

Confidence-Gated Indonesian Grammatical Error Correction Under Synthetic-to-Authentic Domain Shift

Eko Suroso^{1*}, Onok Yayang Pamungkas², Laily Nurlina³, Sudiyono⁴, Jimat Susilo⁵, Eko Priyanto⁶, Hera Septriana⁷

Indonesian Language and Literature Education, Universitas Muhammadiyah Purwokerto, Banyumas, Indonesia^{1, 2, 3, 7}

Social Sciences and Humanities Research Center, National Research and Innovation Agency of the Republic of Indonesia, Indonesia⁴

Indonesian Language Education, Universitas Swadaya Gunung Jati, Cirebon, Indonesia⁵

Pancasila and Citizenship Education, Universitas Muhammadiyah Purwokerto, Banyumas, Indonesia⁶

Abstract—Indonesian grammatical error correction (GEC) requires decisions about when to apply a proposed edit under synthetic-to-authentic domain shift. We evaluate a selective correction framework using a reconstructed, fully executable experiment. An audit of 1,955 authentic records removed 36 incomplete pairs, 14 duplicates, 45 string identities, and five additional token identities, leaving 1,855 pairs. Exact-text groups were partitioned into 1,297 training, 280 validation, and 278 test pairs. Synthetic priors were learned from 200,000 public development pairs. A cross-fitted logistic ranker combined synthetic mappings, authentic templates, spelling retrieval, and local language-model evidence. Thresholds were chosen on validation data under joint degradation and overcorrection budgets. At the 2% budget, the gate accepted five source edits, four correct: precision 80.00%, F0.5 4.03, coverage 1.80%, and word-error reduction 0.88%. No test source was degraded, but one of 278 correct-target controls changed. The precision interval was 37.55–96.38%. Removing the synthetic channel reduced available gold candidates from 110 to 109 and accepted true positives from four to three. Replacing the language-model feature with frozen IndoBERT produced 66.67% precision at 2.16% coverage, with one degraded source and one changed control. The results support restrictive authorization on this corpus, show limited unique synthetic contribution, and expose a substantial candidate-recall bottleneck. They do not establish broad correction ability or readiness for educational deployment.

Keywords—Indonesian grammatical error correction; domain shift; synthetic data; selective prediction; abstention; overcorrection; trustworthy natural language processing; writing feedback

I. INTRODUCTION

Grammatical error correction (GEC) is the computational task of detecting and repairing errors in written text, including orthographic, morphological, syntactic, lexical, and meaning-related problems [1]. The task is technically demanding because a system must not merely generate fluent text. It must identify which segment is erroneous, make the smallest defensible intervention, preserve the writer's intended meaning, and leave already acceptable text unchanged. These requirements are especially important in educational settings. A false correction may teach an incorrect rule, erase a legitimate stylistic choice, or encourage learners to surrender judgment to an automated system. For Indonesian-language education, reliable GEC could

support formative feedback in academic writing, editing practice, and self-regulated revision. Yet an educationally safe system must recognize when it lacks sufficient evidence and defer to a teacher or human editor.

Indonesian GEC has progressed from a correction framework [2] to a Transformer supported by synthetic training pairs [3] and a framework for constructing evaluation corpora [4]. International work addresses complementary limitations: contextual augmentation targets error-distribution mismatch [5], RobustGEC tests sensitivity to contextual perturbations [6], and NaSGEC evaluates domain variation [7]. These developments motivate a separate question for writing assistance: when should a proposed correction actually be applied?

Synthetic corpora provide frequent spelling substitutions, character confusions, and surface edits, but their distribution reflects the corruption procedure. Authentic writing can require phrase restructuring, lexical interpretation, or several interacting changes. Wang et al. [5] combine contextual augmentation with relabeling-based cleaning; MixEdit [8] examines augmented-data affinity and diversity; and Stahlberg and Kumar [9] use tagged corruption to control synthetic error types. These approaches motivate testing distributional compatibility rather than assuming that more synthetic pairs guarantee reliable transfer.

Evaluation on error-containing sentences alone does not measure the cost of changing acceptable text. A selective writing assistant must therefore distinguish applying an edit from withholding it. In this implementation, correction is an accepted edit, retention describes unchanged output, and abstention describes withholding a proposed edit because its score is insufficient. Retention does not certify that a sentence is grammatical. The broader interface concept of correct, retain, or defer must not be interpreted as an evaluated three-class classifier. Varshney et al. [10] show that selective prediction performance can vary across tasks and distribution shifts, motivating direct evaluation of both erroneous inputs and correct-target controls.

This study combines synthetic replacement evidence from the Indonesian Transformer project [3] with the public GEC-GEE release [11]. It measures direct transfer, candidate-source ablations, budget sensitivity, score-tail calibration, and a frozen

*Corresponding author.

contextual scoring alternative. The paired evaluation processes every erroneous source and its corrected target as a no-error control. This makes unnecessary changes observable alongside successful repairs.

For this revision, we reconstructed the computational pipeline because the original executable artifacts and split indices could not be recovered from the submission files. The experiments below constitute a new evaluation and supersede the previous performance values. The audit, tokenizer, ontology rules, grouped partitions, model settings, candidate logs, and policy selections are now explicitly recorded. The historical 4.66% Identity exact-match value and the omitted category case cannot be traced to original predictions; both reporting problems are addressed through consistent token-level filtering and complete denominators in the reconstructed benchmark.

Four questions guide the revised evaluation: what marginal contribution do synthetic candidates make within the combined model (RQ1); how much observed harm is reduced by authorization at controlled coverage (RQ2); which error families and pipeline stages limit useful correction (RQ3); and how do alternative harm budgets and contextual scoring affect the operating point (RQ4)? The contribution is a reproducible evaluation protocol with paired controls and explicit abstention, rather than a claim of a new general-purpose GEC generator.

II. RELATED WORK

A. Modern Grammatical Error Correction

The survey by Bryant et al. [1] organizes sequence generation, tagging, edit-based modeling, and evaluation. MainGEC [12] weights training examples using annotation properties; MixEdit [8] studies augmentation affinity and diversity; TransGEC [13] uses translated text to reduce the style mismatch in augmentation; and Mesham et al. [14] extend the sequence-tagging vocabulary for spelling and morphology. Cao et al. [15] introduce non-autoregressive signals into autoregressive correction. Luhtaru et al. [16] investigate multilingual transfer using pretrained translation models, while MoECE [17] uses expert specialization to improve efficiency and interpretability. These methods concern how candidate corrections are learned or generated; they do not establish the safety of the present authorization policy.

Fei et al. [18] incorporate explanations into GEC systems, while GEE! [19] evaluates whether large language models can locate and explain errors. Peng et al. [20] retrieve demonstrations using grammatical-error representations rather than semantic similarity alone. Li et al. [21] use explanations for demonstration retrieval. These approaches can strengthen candidate support, but an explanation must still be assessed separately from whether the proposed edit is correct.

B. Synthetic Data and Domain Mismatch

Wang et al. [5] combine rule-based substitution, context generation, and relabeling to improve synthetic examples. Stahlberg and Kumar [9] generate low-resource training data with tagged corruption models. In Bangla, Bhattacharyya and Bhattacharyya [22] define error categories and use rule-based noise for instruction-tuning data. Smruti [23] combines a language model with

non-parametric memory for Gujarati and supplies an evaluation dataset. Team Horizon [24] combines synthetic augmentation with multilingual Transformer models for Indic GEC. These are distinct data and modeling strategies; their reported benefits do not directly establish transfer to the Indonesian release evaluated here.

For Indonesian, Musyafa et al. [3] provide a synthetic corpus for Transformer-based correction. Lin et al. [4] propose a separate corpus-construction framework. The authentic data audited here are the Indonesian GEC-GEE release maintained in the syauqiezjut repository [11]; they should not be conflated with the corpus released with [4]. Frequent synthetic substitutions can generate plausible candidates, but their incremental contribution to correct accepted edits must be measured by removing the synthetic channel while retaining the remaining gated pipeline.

C. Robustness and Overcorrection

RobustGEC [6] examines sensitivity to minor contextual perturbations, and NaSGEC [7] documents cross-domain variation. Chang and Zhu [25] use chunk-based prompting to limit unnecessary correction. LM-Combiner [26] employs contextual rewriting to combine correction outputs. Liu et al. [27] introduce a chain-of-task instruction-tuning framework, while Wang et al. [28] investigate multi-pass decoding. These approaches motivate explicit control tests because improved correction or generation scores need not imply fewer unnecessary changes.

D. Evaluation and Selective Prediction

The gec-metrics library [29] provides a common interface for GEC evaluation. IMPARA-GED [30] incorporates grammatical-error detection into reference-free quality estimation, whereas Goto et al. [31] examine reliability problems in reference-free metrics. Varshney et al. [10] compare selective prediction across distribution settings; their subsequent method [32] calibrates predictions using confidence and instance difficulty. Zablotskaia et al. [33] evaluate uncertainty and selective generation for summarization, and Ba et al. [34] use synthetic data for calibration. Guo et al. [35] study post-hoc calibration of neural networks. These findings support reporting ranking quality, calibration, coverage, and observed harm as separate quantities.

E. Indonesian NLP and Educational Writing Support

IndoNLU [36] supplies language-understanding benchmarks and models, and IndoNLG [37] supplies benchmarks for language generation. IndoBERTweet [38] adapts a pretrained encoder to Indonesian Twitter, IndoNLI [39] provides natural-language inference data, and Cendol [40] develops instruction-tuned generative models for Indonesian languages. KEBI [41] addresses spelling detection in scientific papers, while U-Tapis [42] evaluates spelling support for journalism students. Mohsen [43] synthesizes computer-mediated corrective-feedback studies; Zhai and Ma [44] synthesize automated writing evaluation; Fan [45] examines automated feedback in an EFL writing study; and Lee and Moore [46] review generative-AI feedback in higher education. These studies motivate supervised feedback research but do not demonstrate learning benefits for the present system. Koto et al. [47] introduce IndoLEM and a pretrained Indonesian BERT model; the present comparison uses its frozen masked-language-model checkpoint.

III. DATA SOURCES AND QUALITY AUDIT

A. Synthetic Source

Musyafa et al. [3] report a large synthetic corpus generated from Indonesian text. The accessible development and test CSV archives contained 424,430 and 424,419 parsable pairs, respectively, totaling 848,849. Their 47 string identities and 67 token identities differ because tokenization ignores spaces adjacent to punctuation. The archives are audited separately; the synthetic test partition contributes no training evidence.

Using Python random seed 2026, we sampled 200,000 development rows without replacement after checking exact normalized-text overlap with the authentic corpus. No overlapping development pair was found. Minimum-edit alignment supplied single-token replacement counts. A replacement was retained at support ≥ 2 and conditional probability ≥ 0.50 among nonidentity substitutions of its source token, with at most three alternatives per source. This yielded 40,147 retained replacements for 39,985 source tokens. Sample row identifiers and hashes of all three source CSVs are included in the computational supplement.

B. Authentic Source

The GEC-GEE repository [11] describes 1,920 expert-validated pairs from academic publications dated 2020–2024, whereas its downloadable CSV contains 1,955 records. It supplies source, target, label, and explanation fields. It does not supply row-level genre, writer proficiency, register, or document provenance. We therefore describe its stated origin without assigning unsupported demographic or genre proportions. Its scope is one heterogeneous public release.

We normalized text with Unicode NFKC, collapsed whitespace, removed blank source/target pairs, retained the first

duplicate normalized pair, and removed remaining string identities. These sequential exclusions were 36, 14, and 45, yielding 1,860 string-distinct pairs. The case-preserving tokenizer matches Unicode words and individual nonspace punctuation with the regular expression `\w+|[\^\w\s]`. Five additional pairs were identical under these tokens and were explicitly removed, leaving 1,855 pairs. No additional token-pair duplicate remained. Training, exact match, edit extraction, and distance evaluation all use these same token sequences, so Identity has zero exact match on the error benchmark.

C. Ontology and Grouped Partitioning

Existing labels were mapped by published, case-normalized keyword rules to orthography, punctuation, morphology, function word, redundancy, lexical choice, semantic coherence, and syntax, with Other for residual labels. Typographical variants were explicitly standardized. Primary-family priority follows that order; multiple labels are retained for descriptive purposes. For example, *afiksasi* and *reduplikasi* indicate morphology, *preposisi* and *konjungsi* indicate function words, and *pleonasme* indicates redundancy. These are operational groupings of existing annotations, not newly validated linguistic judgments.

Exact token-identical source or target texts were joined into 1,845 connected groups. A seeded search over 400 group partitions minimized size and category-count imbalance relative to a 70/15/15 split, using no model outcomes. The selected partitions contain 1,297 training, 280 validation, and 278 test pairs, with no exact text shared between partitions. Their row IDs and complete label-to-family mappings are supplied. Across the final corpus, sources average 14.00 tokens, with median 12 and 95th percentile 26.3; minimum-edit alignment yields 3,159 gold spans. The new counts supersede the historical tokenizer and split statistics.

TABLE I. SEQUENTIAL CLEANING WITH CONSISTENT TOKEN EVALUATION

Sequential stage	Entering	Removed	Remaining
Release	1,955	0	1,955
Remove empty source or target	1,955	36	1,919
Deduplicate normalized pairs	1,919	14	1,905
Remove string identities	1,905	45	1,860
Remove additional token identities	1,860	5	1,855
Deduplicate token pairs	1,855	0	1,855

Table I accounts for all 100 exclusions without adding overlapping audit categories. The distinction between string identity and token identity is explicit. This prevents an unchanged source from being counted as an exact correction merely because evaluation applies a more permissive normalization. Dataset quality and preprocessing decisions affect downstream inference and must be auditable [48], [49], [50].

IV. PROPOSED FRAMEWORK

A. Design Principle

Candidate discovery proposes a local change; ranking orders proposals; authorization determines which changes are applied. Correction means an accepted edit. Retention describes unchanged output, while abstention records rejection below the

score threshold. Overlap and the five-edit cap are separate rejection reasons; no-candidate retention is also logged. An unchanged output does not certify grammatical correctness. The implementation is therefore an edit-selection policy with rejection records, not an evaluated three-class correct/retain/abstain classifier. Fig. 1 shows the full experimental flow.

B. Tokenization and Alignment

Case-preserving tokens are aligned by unit-cost Levenshtein operations [51] using RapidFuzz 3.14.6. Consecutive non-equal operations are merged into a single span; exact matches use source start/end offsets and the replacement token tuple. Authentic templates are retained when each side contains at most three tokens, including insertion and deletion. Support is the number of training occurrences. Conditional probability divides

template support by all nonidentity rule support for the same source span; insertion rules share an empty-source denominator.

This definition is available in executable code rather than inferred from reported metrics.

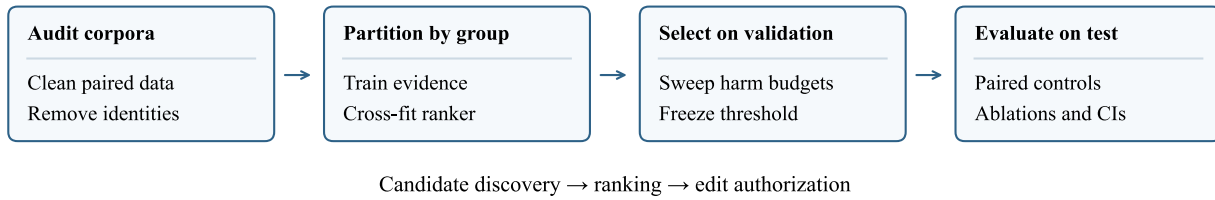


Fig. 1. Reconstructed data flow and separation of candidate discovery, ranking, and authorization.

C. Candidate Sources

Synthetic candidates follow the fixed support/probability filters in Section III. Authentic templates store distributions of the immediately adjacent left and right tokens, including sentence boundaries. Context agreement is half the sum of the two matching-neighbor proportions. An insertion requires at least one matching neighbor. Spelling retrieval searches only the current training-target vocabulary: alphabetic tokens of length ≥ 3 and frequency ≥ 2 , Levenshtein distance ≤ 1 for source length ≤ 5 and ≤ 2 otherwise, normalized similarity ≥ 0.70 , and at most three alternatives ordered by distance, descending frequency, then lexical order. No dictionary or reference target is consulted during inference.

D. Candidate Features and Language Model

Sixteen features encode log support; conditional probability; log source and target frequencies; their smoothed log ratio; character similarity; source and replacement token lengths; character-length difference; punctuation and case-change flags; three candidate-origin flags; context agreement; and local language-model gain. Duplicate candidates with the same offsets and replacement are merged while retaining all origin flags. Source and target frequencies use only the current authentic training fold. The language model uses lowercased target-training bigrams, BOS/EOS and UNK, with add-0.1 smoothing. Gain is the difference in mean log-likelihood within two neighboring tokens on either side. The complete feature order and formulas accompany the code.

E. Cross-Fitted Ranker

Five group-disjoint folds within the training partition induced rules and vocabulary on four folds and generated candidates for held-out sources and their target controls.

Candidate labels indicate exact gold-edit matches. This produced 207,580 candidates and 472 positives. StandardScaler and L2 logistic regression were fitted to these examples with $C=1$, balanced class weights, lbfgs, max_iter=2000, and seed 2026. The final evidence dictionaries used only the complete authentic training partition. Calibration metrics are reported on validation and test candidates, not on the same candidates used to fit the ranker.

F. Validation-Constrained Authorization

Thresholds cover 0.00–1.00 at increments of 0.01, plus 0.995, 0.999, and a retain-all option. Validation F0.5 is maximized subject to the chosen budget; ties favor lower degradation, lower overcorrection, and then a higher threshold. Budgets of 1%, 2%, and 5% are evaluated with overcorrection alone and with a joint bound on degradation and overcorrection. The primary condition is the full model at the joint 2% budget, fixed before inspecting its test outputs. This is an empirical tolerance for comparing policies, not a validated educational safety standard. A budget does not guarantee the same population or test risk.

G. Comparators and Ablations

Table II defines Identity, direct synthetic and authentic rules, a fixed heuristic, an ungated ranker, the primary gate, four ablations, and an IndoBERT feature replacement. Channel ablations remove both candidates unique to the channel and that channel's support/probability/origin evidence in merged candidates; the logistic ranker is refitted. Authentic language statistics remain training-only in every channel ablation. The no-LM condition sets the local gain feature to zero. All active policies apply at most five nonoverlapping edits in one pass, without iterative re-writing.

TABLE II. EVALUATED CORRECTION VARIANTS

Variant	Candidate evidence	Authorization policy
Identity	None	Retain every input
Synthetic direct	Synthetic channel	Probability/support ranking
Authentic direct	Authentic channel	Probability/support ranking
Heuristic	All channels	Fixed weighted score ≥ 0.50
Ungated ranker	All channels	Logistic score with no gate
Confidence gated	All channels	Validation-selected joint 2% gate
Channel ablations	Remove synthetic, authentic, or spelling	Refit and reselect independently
No n-gram feature	Same full candidate pool	Zero LM gain; refit and reselect
IndoBERT feature	Same full candidate pool	Replace LM gain; refit and reselect

Algorithm 1 summarizes the reconstructed inference procedure. The implementation code and per-sample numerical results are provided in the supplementary material described at the end of this paper.

Algorithm 1. Confidence-Gated Selective Correction

Input: source tokens x ; training evidence E ; fitted ranker f ; validation-selected threshold τ

1. Generate local candidates from the permitted channels and deduplicate identical edits.
 2. Compute inference features from x and training evidence only; preserve origin flags.
 3. Score and sort candidates once, breaking ties deterministically by offsets and replacement.
 4. Accept an edit only if its score $\geq \tau$, it does not overlap an accepted edit, and fewer than five edits have been accepted.
 5. Log rejection as below threshold, overlap, or edit cap; distinguish no-candidate retention.
 6. Apply accepted edits from right to left to preserve source offsets.
 7. Return the output and rejection record. Retention does not certify correctness.
-

V. EXPERIMENTAL DESIGN

A. Evaluation Units

The held-out test set contains 278 erroneous sources with 476 gold edit spans and 278 corresponding reference-correct controls. Validation has 280 paired sources/controls. Exact-text grouping prevents the same source or target from appearing in both training and evaluation. Shared words and rules remain possible, as expected for local correction; near-duplicate and provenance overlap are not ruled out. Each variant's threshold is selected on validation before its test evaluation.

B. Edit Metrics

A system edit is a true positive only if its source offsets and replacement tokens exactly equal a gold span. Precision equals TP/predicted edits; recall equals TP/gold edits; $F0.5 = 1.25TP/(predicted\ edits + 0.25\ gold\ edits)$. Scores are displayed on a 0–100 scale. A partly helpful change can lower token distance without matching a complete reference edit. When a system makes no edit, precision is mathematically undefined; tables mark it with a dash, while the evaluation convention sets F0.5 and recall to zero.

C. Sentence and Distance Metrics

Exact match compares output and target tokens. Coverage is the proportion of source sentences changed. Improvement and degradation compare output-to-target Levenshtein distance with the initial source-to-target distance. Overcorrection is any token change on a target control. Word-error reduction is one minus the ratio of aggregate output and initial token distances; character reduction uses the analogous distance on token-normalized strings. We distinguish unconditional degradation over all sources from conditional degradation among changed sources, and report both denominators.

D. Uncertainty and Comparisons

We report Wilson 95% intervals for rates [52], two-sided ex

act McNemar degradation comparisons [53], Holm adjustment across ten comparisons with the primary gate, and 10,000 paired sentence bootstrap samples with seed 2026 for token-distance gain [54]. These are descriptive intervals for this corpus; candidates from the same sentence can be dependent, so candidate-tail Wilson intervals do not establish independent-event calibration guarantees. A matched-coverage diagnostic restricts each ranked comparator to five changed source sentences and one edit per sentence. Two hundred seeded random authorizations provide an additional descriptive comparison, without influencing model selection.

E. Category Accounting and Bottlenecks

Every test pair appears once under its primary family, including Other. The analysis distinguishes gold edits that never enter the candidate pool from available gold edits rejected during authorization. Removing a candidate channel can change both the pool and its scoring evidence, so an end-to-end ablation effect is not automatically a uniquely generated edit. Accepted-candidate origin flags support a narrower attribution analysis.

F. Contextual Feature Comparison

We replace only the local n-gram gain with a frozen IndoBERT masked-span score [47], using indolem/indobert-base-uncased at revision 7ccb3cd0f5b08ffbaa465aade22328e8600e23eb. All edited subword tokens are masked simultaneously, and the feature is the change in their mean log probability between original and proposed spans. Windows use up to six neighboring word tokens per side, capped at 32 subwords per side and 16 edited subwords. Insertions/deletions include a shared neighboring token as an anchor. The trained MLM head is loaded with no missing weights. Candidates, other features, folds, and logistic settings are identical; no neural fine-tuning or candidate enrichment is performed.

G. Reproducible Implementation

The executed environment uses Python 3.12, NumPy 2.5.3, SciPy 1.18.1, scikit-learn 1.9.0, RapidFuzz 3.14.6, PyTorch 2.14.0, and Transformers 4.57.6. IndoBERT runs on CPU with four threads and batches of 32. The heuristic is fixed at 0.30 conditional probability + 0.25 context agreement + 0.20 similarity + 0.15 sigmoid(LM gain) + 0.10 support/(support+2), accepting scores ≥ 0.50 . Direct-rule ties use support after conditional probability. The supplement records exact software versions, hashes, all split IDs, feature definitions, candidate counts, and validation sweeps.

H. Reconstruction and Leakage Controls

The revised protocol was written before model fitting and test-outcome inspection. It replaces undocumented historical choices with explicit settings and uses new grouped partitions. The full n-gram model is the prespecified primary condition, rather than the variant with the highest test score. No test label, explanation, or target enters rule induction, frequency estimation, logistic fitting, or threshold selection. Exact-text overlap checks use corpus strings only for duplicate detection. Unrecoverable historical calibration variants and their test-selected comparison have been removed, and the new values are disclosed as a reconstruction rather than exact reproduction.

VI. RESULTS

A. Main Comparison

Table III presents accuracy and intervention outcomes separately. Identity correctly has zero exact matches. Direct synthetic application changes 63.67% of sources, degrades 53.60%, and changes 58.99% of controls. Its word-error reduction is -25.83%, demonstrating harmful direct transfer in this implementation. These values supersede those obtained with the unrecovered historical preprocessing and candidate filters.

Authentic direct rules change every source and every control, with 95.32% degradation. The ungated ranker degrades

93.17% and also changes every control. The heuristic has the higher unconstrained F0.5 of 8.03, but degrades 53.24% and overcorrects 62.23%. A higher conventional edit score therefore does not imply an acceptable intervention policy.

The primary gate selects threshold 0.999 and changes five sources, with four of five system edits matching the reference. Precision is 80.00%, recall 0.84%, and F0.5 4.03. Coverage is 1.80%, word-error reduction 0.88%, and one source is corrected exactly. All five changed sources become closer to their targets; none degrades. One correct-target control changes (0.36%), so the revised results do not support the earlier claim of zero observed overcorrection.

TABLE III. RECONSTRUCTED TEST PERFORMANCE ON 278 SOURCES AND 278 CONTROLS (PANEL A EDIT AND SENTENCE ACCURACY)

System	Precision (%)	Recall (%)	F0.5	Exact match (%)
Identity	—	0.00	0.00	0.00
Synthetic direct	2.01	1.26	1.79	0.72
Authentic direct	2.31	5.04	2.59	0.00
Heuristic	8.83	5.88	8.03	3.24
Ungated ranker	2.99	6.30	3.34	0.00
Confidence gated	80.00	0.84	4.03	0.36
No synthetic channel	75.00	0.63	3.05	0.36
No authentic channel	18.18	0.84	3.55	0.36
No spelling channel	—	0.00	0.00	0.00
No n-gram feature	—	0.00	0.00	0.00
IndoBERT feature	66.67	0.84	4.00	0.00

PANEL B SENTENCE INTERVENTION AND HARM

System	Degradation (%)	Coverage (%)	Word error reduction (%)	Overcorrection (%)
Identity	0.00	0.00	0.00	0.00
Synthetic direct	53.60	63.67	-25.83	58.99
Authentic direct	95.32	100.00	-114.90	100.00
Heuristic	53.24	70.86	-23.51	62.23
Ungated ranker	93.17	100.00	-102.87	100.00
Confidence gated	0.00	1.80	0.88	0.36
No synthetic channel	0.00	1.44	0.77	0.36
No authentic channel	5.04	7.55	-1.10	5.04
No spelling channel	0.00	0.00	0.00	0.00
No n-gram feature	0.00	0.00	0.00	0.00
IndoBERT feature	0.36	2.16	0.66	0.36

B. Safety and Useful Intervention

Table IV reports the uncertainty in the low-intervention results. The primary precision interval is 37.55–96.38%. Unconditional degradation is 0/278, with an upper bound of 1.36%; conditional degradation is 0/5 changed sources, whose Wilson upper bound is 43.45%. Overcorrection is 1/278, with interval 0.06–2.01%. These denominators prevent the unconditional zero-degradation observation from being interpreted as evidence of near-perfect reliability for accepted corrections.

The primary mean token-distance gain is 0.0288 tokens per source, with bootstrap interval 0.0072–0.0576. Paired harm tests favor the gate over the aggressive direct and ungated systems,

but those comparisons alone confound authorization with coverage. At a fixed five changed sources and one edit each, the full ranking retains four correct edits without degradation. Random selection averages 0.60% precision and 1.61% unconditional degradation over 200 runs. The matched diagnostic is small and descriptive; it does not establish a general safety guarantee or universal superiority at all coverage levels.

Fig. 2 separates coverage from the two harm rates. The primary gate improves observed safety through very restrictive selection. Identity attains no intervention-related harm by doing nothing, so the gain in useful corrections and its uncertainty must accompany any safety comparison.

TABLE IV. UNCERTAINTY AND PAIRED DEGRADATION COMPARISONS

System	Source harm	Controls changed	Precision 95% CI	Mean gain 95% CI	Adjusted p
Synthetic direct	149/278	164/278	0.92–4.31%	-0.9676–0.7194	2.2e-44
Authentic direct	265/278	278/278	1.56–3.41%	-3.9496–3.5396	3.4e-79
Heuristic	148/278	173/278	6.18–12.47%	-0.9245–0.6151	3.9e-44
Ungated ranker	259/278	278/278	2.10–4.23%	-3.5755–3.1259	1.9e-77
Confidence gated	0/278	1/278	37.55–96.38%	0.0072–0.0576	—
IndoBERT feature	1/278	1/278	30.00–90.32%	-0.0072–0.0504	1

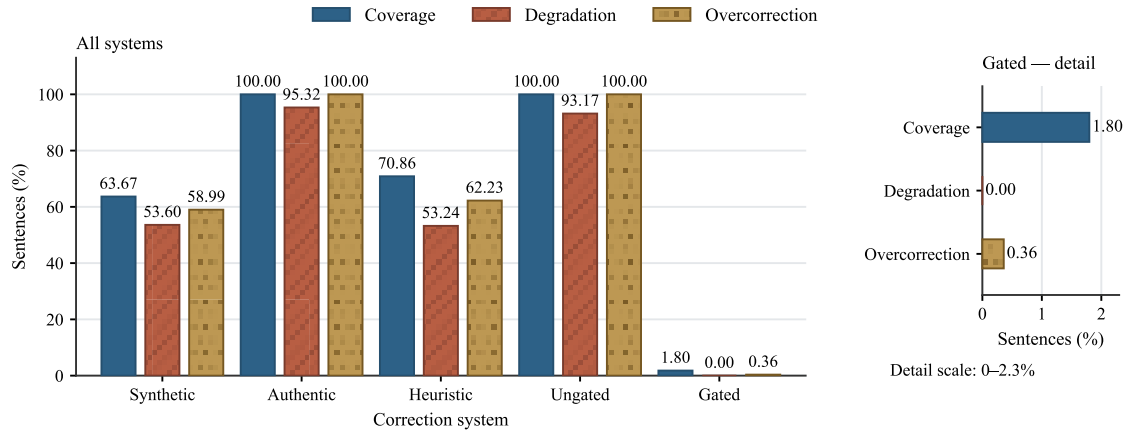


Fig. 2. Sentence-level harm and benefit under synthetic-to-authentic transfer.

C. Budget Sensitivity

Under the full n-gram model, both the overcorrection-only and joint formulations select retain-all at 1%, threshold 0.999 at 2%, and 0.995 at 5%. The resulting test coverage is 0%, 1.80%, and 7.91%; F0.5 is 0, 4.03, and 8.87. At 5%, test degradation

risks to 3.60% and overcorrection to 4.68%. Fig. 3 plots validation frontiers; the full validation/test sweep and six selected policies per model are supplied. The 280 validation controls permit at most 2, 5, or 14 changed controls under the three budgets. These are empirical tolerances, not confidence-bound guarantees.

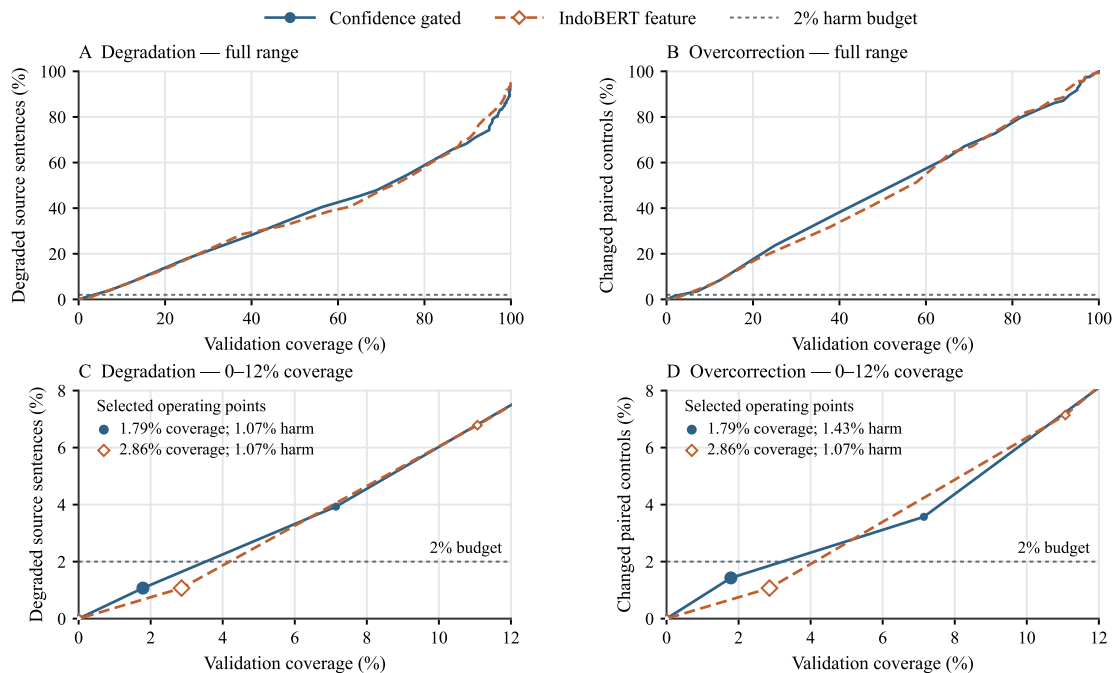


Fig. 3. Validation harm versus coverage for n-gram and IndoBERT scoring. Dots mark the joint 2% policies.

D. Error Families

Table V and Fig. 4 exhaust all 278 test sources: orthography 78, punctuation 15, morphology 56, function word 19, redundancy 23, lexical choice 42, semantic coherence 15, syntax 27, and Other 3. Correct edits occur in orthography (two), redundancy (one), and syntax (one). A lexical-group edit improves

distance without matching a complete gold span. The control overcorrection occurs in the morphology group. These family labels describe each pair's primary annotation, not necessarily the linguistic type of its accepted edit. The earlier inference that all structural categories were deferred is therefore withdrawn.

TABLE V. COMPLETE PRIMARY-FAMILY ACCOUNTING FOR THE GATED SYSTEM

Family	n	TP/edits	Recall (%)	F0.5	Coverage (%)	Controls changed (%)
Orthography	78	2/2	1.38	6.54	2.56	0.00
Punctuation	15	0/0	0.00	0.00	0.00	0.00
Morphology	56	0/0	0.00	0.00	0.00	1.79
Function word	19	0/0	0.00	0.00	0.00	0.00
Redundancy	23	1/1	2.63	11.90	4.35	0.00
Lexical	42	0/1	0.00	0.00	2.38	0.00
Semantic	15	0/0	0.00	0.00	0.00	0.00
Syntax	27	1/1	1.92	8.93	3.70	0.00
Other	3	0/0	0.00	0.00	0.00	0.00

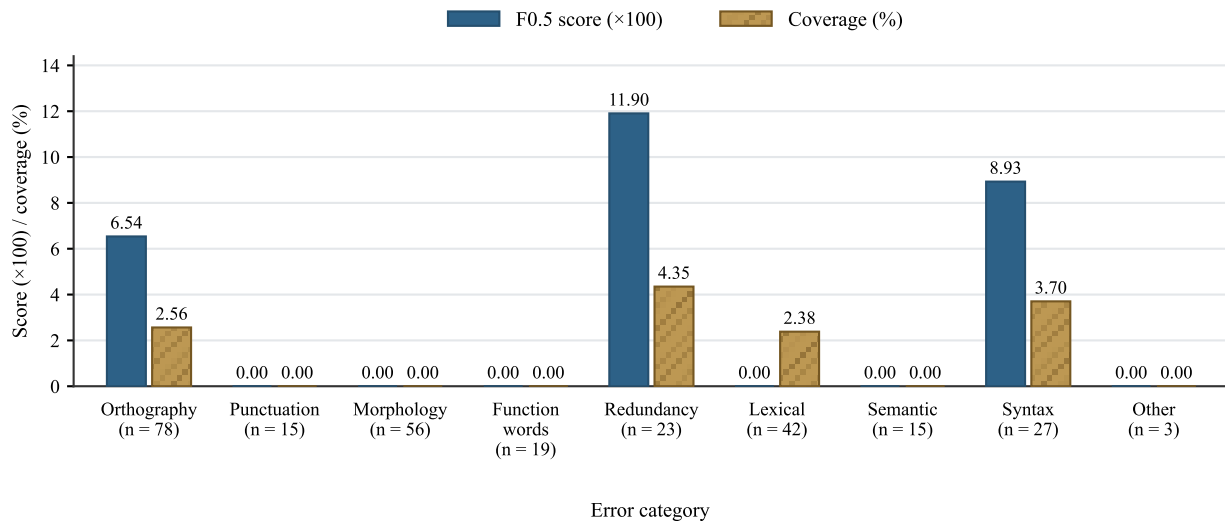


Fig. 4. Category-specific F0.5 and coverage for the confidence-gated system.

E. Candidate Attribution

The full pool contains 110 of 476 gold edits (23.11%), across 84 of 278 source sentences. Without synthetic candidates, 109 gold edits remain available, indicating only one uniquely supplied gold candidate. Of the four accepted gold candidates, three originate only from authentic templates and one is shared by authentic and synthetic channels. Removing synthetic evidence and refitting yields three true positives from four edits, F0.5 3.05 and coverage 1.44%. The one-edit end-to-end change may reflect scoring support; it is not proof that an accepted correction was unavailable without synthetic generation. The revised RQ1 conclusion is therefore limited marginal contribution on this split.

F. Additional Ablations

Removing the authentic channel reduces the gold-candidate ceiling to 26/476 (5.46%). Its validation-selected policy changes

21 test sources and degrades 14, illustrating that an empirical validation budget may fail to transfer even within the same release. Removing spelling retrieval leaves 96/476 gold candidates, but its joint 2% selection is retain-all. The no-LM condition also selects retain-all. These outcomes describe the fixed candidate filters, feature representation, grid, and split; they do not establish that the removed components are universally necessary.

VII. DISCUSSION

A. Synthetic Scale and Generality

The synthetic source provides frequent surface substitutions, but only one unique exact gold candidate beyond the remaining channels on this test set. Its support can still influence a shared candidate's ranking. The controlled ablation therefore narrows the earlier claim that scale benefits candidate discovery: large synthetic resources are insufficient evidence of useful marginal

transfer. Authentic templates supply most of the exact gold-candidate coverage in this reconstructed system.

B. Ranking and Tail Calibration

The n-gram ranker has validation AUC 0.925, average precision 0.047, Brier score 0.100, and ECE 0.181. On test candidates, these are 0.937, 0.093, 0.101, and 0.184. Among the 205 source/control candidates scoring ≥ 0.98 , only 20 match gold edits (9.76%), despite a mean score of 0.989. At ≥ 0.999 , four of six candidates are correct, with interval 30.00–90.32%; one of the two errors is a control edit. The gate succeeds by severe empirical filtering rather than demonstrated probability calibration. A score of 0.999 must not be read as a 99.9% chance of correctness.

C. Coverage and Conditional Risk

Only 110 gold edits enter the full candidate pool, and only four are recovered by the selected policy. Thus, 366/476 gold edits are unavailable to authorization at any threshold. All test sources have at least one candidate, so unchanged output is caused by rejection rather than evidence that a sentence is correct. The small number of accepted edits gives weak precision and conditional-harm estimates. Matched-coverage diagnostics reduce one confound but remain exploratory and depend on the supplied candidate universe.

D. Educational Implications

Mohsen [43] and Zhai and Ma [44] synthesize evidence about corrective feedback and automated writing evaluation, respectively; Fan [45] studies revision with automated feedback, and Lee and Moore [46] review generative feedback in higher education. The current study measures computational outcomes only. A proposed teacher-mediated interface would display candidate evidence and distinguish accepted edits from uncertain retained spans. It should not label every unchanged sentence as correct. Teacher workload, learner understanding, and writing improvement require prospective evaluation.

E. Contextual Scoring

Replacing the n-gram feature with frozen IndoBERT [47] keeps the same candidate pool and fits the same logistic architecture. At the joint 2% policy, its threshold is 0.999, with four correct edits from six, precision 66.67%, F0.5 4.00, and coverage 2.16%. It degrades one source and changes one control, each by 0.36%. Its test ECE is 0.192. The experiment provides no material safety-coverage improvement in this configuration. It evaluates local frozen masked-span scoring, not full-sentence neural generation or a fine-tuned contextual corrector; the small event counts cannot establish equivalence between the alternatives.

F. Measured Technical Bottlenecks

Candidate recall is bounded at 23.11% before authorization, reflecting short-span templates, sparse authentic evidence, and a limited synthetic match to the corpus. Gating then recovers only 4/110 available gold edits. Improving only the ranking feature cannot recover gold edits absent from the pool. Future experiments should therefore separate richer generation from better authorization, maintain correct-target controls, and report both candidate recall and policy-level harm. Neural generator enrichment was not tested here.

G. Criteria for a Supervised Pilot

The present system remains below a defensible deployment threshold. A concrete future screening rule is at least 10% coverage on an independently sampled corpus, at least 90% edit precision by the lower 95% interval, and upper 95% bounds below 1% for both degradation and control overcorrection. These are proposed engineering targets for discussion with teachers, not standards or thresholds validated by this experiment. The current 1.80% coverage and broad precision interval fail that screen. A pilot must also measure correct accepted feedback per teacher-review minute, teacher acceptance, and learner understanding against the existing review process; coverage alone is insufficient.

H. Comparison at Equal Intervention

With exactly five changed source sentences and one edit each, the full ranking yields four correct edits and no degradation. The heuristic also yields four correct edits, but degrades one source; synthetic ranking yields one correct edit and three degraded sources, and authentic ranking yields no correct edit and five degraded sources. The full method therefore does not improve precision over the heuristic in this small matched comparison. These descriptive outcomes use frozen scores and do not replace the validation-selected policies in Table III. Random-authorized controls are not assessed in this source-only diagnostic, and the five-intervention sample cannot establish robust safety superiority.

VIII. LIMITATIONS AND THREATS TO VALIDITY

The first limitation is reconstruction fidelity: new explicit filters, tokenization, ontology mapping, grouped partitions, and ranking settings supersede unrecovered historical choices. New results cannot identify the exact cause of every historical number. Second, all operating-point results come from a single split with a fixed seed. Sentence bootstrap intervals reflect test-sampling variation conditional on that fitted system, not training-seed or corpus-to-corpus variability. Third, the primary gate makes five source edits, yielding broad precision and conditional-risk intervals. Low observed harm therefore cannot establish reliable operation at greater coverage.

Fourth, the GEC-GEE CSV [11] lacks row-level genre, register, proficiency, and document provenance. Its repository description does not establish representative sampling, and ontology mapping was not independently reannotated. Exact-text grouping limits one form of leakage but does not remove near-duplicates, shared source documents, or possible pre-trained-model exposure. Fifth, the synthetic corruption process represents one particular resource. Sixth, the IndoBERT comparison is a frozen local-feature replacement; broader contextual models, fine-tuning, and candidate enrichment may behave differently. Calibration must be re-evaluated under every changed distribution.

Seventh, target controls are reference-correct within the dataset, not universally error-free writing. A single reference may penalize legitimate alternative corrections. Independent no-error texts and multiple references would strengthen ecological validity. Eighth, the empirical 1%, 2%, and 5% budgets do not control unknown population risk, and the proposed pilot targets remain unvalidated. Finally, no teacher-workload, learning, or

longitudinal writing outcome is measured. Those claims require prospective educational evaluation after stronger computational evidence is established.

IX. ETHICS, DATA AVAILABILITY, AND AI DISCLOSURE

This study analyzes publicly released text corpora and includes no new participant recruitment or intervention. The computational supplement contains executable code, source URLs and hashes, row identifiers, ontology mappings, aggregate statistics, and numeric per-sample outcomes. It does not redistribute third-party raw corpora or pretrained weights. Users should obtain those resources from their original providers and follow the corresponding terms. The source references identify the corpus and model used in the reconstruction.

Generative AI assisted with code development, methodological checking, drafting, and language revision. The reconstructed performance values are computed by executable scripts; unverified historical values are not retained as new experimental evidence. The authors remain responsible for checking the revised methods, data provenance, numerical results, and interpretations before submission. No gold correction or benchmark label was generated to improve the reported model performance.

Data and code availability. The accompanying reconstruction package includes preprocessing and model scripts, frozen split IDs and synthetic sample IDs, source hashes, validation selections, ablations, calibration tails, paired uncertainty analyses, figure generation, and verification checks. The package distinguishes the reconstructed evaluation from unrecovered historical artifacts. No funding or competing interests are declared.

X. CONCLUSION

The reconstructed experiment shows that aggressive local correction can introduce substantial harm on this public Indonesian corpus. Validation-constrained authorization reduces intervention to five of 278 sources, obtaining four exact gold edits, 80.00% precision, F0.5 4.03, and 0.88% word-error reduction. No source degrades, but one reference-correct control changes. All results use consistent token filtering and complete category denominators, and supersede the earlier unreproduced performance values.

The synthetic-channel ablation shows little uniquely added gold-candidate recall on this split. Budget sensitivity reveals a direct trade-off between useful coverage and observed harm, while the frozen IndoBERT replacement does not materially improve the selected operating point. The strongest remaining bottleneck is that most gold edits never enter the candidate pool. Selective authorization is therefore a useful research component, but improved generation, external evaluation, stronger uncertainty evidence, and teacher studies are required before the system can support claims of dependable educational deployment.

XI. SUPPLEMENTARY MATERIAL

Supplementary material for this article is available at https://drive.google.com/file/d/11qjUt0yG6TTwe0Nf_QEY5OybPaUcn3aX/view?usp=sharing. The package includes source code for data preprocessing, feature generation, policy selection, and evaluation, together with instructions for reproducing the experiments and the corresponding numerical results. Third-

party datasets and pretrained model weights are not included. The accompanying documentation provides instructions for obtaining these resources from their original providers.

REFERENCES

- [1] C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, and T. Briscoe, "Grammatical Error Correction: A Survey of the State of the Art," *Computational Linguistics*, vol. 49, no. 3, pp. 643-701, 2023, doi: 10.1162/coli_a_00478.
- [2] N. Lin, Y. Gao, C. Wu, and A. Musyafa, "A Framework for Indonesian Grammar Error Correction," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 4, art. 58, 2021, doi: 10.1145/3440993.
- [3] A. Musyafa, Y. Gao, A. Solyman, C. Wu, and S. Khan, "Automatic Correction of Indonesian Grammatical Errors Based on Transformer," *Applied Sciences*, vol. 12, no. 20, art. 10380, 2022, doi: 10.3390/app122010380.
- [4] N. Lin, M. Zeng, W. Huang, S. Jiang, L. Xiao, and A. Yang, "A Simple Yet Effective Corpus Construction Framework for Indonesian Grammatical Error Correction," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 1, art. 5, 2025, doi: 10.1145/3704264.
- [5] Y. Wang, B. Wang, Y. Liu, Q. Zhu, D. Wu, and W. Che, "Improving Grammatical Error Correction via Contextual Data Augmentation," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10898-10910, 2024, doi: 10.18653/v1/2024.findings-acl.647.
- [6] Y. Zhang, L. Cui, E. Zhao, W. Bi, and S. Shi, "RobustGEC: Robust Grammatical Error Correction Against Subtle Context Perturbation," in *Proc. EMNLP 2023*, pp. 16780-16793, 2023, doi: 10.18653/v1/2023.emnlp-main.1043.
- [7] Y. Zhang, B. Zhang, H. Jiang, Z. Li, C. Li, F. Huang, and M. Zhang, "NaSGEC: A Multi-Domain Chinese Grammatical Error Correction Dataset from Native Speaker Texts," in *Findings of ACL 2023*, pp. 9935-9951, 2023, doi: 10.18653/v1/2023.findings-acl.630.
- [8] J. Ye, Y. Li, Y. Li, and H.-T. Zheng, "MixEdit: Revisiting Data Augmentation and Beyond for Grammatical Error Correction," in *Findings of EMNLP 2023*, pp. 10161-10175, 2023, doi: 10.18653/v1/2023.findings-emnlp.681.
- [9] F. Stahlberg and S. Kumar, "Synthetic Data Generation for Low-resource Grammatical Error Correction with Tagged Corruption Models," in *Proc. BEA 2024*, pp. 11-16, 2024.
- [10] N. Varshney, S. Mishra, and C. Baral, "Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings," in *Findings of ACL 2022*, pp. 1995-2002, 2022, doi: 10.18653/v1/2022.findings-acl.158.
- [11] Syauqiezjut, "Indonesian Grammar Error Correction and Explanation Dataset," GitHub repository. Accessed: Sep. 10, 2026. [Online]. Available: <https://github.com/syauqiezjut/indonesian-gec-gee>
- [12] J. Li, Q. Wang, C. Zhu, Z. Mao, and Y. Zhang, "Grammatical Error Correction via Mixed-Grained Weighted Training," in *Findings of EMNLP 2023*, pp. 6027-6037, 2023, doi: 10.18653/v1/2023.findings-emnlp.400.
- [13] T. Fang, X. Liu, D. F. Wong, R. Zhan, L. Ding, L. S. Chao, D. Tao, and M. Zhang, "TransGEC: Improving Grammatical Error Correction with Translationese," in *Findings of ACL 2023*, pp. 3614-3633, 2023, doi: 10.18653/v1/2023.findings-acl.223.
- [14] S. Mesham, C. Bryant, M. Rei, and Z. Yuan, "An Extended Sequence Tagging Vocabulary for Grammatical Error Correction," in *Findings of EACL 2023*, pp. 1608-1619, 2023, doi: 10.18653/v1/2023.findings-eacl.119.
- [15] H. Cao, Z. Cao, C. Hu, B. Hou, T. Xiao, and J. Zhu, "Improving Autoregressive Grammatical Error Correction with Non-autoregressive Models," in *Findings of ACL 2023*, pp. 12014-12027, 2023, doi: 10.18653/v1/2023.findings-acl.760.
- [16] A. Luhtaru, E. Korotkova, and M. Fishel, "No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models," in *Proc. EACL 2024*, pp. 1209-1222, 2024, doi: 10.18653/v1/2024.eacl-long.73.

- [17] M. R. Qorib, A. F. Aji, and H. T. Ng, "Efficient and Interpretable Grammatical Error Correction with Mixture of Experts," in Findings of EMNLP 2024, pp. 17127-17138, 2024, doi: 10.18653/v1/2024.findings-emnlp.997.
- [18] Y. Fei, L. Cui, S. Yang, W. Lam, Z. Lan, and S. Shi, "Enhancing Grammatical Error Correction Systems with Explanations," in Proc. ACL 2023, pp. 7489-7501, 2023, doi: 10.18653/v1/2023.acl-long.413.
- [19] Y. Song, K. Krishna, R. Bhatt, K. Gimpel, and M. Iyyer, "GEE! Grammar Error Explanation with Large Language Models," in Findings of NAACL 2024, pp. 754-781, 2024, doi: 10.18653/v1/2024.findings-naacl.49.
- [20] G. Peng, W. Li, W. Luo, and H. Wang, "Encode Errors: Representational Retrieval of In-Context Demonstrations for Multilingual Grammatical Error Correction," in Findings of ACL 2025, pp. 21166-21180, 2025, doi: 10.18653/v1/2025.findings-acl.1090.
- [21] W. Li, W. Luo, G. Peng, and H. Wang, "Explanation-Based In-Context Demonstrations Retrieval for Multilingual Grammatical Error Correction," in Proc. NAACL 2025, pp. 4881-4897, 2025, doi: 10.18653/v1/2025.naacl-long.251.
- [22] P. Bhattacharyya and A. Bhattacharya, "Leveraging LLMs for Bangla Grammar Error Correction: Error Categorization, Synthetic Data, and Model Evaluation," in Findings of ACL 2025, pp. 8220-8239, 2025, doi: 10.18653/v1/2025.findings-acl.431.
- [23] V. Dobariya, J. Baxi, B. Gambhava, and B. Bhatt, "Smruti: Grammatical Error Correction for Gujarati using LLMs with Non-Parametric Memory," in Findings of IJCNLP-AACL 2025, pp. 473-485, 2025, doi: 10.18653/v1/2025.findings-ijcnlp.28.
- [24] M. Dhamecha, S. Jaat, G. Damor, and P. Mishra, "Team Horizon at Bhasha Task 1: Multilingual IndicGEC with Transformer-based Grammatical Error Correction Models," in Proc. Bhasha 2025, pp. 142-146, 2025, doi: 10.18653/v1/2025.bhasha-1.14.
- [25] X. Chang and J. Zhu, "A Chunk-based Chain of Thought Prompting Method for Mitigating Over-Correction in Chinese Grammatical Error Correction," in Proc. CCL 2025, pp. 831-841, 2025.
- [26] Y. Wang, B. Wang, Y. Liu, D. Wu, and W. Che, "LM-Combiner: A Contextual Rewriting Model for Chinese Grammatical Error Correction," in Proc. LREC-COLING 2024, pp. 10675-10685, 2024.
- [27] X. Liu, B. Xu, M. Yang, H. Cao, C. Zhu, T. Zhao, and W. Lu, "A Chain-of-Task Framework for Instruction Tuning of LLMs Based on Chinese Grammatical Error Correction," in Proc. COLING 2025, pp. 8623-8639, 2025.
- [28] X. Wang, L. Mu, J. Zhang, and H. Xu, "Multi-pass Decoding for Grammatical Error Correction," in Proc. EMNLP 2024, pp. 9904-9916, 2024, doi: 10.18653/v1/2024.emnlp-main.553.
- [29] T. Goto, Y. Sakai, and T. Watanabe, "gec-metrics: A Unified Library for Grammatical Error Correction Evaluation," in Proc. ACL 2025: System Demonstrations, pp. 524-534, 2025, doi: 10.18653/v1/2025.acl-demo.50.
- [30] Y. Sakai, T. Goto, and T. Watanabe, "IMPARA-GED: Grammatical Error Detection is Boosting Reference-free Grammatical Error Quality Estimator," in Findings of ACL 2025, pp. 25647-25654, 2025, doi: 10.18653/v1/2025.findings-acl.1315.
- [31] T. Goto, Y. Sakai, and T. Watanabe, "Reliability Crisis of Reference-free Metrics for Grammatical Error Correction," in Findings of EMNLP 2025, pp. 24913-24926, 2025, doi: 10.18653/v1/2025.findings-emnlp.1356.
- [32] N. Varshney, S. Mishra, and C. Baral, "Towards Improving Selective Prediction Ability of NLP Systems," in Proc. 7th Workshop on Representation Learning for NLP, pp. 221-226, 2022, doi: 10.18653/v1/2022.repl4nlp-1.23.
- [33] P. Zablotskaia, D. Phan, J. Maynez, S. Narayan, J. Ren, and J. Liu, "On Uncertainty Calibration and Selective Generation in Probabilistic Neural Summarization: A Benchmark Study," in Findings of EMNLP 2023, pp. 2980-2992, 2023, doi: 10.18653/v1/2023.findings-emnlp.197.
- [34] Y. Ba, M. V. Mancenido, and R. Pan, "Fill In The Gaps: Model Calibration and Generalization with Synthetic Data," in Proc. EMNLP 2024, pp. 17211-17225, 2024, doi: 10.18653/v1/2024.emnlp-main.955.
- [35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, vol. 70, pp. 1321-1330, 2017.
- [36] B. Wilie, et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proc. AACL-IJCNLP 2020, pp. 843-857, 2020, doi: 10.18653/v1/2020.aacl-main.85.
- [37] S. Cahyawijaya, et al., "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," in Proc. EMNLP 2021, pp. 8875-8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.
- [38] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in Proc. EMNLP 2021, pp. 10660-10668, 2021, doi: 10.18653/v1/2021.emnlp-main.833.
- [39] R. Mahendra, A. F. Aji, S. Louvan, F. Rahman, and C. Vania, "IndoNLI: A Natural Language Inference Dataset for Indonesian," in Proc. EMNLP 2021, pp. 10511-10527, 2021, doi: 10.18653/v1/2021.emnlp-main.821.
- [40] S. Cahyawijaya, et al., "Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages," in Proc. ACL 2024, pp. 14899-14914, 2024, doi: 10.18653/v1/2024.acl-long.796.
- [41] T. M. Fahrudin, I. Sa'diyah, Latipah, I. Z. A. Illah, C. C. B. Lirna, and B. S. Acarya, "KEBI 1.0: Indonesian Spelling Error Detection System for Scientific Papers Using Dictionary Lookup and Peter Norvig Spelling Corrector," Lontar Komputer, vol. 12, no. 2, pp. 78-90, 2021, doi: 10.24843/LKJITI.2021.v12.i02.p02.
- [42] N. Mediyawati, J. C. Young, and S. B. Nusantara, "U-Tapis: Automatic Spelling Filter as an Effort to Improve Indonesian Language Competencies of Journalistic Students," Cakrawala Pendidikan, vol. 40, no. 2, pp. 402-412, 2021, doi: 10.21831/cp.v40i2.34546.
- [43] M. A. Mohsen, "Computer-Mediated Corrective Feedback to Improve L2 Writing Skills: A Meta-Analysis," Journal of Educational Computing Research, vol. 60, no. 5, pp. 1253-1276, 2022, doi: 10.1177/07356331211064066.
- [44] N. Zhai and X. Ma, "The Effectiveness of Automated Writing Evaluation on Writing Quality: A Meta-Analysis," Journal of Educational Computing Research, vol. 61, no. 4, pp. 875-900, 2023, doi: 10.1177/07356331221127300.
- [45] N. Fan, "Exploring the Effects of Automated Written Corrective Feedback on EFL Students' Writing Quality: A Mixed-Methods Study," SAGE Open, vol. 13, no. 2, 2023, doi: 10.1177/21582440231181296.
- [46] S. S. Lee and R. L. Moore, "Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review," Online Learning, vol. 28, no. 3, 2024, doi: 10.24059/olj.v28i3.4593.
- [47] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proc. COLING, pp. 757-770, 2020, doi: 10.18653/v1/2020.coling-main.66.
- [48] S. Swayamdiptha, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics," in Proc. EMNLP 2020, pp. 9275-9293, 2020, doi: 10.18653/v1/2020.emnlp-main.746.
- [49] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," Journal of Artificial Intelligence Research, vol. 70, pp. 1373-1411, 2021, doi: 10.1613/jair.1.12125.
- [50] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. Aroyo, "Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI," in Proc. CHI 2021, art. 39, 2021, doi: 10.1145/3411764.3445518.
- [51] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.
- [52] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, vol. 22, no. 158, pp. 209-212, 1927.
- [53] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," Psychometrika, vol. 12, pp. 153-157, 1947.
- [54] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.