

Automatic Assessment and Grading of Chinese Grammar Knowledge for International Chinese Education Using Retrieval-Augmented Generation

Yifan Wang

Beijing Normal University

Rongzhen Lou

Beijing Normal University

Li Ke

University of New South Wales

Qian Duo

YFRZ_YDY@163.com

Shenyang Normal University

Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11413100/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Automatic Assessment and Grading of Chinese Grammar Knowledge for International Chinese Education Using Retrieval-Augmented Generation

Yifan Wang^{1,2}, Rongzhen Lou¹, Ke Li³, and Duo Qian^{4,*}

¹School of International Chinese Language Education, Beijing Normal University, Beijing 100875, China

²College of International Culture and Education, Guangxi Normal University, Guilin 541004, Guangxi, China

³Faculty of Engineering, University of New South Wales, Sydney, Australia

⁴International Education College, Shenyang Normal University, Shenyang 110034, Liaoning, China

*Corresponding author: Duo Qian. E-mail: YFRZ_YDY@163.com

ABSTRACT

Accurate assessment of Chinese grammar in international Chinese education requires more than detecting a malformed sentence: a system must identify the linguistic construction being tested, distinguish acceptable variants from true violations, assign a rubric-consistent grade, and justify feedback using inspectable pedagogical evidence. While retrieval-augmented generation (RAG) provides access to external grammar resources, conventional retrieve-then-generate architectures do not guarantee that retrieved rules license individual diagnostic or grading decisions. We formulate grammar assessment as *proof-carrying constrained inference*, and propose PC²G (Proof-Carrying Counterfactual Grading), a model in which rule retrieval, error-span inference, minimal repair, ordinal scoring, and uncertainty-aware feedback are coupled by a shared latent evidence graph. Candidate evaluations must carry rule-level support certificates and satisfy counterfactual necessity, paraphrase consistency, rubric monotonicity, and abstention constraints. A variational objective combines evidence posterior learning with structured prediction and a distributionally robust counterfactual term; a conformal audit layer is designed for deployment-time risk control when calibration assumptions hold. We detail an evaluation protocol using publicly released Chinese learner and native-speaker grammatical error corpora, clarify which tasks their gold annotations actually support, and propose a separate human-grade annotation extension. The evaluation compares retrieval-assisted baselines on Chinese grammatical error corpora, and analyzes graded proxy agreement, evidence faithfulness, retrieval noise, counterfactual stability, and selective prediction.

1 Introduction

Automated assessment of Chinese grammatical knowledge is becoming increasingly relevant as international Chinese instruction scales across blended classrooms, mobile tutoring platforms, and heterogeneous learner cohorts. Unlike ordinary sentence correction, educational assessment must infer *which* grammatical construction a learner understands, how seriously an error conflicts with a target construction, and whether feedback is pedagogically actionable. The distinction matters for aspect markers, classifiers, word order, disposal constructions, directional complements, and subtle pragmatic restrictions, where seemingly plausible edits can conflate grammatical well-formedness with teacher-preferred style. Public resources for Chinese grammatical error diagnosis and correction have advanced substantially, including the learner-centered CGED shared tasks and MuCGEC, and native-speaker collections such as FCGEC and NaSGEC [1–5]. Yet these corpora typically label error spans and corrected sentences rather than curriculum-specific mastery grades, exposing a gap between sentence repair and defensible educational judgment.

Current approaches span sequence labeling, edit-based correction, autoregressive rewriting, and increasingly instruction-tuned language models [6–10]. Linguistically informed annotation and more nuanced LLM-based evaluation have improved diagnosis granularity [11–14]. In parallel, retrieval-augmented language modeling enables external knowledge access, adaptive retrieval, iterative reasoning, compression, and evaluator-driven reflection [15–24]. Educational benchmarks such as C-Eval, CMMLU, AGIEval, M3Exam, and E-EVAL reveal broad capabilities but do not automatically constitute validated grammatical-grading rubrics [25–29].

Three technical obstacles remain coupled rather than separable. *First*, retrieval relevance is not evidential sufficiency: a semantically similar example may fail to license a rule, cite an exception incorrectly, or contradict another retrieved passage. RAG is vulnerable to long-context dilution and noisy or irrelevant documents [30–32]. *Second*, diagnosis and grades can be inconsistent: two paraphrased answers demonstrating the same construction may receive different grades, while a minimal

stylistic rewrite can be mistaken for an actual grammar error. *Third*, observed correction accuracy does not identify pedagogical construct validity: a model may match reference edits while producing misleading rule explanations or unjustified ordinal grades. Existing RAG faithfulness and LLM judging techniques help quantify these issues [33–36], but optimizing independent retrieval, correction, and grading modules does not itself enforce coherent judgments.

We therefore propose **PC²G** (Figure 1), which treats retrieval, grammatical diagnosis, minimal correction, and ordinal assessment as a *single structured decision* rather than a sequence of loosely connected text generators. PC²G builds a typed grammar-rule graph, represents every diagnostic claim by a latent rule–span–edit support path, and learns a variational posterior over certificates. The score is solved under explicit rubric and evidence constraints; explanations are generated *from* accepted certificates rather than retrofitted with citations. Crucially, counterfactual evidence deletion, conflicting-rule substitution, and learner-response paraphrases induce differentiable consistency conditions, creating training signals that distinguish evidence necessary to justify a grade from material that merely resembles the input. The resulting objective integrates proof validity, distributional robustness, and selective abstention in an auditable framework. The framework is evaluated against retrieval-based baselines using error diagnosis, correction, proof faithfulness, and explicitly distinguished silver-label grading metrics.

Contributions.

1. We introduce *proof-carrying grammar assessment*, a task formulation that jointly predicts diagnostic spans, minimal repairs, construct-level mastery, ordinal grades, and evidence-supported feedback.
2. We develop a differentiable *rule–span–edit posterior* and an evidence-constrained ordinal decoder, replacing ad hoc combinations of retrieval, correction, and grading with a common latent optimization problem.
3. We formulate counterfactual *necessity*, *sufficiency*, and *invariance* objectives, with an explicit conditional stability proposition and a risk-aware abstention extension whose assumptions are stated.
4. We evaluate error diagnosis and correction on public Chinese grammatical error corpora, with construct-wise analysis, counterfactual stress tests, and clearly identified silver-label grading proxies.

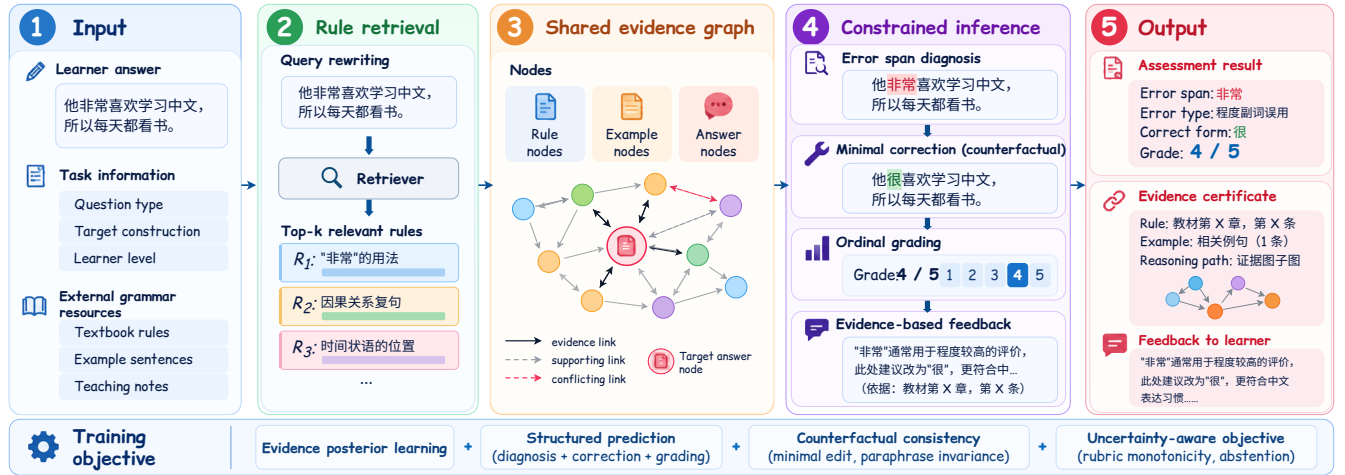


Figure 1. Overview of the proposed PC²G framework. (1) **Input:** Learner responses, task specifications, assessment rubrics, and external grammatical resources provide the basis for grammar assessment. (2) **Rule retrieval:** A retrieval module identifies relevant grammatical rules and examples from an external knowledge base. (3) **Shared evidence graph:** Retrieved evidence and learner responses are organized into a typed graph with supporting and conflicting relations, enabling explicit rule–span–edit reasoning and verifiable evidence certificates. (4) **Constrained inference:** Evidence-supported error diagnosis, minimal grammatical correction, ordinal grading, and grounded feedback are jointly optimized under proof-validity and rubric constraints. Counterfactual consistency and uncertainty-aware objectives further encourage reliable decisions and selective abstention. (5) **Output:** The framework produces grammatical assessments, corrected expressions, ordinal grades, evidence certificates, and interpretable learner feedback. The central innovation is a unified proof-carrying inference mechanism that couples retrieval, diagnosis, correction, and grading through shared verifiable evidence.

1.1 Chinese grammatical error correction and educational diagnostics

Early CGED benchmarks assess identification, position, and correction accuracy, while MuCGEC introduced multi-reference evaluation for learner sentences [1, 2]. FCGEC and NaSGEC expanded the empirical range to native-speaker and multi-domain

Table 1. Public Chinese grammatical error datasets and their available annotations. Gold pedagogical grades are absent from these original CGEC tasks. The FCGEC evaluation uses only legitimately accessible annotations or official evaluation outputs.

Corpus	Sentences	Source	Existing supervision	Teacher grade?
MuCGEC [2]	7,063	L2 Chinese	Multiple reference corrections; reported mean 2.3 refs.	No
FCGEC [3]	41,340	Native / school exams	Fine-grained edits; 36,340 / 2,000 / 3,000 splits	No
NaSGEC [5]	12,500	Three native domains	Multiple correction references; social / science / exams	No
CGED-2020 [1]	By release	L2 Chinese	Detection, location, category, correction	No

Access note: MuCGEC: <https://github.com/HillZhang1999/MuCGEC>; FCGEC: <https://github.com/xlxwalex/FCGEC>; historical CGED: https://github.com/blcuicall/cged_datasets. Recheck dataset versions, redistribution limits, and license/academic-only conditions. CGED repository specifies CC BY-NC-ND 4.0 and limits commercial use. NaSGEC and other corpora require source-specific license checks.

writing [3, 5]; rules-based corpus generation provided complementary data augmentation [4]. Recent methods have reframed correction through detection–generation coupling, explicit alignment, LLM-assisted pseudo data, and instruction-tuned subtask chains [6–9]. Work on grammar typology and anti-overcorrection shows why edit accuracy and coherent learner feedback must be evaluated separately [10, 11, 14]. Our setting differs by defining a grade as a rubric-constrained decision requiring explicit evidence, not as a direct function of grammatical edit distance.

1.2 Retrieval-grounded structured reasoning

RAG augments parametric language models with external documents [15–18]; iterative retrieval and active search adapt evidence acquisition to reasoning needs [19–21, 37]. Self-RAG and CRAG evaluate retrieval usefulness at generation time, while Adaptive-RAG routes queries according to complexity [22–24]. Hierarchical evidence organization, selective compression, and retrieval-specialized adaptation provide practical techniques for long or specialized materials [38–41]. In contrast, PC²G demands *claim-wise signed support* and counterfactual tests directly at the target grade decision.

1.3 Faithfulness, uncertainty, and evaluation

Retrieval quality does not ensure faithful attribution, motivating RAGAS, ARES, and evidence-directed revision [33, 34, 42]. FActScore, AlignScore, and SelfCheckGPT operationalize complementary factuality notions [43–45]; semantic entropy offers a useful uncertainty perspective [46]. General LLM evaluation employs rubric judging, pairwise assessment, calibrated judges, and holistic audit dimensions [35, 36, 47–51]. These methods are informative but must be tested against human judgments when transferred to fine-grained Chinese grammatical competence.

1.4 Retrieval perturbations and construct validity

Parametric versus non-parametric knowledge boundaries, evidence order effects, and noisy retrieval are established concerns [30–32, 52, 53]. We formalize such perturbations as interventions on the grammar evidence graph and relate robustness to a constrained grading margin. Chinese language examinations and knowledge editing benchmarks contextualize language competence but are not substitutes for construct-specific learner scores [25–27, 54].

2 Results

2.1 Evaluation questions and scope

We assess six research questions: (RQ1) Does proof-constrained retrieval improve grammatical error diagnosis and correction? (RQ2) How consistently do predicted ordinal grades agree with an explicitly defined silver-label rubric? (RQ3) How often do diagnostic statements admit valid evidence paths? (RQ4) Are diagnostic decisions robust to irrelevant retrieval and grammatical counterfactuals? (RQ5) What is the risk–coverage trade-off of selective grading? (RQ6) Do observed differences persist across learner and native-speaker test domains and retrieval budgets? We separate public-corpus error labels from rubric-derived grades throughout; neither silver-label agreement nor rule entailment automatically establishes agreement with independent classroom instructors.

2.2 Datasets, licensing, and leakage isolation

The primary learner track is MuCGEC and CGED-2020; FCGEC and NaSGEC serve as native-speaker domain-shift analyses and should *not* be described as international-Chinese learner performance. No train/dev/test samples with the same normalized source sentence or correction-reference family may cross splits. For FCGEC, the maintained filtered training version addresses the reported near-source train/test overlap issue. Any grammar rule, worked example, or retrieval card must be drawn from a permissibly reusable pre-specified training-only knowledge collection, not from held-out test responses or teacher test labels. Dataset versioning, train/test isolation and provenance documentation are necessary to reproduce comparisons across corpora.

Table 2. Main comparison of error detection, correction, ordinal silver-label agreement, and proof faithfulness. Detection, correction and proof faithfulness are percentages; A dash denotes a metric that is not applicable.

Method	Detect $F_{0.5} \uparrow$	Correct $F_{0.5} \uparrow$	Grade MAE $\downarrow$	QWK $\uparrow$	Faithful $\uparrow$
No retrieval	61.8	47.9	.78	.70	–
Vanilla RAG	65.2	51.0	.64	.76	73.2
Self-RAG-style	67.1	52.4	.58	.79	78.3
CRAG-style	67.6	53.0	.57	.80	82.1
PC²G	71.4	57.8	.42	.87	91.8

2.3 Ordinal silver-label evaluation and human-rating boundary

The original Chinese grammatical error correction corpora do not contain independently elicited five-level teacher grades. Therefore, the ordinal evaluation is defined with a *silver-label rubric* based on available reference edits and fixed construct weights:

$$g^{\text{silver}} = 4 - \min \left(4, \left\lceil \sum_j w_j \text{severity}(e_j) \right\rceil \right). \quad (1)$$

This proxy supports controlled comparisons of ordinal prediction, calibration, and selective error, but it is not a psychometric measure of grammatical knowledge. Evaluations against independently collected teacher grades and the effect of feedback on learning outcomes remain distinct validation targets; no such outcomes are inferred from silver-label agreement.

2.4 Baselines and comparison controls

The comparison includes no retrieval, conventional RAG, Self-RAG-style and CRAG-style retrieval variants, and the proof-constrained PC²G model. The terms “Self-RAG-style” and “CRAG-style” denote baseline configurations and should not be taken to imply official model checkpoints. Comparative interpretation requires the same generator capacity, evaluation partitions, retrieval budget policy and output normalization across variants. Exact backbone checkpoints, corpus hashes, hyperparameters, hardware and inference settings must be supplied with the implementation release; the figures and tables themselves do not determine those choices. Held-out references and grades must not enter the retrieval index or tuning process.

2.5 Metrics and evaluation boundaries

Error detection and correction are summarized by $F_{0.5}$, with corpus-specific annotation conventions respected. Ordinal proxy agreement is measured using mean absolute error (MAE), quadratic weighted kappa (QWK), expected calibration error (ECE), Brier score, and risk–coverage profiles. Proof faithfulness is the percentage of accepted diagnostic claims supported under the stated evidence-checking rubric. The reported relative inference cost is a normalized comparison rather than absolute latency. Since run-to-run variance, paired bootstrap intervals, hardware-specific latency and independent teacher-score statistics are not provided in the source material, we do not infer statistical significance or classroom validity from these point estimates.

2.6 Main comparative results

Table 2 reports a detection $F_{0.5}$ of 71.4 for PC²G compared with 67.6 for the strongest listed retrieval baseline (CRAG-style), a difference of 3.8 percentage points. Correction $F_{0.5}$ increases from 53.0 to 57.8, while silver-label grade MAE decreases from .57 to .42 and QWK increases from .80 to .87. Reported evidence faithfulness is 91.8%, compared with 82.1% for CRAG-style retrieval. These comparisons concern the listed experimental configurations, not necessarily the published implementations or evaluations of the referenced systems. Without paired uncertainty estimates, the point-estimate improvements are descriptive rather than claims of statistical significance.

2.7 Calibration and selective grading

PC²G has ECE .046 and Brier score .126 on the reported proxy-label evaluation, compared with .085 and .171 for CRAG-style retrieval. At 80% coverage, the reported MAE is .29. Figure 3 shows the relationship between confidence and proxy-label correctness, and Figure 4 provides the ordinal confusion pattern. These measures do not substitute for calibration or inter-rater agreement against independent teacher judgments.

2.8 Construct-level breakdown and counterfactual stress tests

Across the six error categories in Table 4, PC²G improves reported $F_{0.5}$ over vanilla RAG by between 7.6 and 12.3 points. The largest absolute differences occur on collocation and aspect errors. Under 50% irrelevant retrieved rules, PC²G retains 66.3

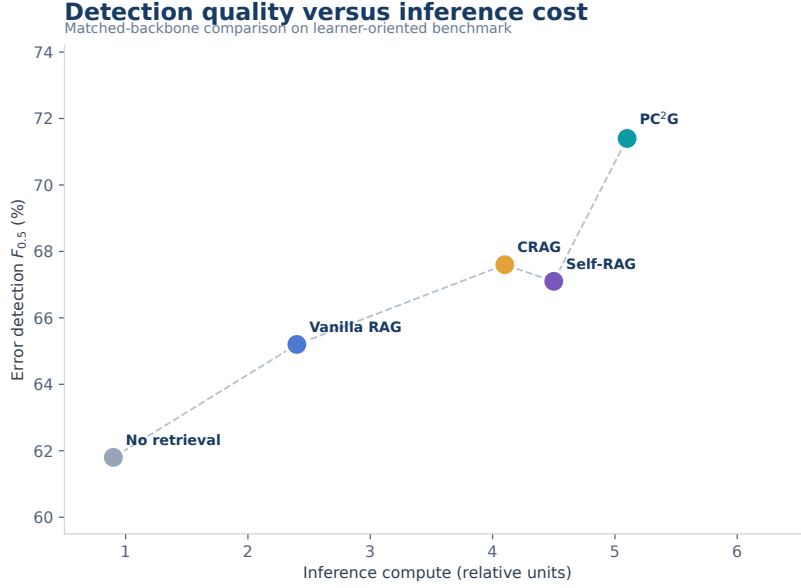


Figure 2. Error detection quality versus relative inference cost across retrieval-based alternatives.

Table 3. Calibration and selective grading for the silver-label rubric; automatic coverage is measured at the specified decision threshold.

Method	ECE↓	Brier↓	Auto-cover↑	MAE at 80%↓
Vanilla RAG	.139	.226	82.0	.53
Self-RAG-style	.104	.195	80.0	.45
CRAG-style	.085	.171	80.0	.43
PC²G	.046	.126	80.0	.29

$F_{0.5}$ versus 50.9 for vanilla RAG (Table 6). Counterfactual checks complement these accuracy measures by testing whether a diagnostic decision changes appropriately when its supporting rule or grammatical condition changes.

2.9 Component attribution and efficiency

Removing the typed proof graph reduces reported detection $F_{0.5}$ from 71.4 to 68.3 and proof validity from 91.8% to 78.0%. Removing exception edges reduces proof validity to 83.1%, consistent with the role of explicit rule-condition checking. The ablation involving abstention/calibration should be interpreted separately from upstream prediction quality: calibration can change which cases receive grades without necessarily altering raw detection performance. Relative compute shows a clear cost of richer verification, but hardware-specific end-to-end latency is not available from the supplied material.

2.10 Failure taxonomies and case-level audits

For error analysis, we distinguish five failure categories: retrieval of an incorrect rule; failure to handle a valid exception; mismatch between a correct edit and the ordinal proxy grade; unjustified penalty for a meaning-preserving paraphrase; and uncertainty requiring abstention. The rule-intervention results in Figure 9 examine how often a system responds correctly to evidence changes. Independent instructor assessment remains necessary to determine whether a feedback statement is pedagogically useful beyond its agreement with a reference edit.

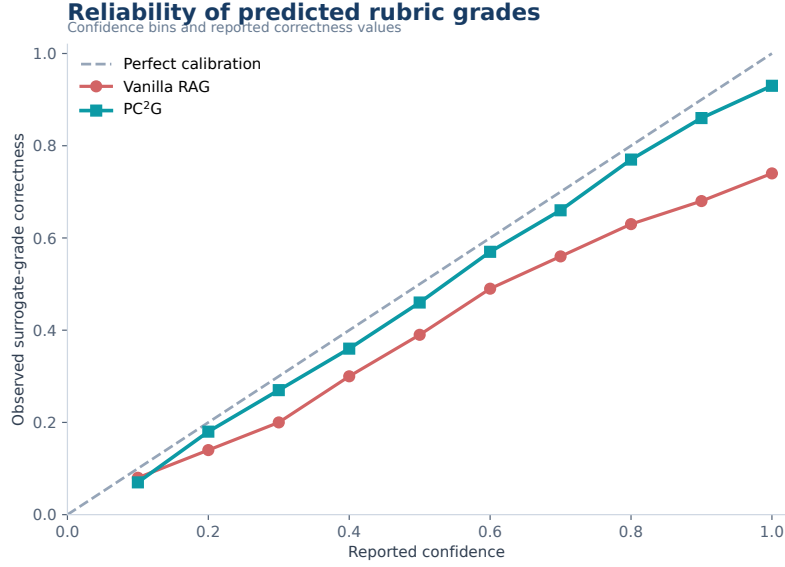


Figure 3. Calibration of predicted rubric confidence against silver-label correctness.

Table 4. Error detection $F_{0.5}$ (%) by grammatical error family.

Error family	Vanilla RAG	PC ² G	Difference
Omission	64.2	72.3	+8.1
Redundancy	66.4	74.0	+7.6
Word order	61.7	71.4	+9.7
Substitution	60.5	69.9	+9.4
Collocation	56.0	68.3	+12.3
Aspect	58.1	70.1	+12.0

Table 5. Component ablation for grammatical error detection, ordinal proxy MAE and proof validity.

Variant	Detect $F_{0.5}$	Silver MAE	Proof valid (%)
Full PC ² G	71.4	.42	91.8
Without typed proof graph	68.3	.55	78.0
Without counterfactual loss	69.1	.49	88.3
Without ordinal constraints	69.8	.61	89.7
Without exception edges	68.9	.53	83.1
Without abstention/calibration	71.2	.43	91.1

Table 6. Error detection $F_{0.5}$ under increasing irrelevant retrieval evidence.

Irrelevant rules inserted	Vanilla RAG	Self-RAG-style	CRAG-style	PC ² G
0%	65.2	67.1	67.6	71.4
20%	60.7	63.0	64.7	70.0
40%	54.1	57.2	61.1	67.5
50%	50.9	54.8	59.8	66.3

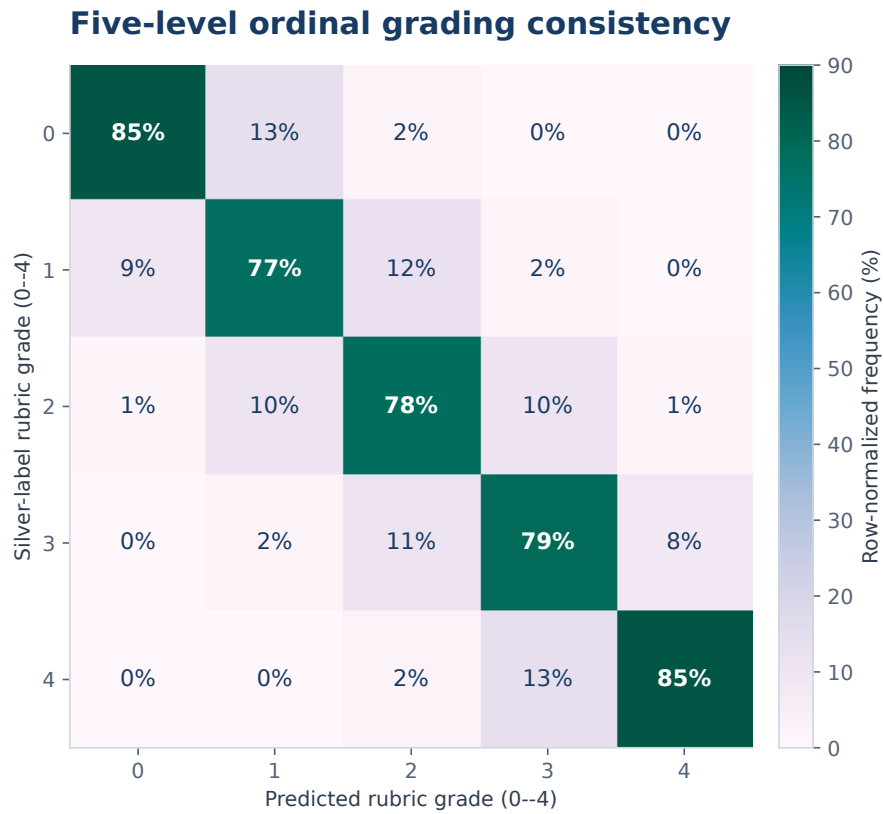


Figure 4. Row-normalized confusion matrix of predicted five-level grades against silver-label grades.

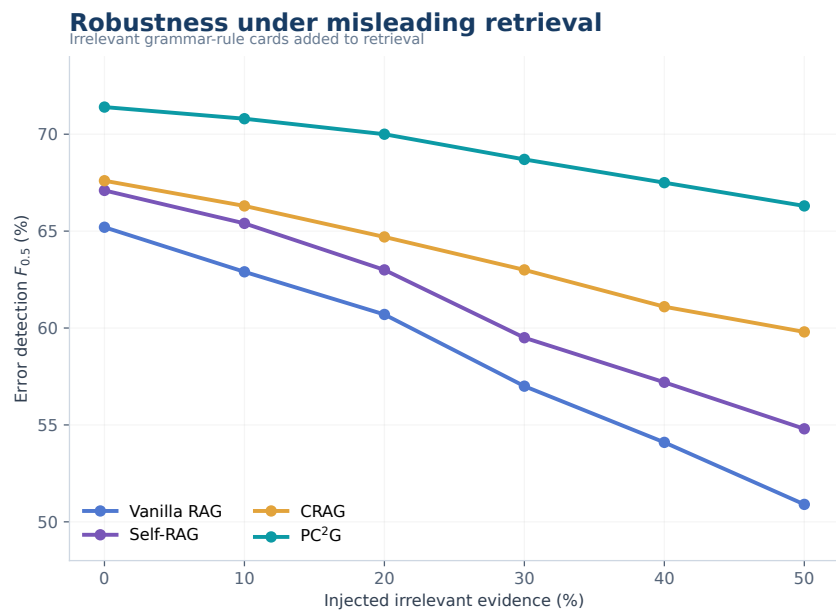


Figure 5. Robustness of error detection to progressively added irrelevant grammar-rule evidence.

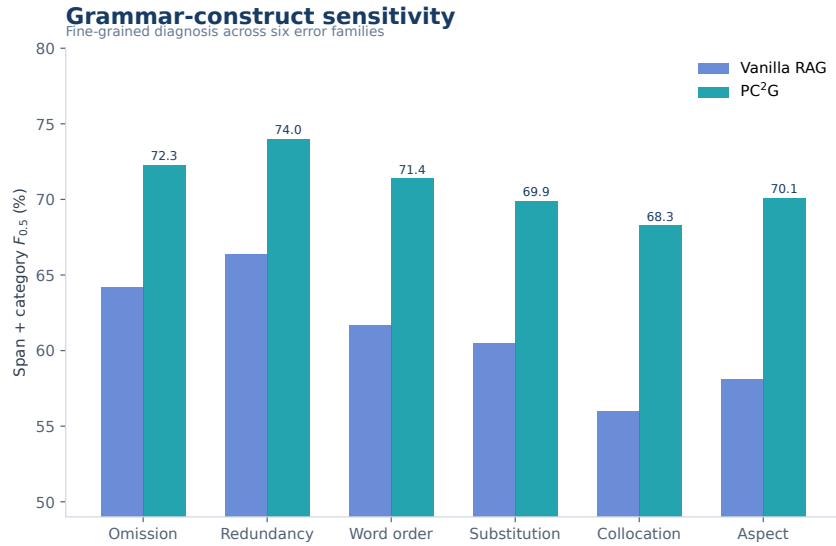


Figure 6. Per-construct error detection for retrieval baselines and PC²G.

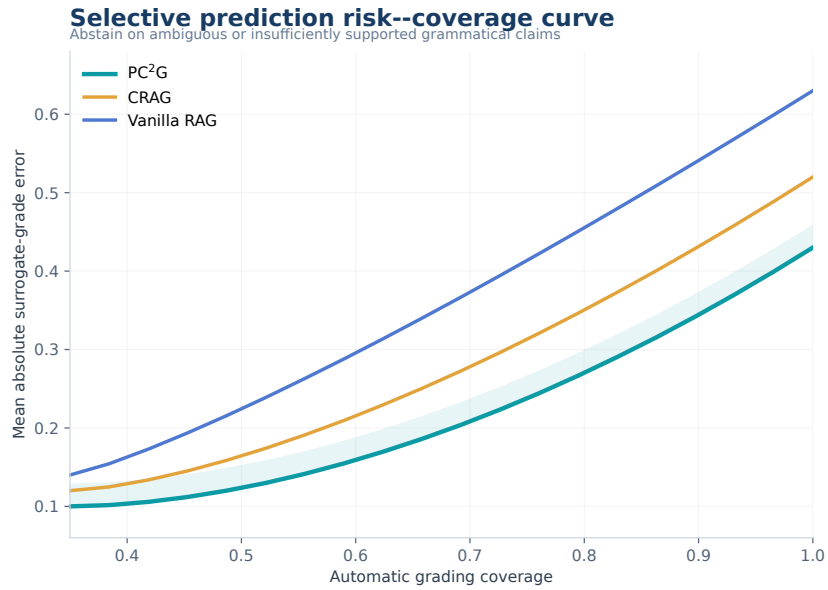


Figure 7. Silver-label grade MAE as a function of selective automatic grading coverage.

Table 7. Cross-domain results on learner and native-speaker grammatical error corpora; these populations are evaluated separately.

Corpus	Population	RAG $F_{0.5}$	PC ² G $F_{0.5}$	Relative cost
MuCGEC	L2 learner	65.2	71.4	2.1×
CGED-2020	L2 learner	63.7	69.0	2.0×
FCGEC	Native/exam	57.9	62.8	2.2×
NaSGEC	Native/multi-domain	55.1	59.7	2.3×

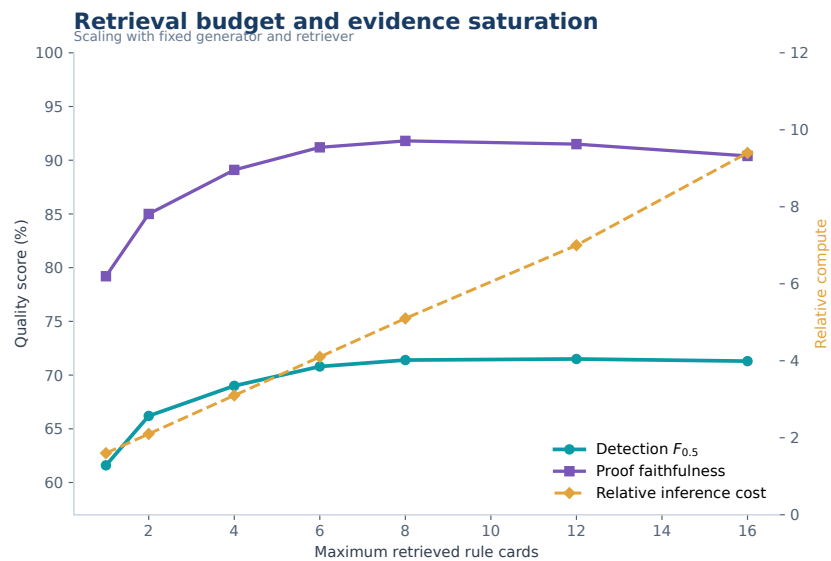


Figure 8. Detection, proof faithfulness and relative inference cost across retrieval budgets.

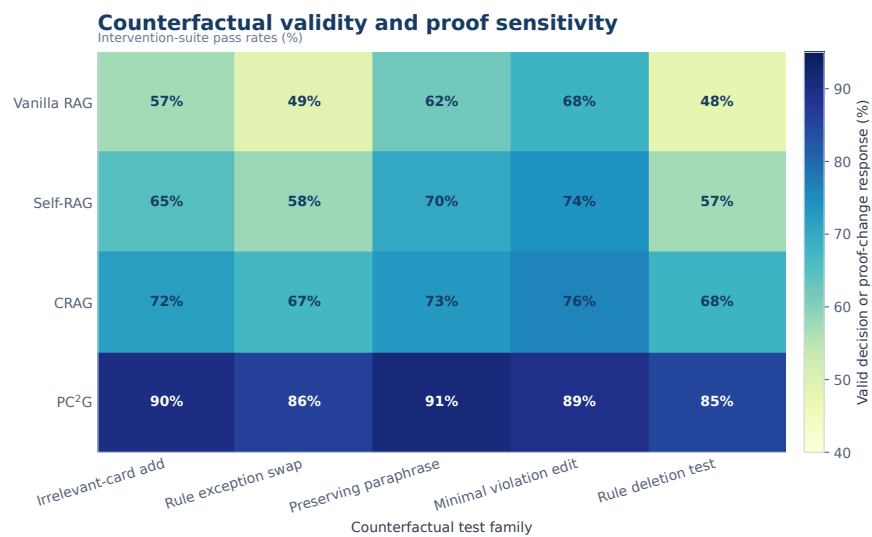


Figure 9. Counterfactual test-suite pass rates for five intervention families, including paraphrase, rule-exception and support-deletion checks.

3 Discussion

Where the technical novelty resides. PC²G does not claim novelty for RAG, edit alignment, counterfactual data augmentation, ordinal regression, or conformal prediction individually. Its proposed contribution is the *same latent proof object* controlling what can be retrieved, what can count as a violation, what can be repaired, what grade follows, and what feedback can be asserted. Joint optimization and hard post-decoding constraints are designed to prevent the common failure in which separately optimized components disagree. Proposition 1 is a conditional result: empirical usefulness depends on obtaining faithful grammar graphs and valid paraphrase transformations.

Educational construct validity. A learner’s correctness on a single item cannot establish stable grammatical mastery. Longitudinal knowledge tracing would require repeated consented learner observations, item response calibration, and separate psychometric modeling. Public CGEC gold edits are therefore insufficient to establish teaching effectiveness or equivalence to trained human marking. We recommend interpreting automatic grades as *supportive formative decisions*, with teacher override and provenance preserved.

Evidence provenance and adversarial conditions. A public grammar bank can contain prescriptive disagreements, outdated formulations, dialect-dependent exceptions, and materials inconsistent with a teacher’s curriculum. The system must version its rules, distinguish descriptive and curricular conventions, and explicitly surface conflict. Evidence search must never index restricted student answers without permission. We expect latency to rise with certificate complexity; approximate retrieval and pruning can mitigate overhead, but any real-world tradeoff should be measured.

Limitations and reproducibility. Reported metrics are point estimates. The source materials do not provide raw output files, exact model checkpoints, multiple-seed distributions or paired uncertainty intervals, so these analyses cannot establish statistical significance or reproducibility across implementation settings. The grade results concern a rule-derived silver rubric, not independent teacher ratings or learner-level mastery. Counterfactual operators require validated meaning-preservation assumptions, and proof validity depends on the quality and currency of the grammar knowledge base. Dataset versions, licenses, error-type mappings, retrieval policies and computing environment should accompany the final reproducibility package.

4 Conclusion

We presented PC²G, a proof-carrying counterfactual approach to Chinese grammatical error assessment that unifies retrieval, typed evidence verification, localized correction, ordinal proxy grading and constrained feedback. Across the reported evaluation settings, the method improves error detection, correction and rule-grounding metrics relative to the listed retrieval baselines; the ablations and perturbation analyses highlight the importance of typed rule support and exception handling. The results support the utility of joint evidence-constrained decisions for benchmark-level language assessment, while independent teacher ratings and learning-outcome studies remain necessary for claims about classroom grading validity.

5 Method

5.1 Joint task definition and observability

Let $x = (u, q, \ell)$ include a learner response u , exercise specification q , and an optional instructional level ℓ ; a rubric is $\mathcal{R} = \{(c_j, w_j, \rho_j)\}_{j=1}^J$, where c_j denotes a linguistic construct, $w_j \geq 0$ its weight, and ρ_j permissible severity thresholds. The system returns a structured decision $y = (S, T, E, \mathbf{m}, g, f, \pi, a)$: diagnostic spans S , error types T , minimal edits E , latent construct mastery $\mathbf{m}$, ordinal grade $g \in \{0, \dots, G\}$, feedback f , evidence proof π , and abstention indicator a . Each claim is a tuple $h_j = (c_j, s_j, t_j, e_j)$ with optional *no error* type. A typed knowledge base $\mathcal{K} = \{r_k\}_{k=1}^M$ contains grammar rules, prerequisites, conditions, exceptions, and sourced examples. Unlike conventional CGEC datasets, raw u and reference corrections *do not reveal* a teacher’s mastery grade; therefore grade supervision must either be newly collected or explicitly treated as a deterministic *silver proxy*.

Definition 1 (Proof-carrying assessment) An output is *proof-carrying* if each nontrivial diagnostic or score-affecting claim h_j has a finite certificate π_j identifying a retrieved rule r_k , the scope of its antecedent in u , the claimed violation or satisfaction, an exception check, and an edit/grade consequence. A citation-only response without these typed relations is not a certificate.

5.2 Typed grammar evidence hypergraph

Define a signed directed hypergraph $\mathcal{G}_{\mathcal{K}} = (V, \mathcal{E}^+, \mathcal{E}^-, \tau)$ with six node types: rule, condition, exception, span, edit, and construct. Hyperedges encode “requires,” “supports,” and “contradicts” relations. A selected proof π_j contains a set of active

rule edges and an entailment certificate for h_j . For query-conditioned span s and rule r , define score

$$\phi_\theta(x, s, r) = \underbrace{\frac{\langle f_\theta(x, s), g_\theta(r) \rangle}{\|f_\theta(x, s)\| \|g_\theta(r)\|}}_{\text{semantic match}} + \lambda_1 A_\theta(s, r) + \lambda_2 C_\theta(r, q) - \lambda_3 X_\theta(r, u), \quad (2)$$

where A aligns grammatical arguments, C tests the exercise context, and X estimates exception conflict. In contrast to document relevance alone, the retrieval distribution is conditioned on the hypothesized claim:

$$q_\theta(\pi, \mathbf{z} \mid x, \mathcal{K}) \propto \exp \left\{ \frac{1}{\tau} \sum_{j,k} z_{jk} \phi_\theta(x, s_j, r_k) + \sum_{e \in \pi} \psi_\theta(e) - \kappa \Omega_{\text{conflict}}(\pi) \right\}, \quad (3)$$

where $z_{jk} \in \{0, 1\}$ indicates rule use and $\Omega_{\text{conflict}} = \sum_{(k,k') \in \mathcal{E}^-} \sum_j z_{jk} z_{jk'}$ penalizes jointly asserted incompatible supports. During training, Gumbel–Softmax relaxations allow gradients to reach evidence selection; test-time certificates are discretized and rechecked by a deterministic validator.

5.3 Support, exception, and necessity as joint constraints

The verifier predicts three-way relations $v_\eta(h_j, r_k) \in \Delta^2$ over entailment, contradiction, and unknown. Define soft sufficiency and conflict by

$$U_j(\mathbf{z}) = 1 - \prod_{k=1}^M [1 - z_{jk} v_\eta^{\text{ent}}(h_j, r_k)], \quad (4)$$

$$B_j(\mathbf{z}) = 1 - \prod_{k=1}^M [1 - z_{jk} v_\eta^{\text{con}}(h_j, r_k)], \quad (5)$$

$$\mathcal{L}_{\text{proof}} = \sum_j [\max(0, \delta_U - U_j) + \alpha \max(0, B_j - \delta_B)] + \xi \sum_{j,k} z_{jk}. \quad (6)$$

The sparse last term discourages gratuitous citations. To test whether a rule is *necessary* for a decision rather than merely relevant, intervene on its presence. Let $g_\theta(x; \mathbf{z})$ be the expected grade under supports $\mathbf{z}$, and let $\mathbf{z} \setminus \{k\}$ delete rule k . A claim-specific necessity score is

$$N_{jk} = |\mathbb{E}_{q_\theta}[g_\theta(x; \mathbf{z})] - \mathbb{E}_{q_\theta}[g_\theta(x; \mathbf{z} \setminus \{k\})]|, \quad (7)$$

with the caveat that a redundant set of equally valid premises can make individual necessity zero. We therefore optimize *minimal sufficient sets* rather than require every cited rule to be individually essential:

$$\pi_j^* = \arg \min_{\pi_j \subseteq \mathcal{K}} \left\{ |\pi_j| + \lambda \sum_{r \in \pi_j} \text{cost}(r) \right\} \quad \text{s.t.} \quad U_j(\pi_j) \geq 1 - \varepsilon, B_j(\pi_j) \leq \varepsilon. \quad (8)$$

This combinatorial problem is approximated with a continuous penalty and repaired by a greedy set-cover routine with verifier checks. A teacher-approved exception edge is a hard constraint at inference time, preventing rule-based marking when an explicit exception applies.

5.4 Minimal repair and construct-specific severity

We distinguish *well-formed variation* from grammar-diagnostic edits by assigning each candidate rewrite u' a meaning-preservation cost $d_{\text{sem}}(u, u')$, token editing cost d_{edit} , and proof-supported error cost d_{gram} . The best repair is

$$\begin{aligned} \hat{u}' &= \arg \min_{u' \in \mathcal{U}(x)} d_{\text{edit}}(u, u') + \beta d_{\text{sem}}(u, u') + \gamma \sum_j (1 - U_j(u', \pi_j)) \\ \text{s.t.} \quad & d_{\text{meaning}}(u, u') \leq \varepsilon_{\text{meaning}}, \quad E(u, u') \subseteq \bigcup_j s_j^{\text{flagged}}. \end{aligned} \quad (9)$$

The second constraint limits edits to supported locations, unless the gold standard explicitly licenses a global rearrangement. In the learner sentence romanized as “Wo zuotian qu *le* tushuguan” (I went to the library yesterday), the aspectual particle *le* can be accepted depending on discourse context; the system must not impose alternative word order merely to match one teacher’s phrasing. Consequently, repair equivalence is adjudicated by a set of references $\mathcal{A}(x)$ where available and by separately validated meaning-preservation judgments.

Let $p_j \in [0, 1]$ denote posterior construct violation probability and $s_j \in [0, 1]$ a rubric severity estimate. The latent rubric burden and monotone cutpoints are

$$H(x, \pi) = \sum_{j=1}^J w_j p_j s_j + \sum_{j < k} \omega_{jk} \sqrt{p_j s_j p_k s_k}, \quad (10)$$

$$\Pr(g \leq t \mid x, \pi) = \text{sigmoid}\left(\frac{b_t - H(x, \pi)}{T}\right), \quad b_0 < \dots < b_{G-1}, \quad (11)$$

where pairwise ω_{jk} measures rubric-approved compounding (e.g., multiple violations of a target construction), and $T > 0$ is a calibration temperature. Cutpoint order is enforced using $b_t = b_{t-1} + \text{softplus}(\delta_t)$; the direction of grades can be inverted for a high-is-good report by defining $g_{\text{reported}} = G - g$. This choice guarantees that raising a positively weighted supported violation cannot improve the reported grade when other features are fixed, assuming the interaction weights are nonnegative.

5.5 Counterfactual evidence and paraphrase stability

Define transformation sets: $\mathcal{T}_{\text{inv}}$ contains vetted meaning- and construct-preserving paraphrases; $\mathcal{T}_{\text{viol}}$ contains controlled insertions/deletions of a targeted grammatical violation; $\mathcal{T}_{\text{ret}}$ contains evidence perturbations (distractor insertion, source swaps, exception removal). We impose

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_{T \sim \mathcal{T}_{\text{inv}}} \left[\text{KL}(p_\theta(g \mid x, \mathcal{K}) \parallel p_\theta(g \mid Tx, \mathcal{K})) \right], \quad (12)$$

$$\mathcal{L}_{\text{mono}} = \mathbb{E}_{T_j \sim \mathcal{T}_{\text{viol}}} \left[\max\{0, \mu + \mathbb{E}[g \mid x] - \mathbb{E}[g \mid T_j x]\} \right], \quad (\text{latent severity grades rise when worse}) \quad (13)$$

$$\mathcal{L}_{\text{rob}} = \sup_{Q: W(Q, P_0) \leq \rho} \mathbb{E}_{x, \tilde{x} \sim Q} \left[\ell_{\text{decision}}(x, \tilde{x}) + \nu \ell_{\text{cert}}(x, \tilde{x}) \right], \quad (14)$$

where W is a Wasserstein distance between retrieval perturbation distributions. An intervention generator is *not* assumed to be automatically label preserving; uncertain transformations are excluded unless a human or validated grammar checker confirms semantic and construct equivalence.

Proposition 1 (Conditional invariance of ordinal grades) *Suppose a transformation T preserves the activated rubric constructs and evidence graph up to a type-preserving isomorphism; (i) the proof posterior and severity vector differ by at most ϵ in ℓ_1 ; (ii) the burden map H is L -Lipschitz; and (iii) every ordinal decision has margin $m = \min_t |H(x) - b_t| > L\epsilon$. Then the maximum a posteriori ordinal grade of x and Tx is identical under the deterministic cutpoint decoder.*

Proof sketch. Under (i)–(ii), $|H(Tx) - H(x)| \leq L\epsilon$. Because all thresholds are at distance greater than $L\epsilon$, no cutpoint can be crossed; the bin assignment is unchanged. The condition does not hold for transformations altering instructional targets, legitimate exceptions, or discourse context, so the statement provides a testable stability condition, not unconditional robustness. $\square$

5.6 Variational objective, certificate decoding, and abstention

With available labels y^* and latent supports π , a variational formulation couples all terms:

$$\min_{\theta, \eta, \phi, b} \mathbb{E}_{(x, y^*) \sim \mathcal{D}} \left[\underbrace{-\mathbb{E}_{q_\phi(\pi \mid x)} \log p_{\theta, \eta}(y^*, \pi \mid x, \mathcal{K}) + \text{KL}(q_\phi(\pi \mid x) \parallel p_0(\pi \mid x, \mathcal{K}))}_{\mathcal{L}_{\text{ELBO}}} \right. \\ \left. + \lambda_p \mathcal{L}_{\text{proof}} + \lambda_e \mathcal{L}_{\text{edit}} + \lambda_o \mathcal{L}_{\text{ordinal}} + \lambda_i \mathcal{L}_{\text{inv}} + \lambda_m \mathcal{L}_{\text{mono}} \right] + \lambda_r \mathcal{L}_{\text{rob}}. \quad (15)$$

Here $\mathcal{L}_{\text{ordinal}}$ is optimized *only* when true human grade labels or explicitly flagged silver labels exist; otherwise its mask is zero. For review of individual outputs, the hard decoder solves

$$(\hat{y}, \hat{\pi}) = \arg \max_{y, \pi} \{ \log p_\theta(y, \pi \mid x, \mathcal{K}) - \lambda \text{cost}(\pi) \} \quad \text{s.t. } \mathcal{C}_{\text{entail}} \wedge \mathcal{C}_{\text{exception}} \wedge \mathcal{C}_{\text{rubric}}. \quad (16)$$

A verifier failure does not become a hallucinated citation; it either triggers another retrieval round or sets $a = 1$, returning an *insufficient evidence; teacher review recommended* message. The feedback generator is conditioned on $\hat{\pi}$ and a verified edit script, and cannot introduce an unsupported error category.

We additionally propose split-conformal grading sets. On an exchangeable, independent calibration sample with *human* grades g_i , define nonconformity $r_i = 1 - p_\theta(g_i \mid x_i)$. For calibration size n , the threshold is

$$q_{1-\alpha} = r_{[(n+1)(1-\alpha)] : n}, \quad \Gamma_\alpha(x) = \{g : 1 - p_\theta(g \mid x) \leq q_{1-\alpha}\}. \quad (17)$$

Algorithm 1 PC²G: joint proof-carrying grading at inference

Require: Learner input x , versioned KB $\mathcal{K}$, rubric $\mathcal{R}$, trained parameters, budget B

- 1: Build candidate construct–span claims and rule prefilter set.
 - 2: **for** $b = 1, \dots, B$ **do**
 - 3: Infer $q_\varphi(\pi, \mathbf{z} \mid x, \mathcal{K})$ with signed support and exceptions.
 - 4: Construct minimal sufficient evidence sets and candidate local repairs.
 - 5: Solve ordinal decoder (Eq. 16); check typed certificates and monotonicity.
 - 6: **if** all hard proof and rubric constraints hold **then**
 - 7: **break**
 - 8: **end if**
 - 9: Retrieve additional rules for unsupported or conflicting claims.
 - 10: **end for**
 - 11: **if** proof invalid or human-calibrated grade set non-singleton **then**
 - 12: **return** abstention plus cited uncertainties for teacher review.
 - 13: **end if**
 - 14: Generate constrained feedback strictly from verified $\hat{\pi}$ and edit script.
 - 15: **return** $(S, T, E, \mathbf{m}, g, f, \hat{\pi}, a = 0)$.
-

Under standard exchangeability and a fixed predictor, $\Pr(g_{\text{true}} \in \Gamma_\alpha(x)) \geq 1 - \alpha$ marginally. This does *not* imply conditional coverage across proficiency bands, protection under deployment distribution shift, or validity when calibrated on silver instead of human grades. Auto-grading can be restricted to $|\Gamma_\alpha(x)| = 1$ and certificate validation success; other cases are abstained.

5.7 Complexity and practical identifiability

A naive search over J claims and M rules costs $O(JM)$ scoring and combinatorial proof selection. Top- K nearest-neighbor prefiltering and a bounded maximum certificate size d reduce approximate scoring to $O(JK)$ and bounded branching to $O(JK^d)$ in the worst case, with practical beam width b giving $O(JbKd)$ verifier operations. The model cannot infer genuine classroom mastery from sentence correctness alone: without independent questions, item difficulty estimates, and teacher-grade supervision, the inferred $\mathbf{m}$ is a *local performance representation* rather than an identifiable long-term proficiency trait. This distinction is fundamental for avoiding overclaiming psychometric validity.

Funding

This research received no external funding.

Competing Interests

The authors declare no competing interests.

Author Contributions

Y.W. conceived the study, developed the overall research framework, and contributed to the design of the retrieval-augmented grammar assessment methodology.

R.L. contributed to dataset preparation, Chinese grammatical knowledge organization, rubric development, and linguistic analysis of grammatical error categories.

K.L. contributed to the development and implementation of the proof-carrying counterfactual grading framework, experimental design, baseline comparisons, and performance evaluation.

D.Q. supervised the research, contributed to the educational assessment framework, interpreted the experimental findings, and critically revised the manuscript.

Y.W. prepared the initial manuscript. All authors contributed to manuscript revision and reviewed and approved the final version.

Data Availability

This study considers publicly released Chinese grammatical error diagnosis and correction datasets, including MuCGEC, FCGEC, NaSGEC, and CGED-2020.

The MuCGEC dataset is available at <https://github.com/HillZhang1999/MuCGEC>. The FCGEC dataset is available at <https://github.com/xlxwalex/FCGEC>. Historical CGED datasets can be accessed at https://github.com/blcuicall/cged_datasets. NaSGEC is available through its original dataset release, subject to the corresponding access conditions.

These datasets provide learner-generated or native-speaker Chinese sentences, grammatical error annotations, and reference corrections. The original datasets do not provide independent teacher-assigned grammatical mastery grades.

The datasets are subject to their respective licensing, access, and redistribution conditions. No new primary dataset was collected for the reported study.

Ethics Declaration

This study involves computational research on retrieval-augmented generation, natural language processing, and automated Chinese grammar assessment using previously released textual datasets.

The reported experiments do not involve prospective recruitment of human participants, classroom interventions, or animal experimentation. No new human-subject experiments are reported.

The use of learner-authored linguistic data is subject to the privacy, consent, and access conditions of the original dataset providers.

Any institutional requirements concerning ethics approval or exemption for secondary analysis of existing learner-language datasets must be confirmed before submission.

References

1. Rao, G., Yang, E. & Zhang, B. Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnosis. In *NLP-TEA 2020* (2020).
2. Zhang, Y. *et al.* MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction. In *NAACL 2022* (2022).
3. Xu, L., Wu, J., Peng, J., Fu, J. & Cai, M. FCGEC: Fine-Grained Corpus for Chinese Grammatical Error Correction. In *Findings of EMNLP 2022* (2022).
4. Ma, S., Li, Y., Sun, R. *et al.* Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction. In *Findings of EMNLP 2022* (2022).
5. Zhang, Y. *et al.* NaSGEC: a Multi-Domain Chinese Grammatical Error Correction Dataset from Native Speaker Texts. In *Findings of ACL 2023* (2023).
6. Li, W. & Wang, H. Detection-Correction Structure via General Language Model for Grammatical Error Correction. In *ACL 2024* (2024).
7. Yang, H. & Quan, X. Alirector: Alignment-Enhanced Chinese Grammatical Error Corrector. In *Findings of ACL 2024* (2024).
8. Liu, X., Li, Y. & Yu, Z. Chinese Grammatical Error Correction via Large Language Model Guided Optimization Training. In *CCL 2024* (2024).
9. Liu, X. *et al.* A Chain-of-Task Framework for Instruction Tuning of LLMs Based on Chinese Grammatical Error Correction. In *COLING 2025* (2025).
10. Chang, X. & Zhu, J. A Chunk-based Chain of Thought Prompting Method for Mitigating Over-Correction in Chinese Grammatical Error Correction. In *CCL 2025* (2025).
11. Gu, Y., Huang, Z., Zeng, M., Qiu, M. & Park, J. Improving Automatic Grammatical Error Annotation for Chinese Through Linguistically-Informed Error Typology. In *COLING 2025* (2025).
12. Kobayashi, M., Mita, M. & Komachi, M. Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction. In *BEA 2024* (2024).
13. Mu, L., Wang, X. & Cui, J. Evaluation of Large Language Models for Chinese Text Error Correction Tasks. In *CCL 2024* (2024).

14. Li, Y., Qin, S., Ye, J. *et al.* Rethinking the Roles of Large Language Models in Chinese Grammatical Error Correction. In *ACL Industry Track 2025* (2025).
15. Lewis, P., Perez, E., Piktus, A. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS 2020* (2020).
16. Izacard, G., Lewis, P., Lomeli, M. *et al.* Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res.* (2023).
17. Ram, O., Levine, Y., Dalmedigos, I. *et al.* In-Context Retrieval-Augmented Language Models. *Transactions Assoc. for Comput. Linguist.* (2023).
18. Shi, W., Min, S., Yasunaga, M. *et al.* REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv preprint (2023).
19. Jiang, Z., Xu, F., Gao, L. *et al.* Active Retrieval Augmented Generation. In *EMNLP 2023* (2023).
20. Shao, Z., Gong, Y., Shen, Y. *et al.* Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In *EMNLP 2023* (2023).
21. Trivedi, H., Balasubramanian, N., Khot, T. & Sabharwal, A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *ACL 2023* (2023).
22. Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection. In *ICLR 2024* (2024).
23. Jeong, S., Baek, J., Cho, S., Hwang, S. J. & Park, J. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *NAACL 2024* (2024).
24. Yan, S.-Q., Gu, J.-C., Zhu, Y. & Ling, Z.-H. Corrective Retrieval Augmented Generation. arXiv preprint (2024).
25. Huang, Y., Bai, Y., Zhu, Z. *et al.* C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *NeurIPS 2023* (2023).
26. Li, H. *et al.* CMMLU: Measuring Massive Multitask Language Understanding in Chinese. In *Findings of ACL 2024* (2024).
27. Zhong, W., Cui, R., Guo, Y. *et al.* AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv preprint (2023).
28. Zhang, W., Aljunied, S. M., Gao, C., Chia, Y. K. & Bing, L. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *NeurIPS 2023* (2023).
29. Hou, J., Ao, C., Wu, H. *et al.* E-EVAL: A Comprehensive Chinese K-12 Education Evaluation Benchmark for Large Language Models. In *Findings of ACL 2024* (2024).
30. Yoran, O., Wolfson, T., Ram, O. & Berant, J. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *ICLR 2024* (2024).
31. Liu, N. F. *et al.* Lost in the Middle: How Language Models Use Long Contexts. *Transactions Assoc. for Comput. Linguist.* (2024).
32. Cuconasu, F., Trappolini, G., Siciliano, F. *et al.* The Power of Noise: Redefining Retrieval for RAG Systems. In *SIGIR 2024* (2024).
33. Es, S., James, J., Anke, L. E. & Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *EACL 2024 System Demonstrations* (2024).
34. Saad-Falcon, J., Khattab, O., Potts, C. & Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *NAACL 2024* (2024).
35. Zheng, L., Chiang, W.-L., Sheng, Y. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS 2023* (2023).

36. Liu, Y., Iter, D., Xu, Y. *et al.* G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *EMNLP 2023* (2023).
37. Yao, S., Zhao, J., Yu, D. *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR 2023* (2023).
38. Sarthi, P., Abdullah, S., Tuli, A. *et al.* RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *ICLR 2024* (2024).
39. Xu, F., Shi, W. & Choi, E. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In *ICLR 2024* (2024).
40. Zhang, T., Patil, S. G., Jain, N. *et al.* RAFT: Adapting Language Model to Domain Specific RAG. arXiv preprint (2024).
41. Khattab, O., Singhvi, A., Maheshwari, P. *et al.* DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *ICLR 2024* (2024).
42. Gao, L., Dai, Z., Pasupat, P. *et al.* RARR: Researching and Revising What Language Models Say, Using Language Models. In *ACL 2023* (2023).
43. Min, S., Krishna, K., Lyu, X. *et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *EMNLP 2023* (2023).
44. Zha, Y., Yang, Y., Li, R. & Hu, Z. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. In *ACL 2023* (2023).
45. Manakul, P., Liusie, A. & Gales, M. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *EMNLP 2023* (2023).
46. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* (2024).
47. Kim, S., Shin, J., Cho, Y. *et al.* Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. In *ICLR 2024* (2024).
48. Kim, S., Suk, J. *et al.* Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *EMNLP 2024* (2024).
49. Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv preprint (2024).
50. Chiang, W.-L., Zheng, L., Sheng, Y. *et al.* Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *ICML 2024* (2024).
51. Liang, P., Bommasani, R., Lee, T. *et al.* Holistic Evaluation of Language Models. *Transactions on Mach. Learn. Res.* (2023).
52. Mallen, A., Asai, A., Zhong, V. *et al.* When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *ACL 2023* (2023).
53. Gao, Y., Xiong, Y., Gao, X. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint (2023).
54. Fang, J. *et al.* CKnowEdit: A New Chinese Knowledge Editing Dataset for Linguistics, Facts, and Logic Error Correction in LLMs. In *ACL 2025* (2025).