed, conservative, explained or withheld, and scored as feedback rather than only as reference-matching correction.

10 Limitations

This is a position paper. We present no new experiments, annotation, or evidence that teachable-edit records improve uptake relative to rewrites. The WCF literature is mixed on direct versus indirect and on focused versus comprehensive correction. The running example is visual CGEC for CFL because that setting publishes multimodal GEC numbers; we do not empirically show transfer to speech, slides, or other scripts. The classifier example is constructed, not from VisCGEC. We do not analyze commercial product copy. Teacher-in-the-loop may be unavailable where automation is most tempting. Our Arabic discussion is illustrative and text-grounded: we identify no Arabic visual GEC resource with annotations equivalent to VisCGEC's character boxes, and we evaluate neither handwritten Arabic recognition nor Arabic localization. Connected writing, ligatures, optional diacritics, and variation between Modern Standard Arabic and regional varieties may require revisions to the pointer and type representations.

11 Ethical Considerations

A false metalinguistic explanation is educational harm. Systems offered to learners should abstain when they cannot ground a rule in the proposed edit (Caines et al., 2023; Fei et al., 2023).

Visual recognition and page-based tutors are sensitive to image quality and writing variation. Large gaps appear between OCR and weaker recognizers. Inclusive multimodal learning papers should report failure modes by input quality rather than only mean $F_{0.5}$.

The level tag can be repurposed for placement, streaming, or assessment. We intend it as a property of the feedback, not of the learner.

A homework photograph can identify a hand, a school, and a proficiency trajectory. Localization makes the image necessary; that necessity raises the duty of care. Teacher-in-the-loop review matches current published visual CGEC quality and this risk profile better than fully automatic page correction.

12 Declaration on Generative AI

A generative AI assistant was used for language editing and drafting support based on outlines and content specified by the author. The author developed the paper's claims, proposed record, evaluation framework, and substantive arguments. No data, annotations, experiments, or measurements were produced using generative AI. The author reviewed and verified all AI-assisted material and takes responsibility for the final manuscript.

References

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

John Bitchener and Ute Knoch. 2008. The value of written corrective feedback for migrant and international students. *Language Teaching Research*, 12(3):409–431.

John Bitchener and Neomy Storch. 2016. *Written Corrective Feedback for L2 Development*. Multilingual Matters.

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

Christopher Bryant, 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. MIT Press, Cambridge, MA.

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 Workshop on Empowering Education with LLMs – the Next-Gen Interface and Content Generation*, volume 3487 of *CEUR Workshop Proceedings*, pages 173–197, Tokyo, Japan. CEUR-WS.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

Rod Ellis. 2009. A typology of written corrective feedback types. *ELT Journal*, 63(2):97–107.

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

Dana R. Ferris. 2002. *Treatment of Error in Second Language Student Writing*. University of Michigan Press, Ann Arbor.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th ed. SAGE Publications, Thousand Oaks, CA.

Icy Lee. 2019. Teacher written corrective feedback: Less is more. *Language Teaching*, 52(4):524–536.

Roy Lyster and Leila Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1):37–66.

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.

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

Hamdy Mubarak, Majd Hawasly, and Abubakr Mohamed. 2026. Nahw: A comprehensive benchmark of Arabic grammar understanding, error detection, correction, and explanation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6310–6328, Rabat, Morocco. Association for Computational Linguistics.

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

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.

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

Mengjie Qian, Rao Ma, Stefano Bannò, Mark J. F. Gales, and Kate M. Knill. 2025. End-to-end spoken grammatical error correction. *arXiv preprint arXiv:2506.18532*.

Hendrik Schuff, Heike Adel, Peng Qi, and Ngoc Thang Vu. 2022. Challenges in explanation quality evaluation. *arXiv preprint arXiv:2210.07126*.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Ryszard Staruch, Filip Graliński, 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.

Tolgahan Türker and Gülşen Eryigit. 2026. Instruction-following LLMs for grammatical error correction: Analyzing neutral-anchored instructional sensitivity across editing modes. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 234–247, San Diego, California, USA. Association for Computational Linguistics.

Xiaoman Wang, Dan Yuan, Xin Liu, Yike Zhao, Xiaoxiao Zhang, Xizhi Chen, and Yunshi Lan. 2025. VisCGEC: Benchmarking the visual Chinese grammatical error correction. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5054–5068, Albuquerque, New Mexico. Association for Computational Linguistics.

Weicheng Xing, Tianqing Zhu, Jenny Wang, and Bo Liu. 2024. A survey on MLLMs in education: Application and future directions. *Future Internet*, 16(12):467.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, and Wenhao Jiang. 2025. EXCGEC: A benchmark for edit-wise explainable Chinese grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25678–25686.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Yuhao Zhan, Yuqing Zhang, Jing Yuan, Qixiang Ma, Zhiqi Yang, Yu Gu, Zemin Liu, and Fei Wu. 2026. JELV: A judge of edit-level validity for evaluation and automated reference expansion in grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34611–34619.