

Natural, Synthetic, or Mixed Supervision for Open-Weight Arabic Grammatical Error Correction

Mohammed Alnuwaiser **Mohammed Albalawi** **Thikra Alghamdi**
ALBA7OOTH Research Lab ALBA7OOTH Research Lab ALBA7OOTH Research Lab
momoalnuw@gmail.com malbowi6@gmail.com thikramans1455@gmail.com

Abstract

Modern Standard Arabic grammatical error correction (GEC) remains underexplored for open-weight instruction-tuned large language models (LLMs). We compare natural, synthetic, and mixed training views for a highlighted-token correction task while holding the model, training-set size, and evaluation protocol fixed. We fine-tune Gemma 3 4B with QLoRA on three equal 2,000-record views and evaluate once on 511 Nahw-Passage records. Mixed supervision reaches 31.70% exact-match accuracy, compared with 20.55% for synthetic-only and 28.38% for natural-only; the mixed–natural difference is not established. Post-hoc robustness analyses preserve the mixed-over- synthetic direction, but checkpoint selection enlarges its measured size. Auxiliary diagnostics reveal a trade-off: synthetic- only changes fewer already-correct tokens, whereas mixed supervision retains more ArabicMMLU capability. The result concerns these particular data views: source also varies with domain, length, and error distribution; correction quality was not independently assessed. One model and strict single-reference scoring further limit generality.

1 Introduction

Grammatical error correction requires a model to produce a valid correction, not merely recognize a grammatical rule. This distinction is consequential for Modern Standard Arabic (MSA), whose rich morphology and agreement systems create substantial correction ambiguity. Arabic GEC research has benefited from the QALB shared tasks (Mohit et al., 2014; Rozovskaya et al., 2015), specialized sequence-to-sequence systems (Nagoudi et al., 2022; Alhafni et al., 2023), and recent prompting and instruction-tuning studies (Kwon et al., 2023). Nevertheless, the effect of *supervision source* is difficult to isolate when model families, corpus sizes, tasks, and evaluation protocols

vary together. Even at equal record count, comparing existing corpora cannot separate the origin of supervision from other corpus differences.

Nahw provides a recent benchmark spanning Arabic grammar understanding, error detection, correction, and explanation (Mubarak et al., 2026). Its fine-tuning study focuses on grammar understanding, leaving direct fine-tuning for its practical correction task open. We address that gap with a controlled extension to highlighted-token correction. Using one fixed Gemma 3 4B instruction-tuned LLM revision (Gemma Team, 2025), we compare equal-size natural-only, synthetic-only, and mixed training views while holding the training and evaluation interfaces fixed.

Our primary question is whether a fixed natural–synthetic mixture improves over size-matched synthetic-only supervision. We specify and freeze this comparison before either adapter accesses Nahw-Passage, retain row-level outputs privately, and reproduce the paired statistics from aligned predictions. We also report the earlier natural-only result and later prompt-only baselines as explicitly staged comparisons rather than presenting all five systems as one simultaneous experiment.

The study is a controlled case study rather than a new model architecture. Its central contribution is a matched comparison of three fixed training views under one task and one open-weight LLM. The source corpora also differ in domain, length, and error distribution, with potentially different label quality; the comparison does not identify a causal effect of supervision origin alone.

Our contributions are:

- a controlled GEC fine-tuning extension of Nahw using an open-weight instruction-tuned LLM;
- an equal-record comparison of natural, released-synthetic, and fixed mixed super-

vision with deterministic data selection and overlap controls;

- a frozen 511-record paired evaluation showing a large mixed-over- synthetic-only advantage;
- a post-hoc five-seed cohort in which the mixed arm exceeds the synthetic-only arm under dev-selected and matched fixed checkpoints; and
- an auditable, privacy-conscious protocol that records exact model, data, checkpoint, decoding, and artifact identities without publishing restricted corpus text or model responses.

2 Related Work

Arabic GEC. The QALB 2014 and 2015 shared tasks established standard Arabic text correction datasets and evaluation procedures (Mohit et al., 2014; Rozovskaya et al., 2015). Later work improved Arabic grammatical error detection and correction through pretrained Arabic models, morphology-aware components, and multitask learning (Alhafni et al., 2023). AraT5 demonstrated the utility of Arabic-focused text-to-text pretraining for generation tasks, including GEC (Nagoudi et al., 2022).

Recent Arabic text-editing work derives language-independent edit tags from training data, improving both correction quality and decoding efficiency (Alhafni and Habash, 2025). That work advances the editing formulation. Our question is complementary: with one model revision, fine-tuning recipe, record count, task interface, and evaluation fixed, how does the source composition of supervision affect correction?

Kwon et al. (2023) compare prompted instruction-following models with fully fine-tuned Arabic GEC systems and explore synthetic augmentation. Their results motivate both prompt-only baselines and synthetic supervision. On QALB-2014, they report M2 $F_{0.5}$ scores of 67.82 for five-shot GPT-4, 70.81 for an AraT5v2 baseline, and 75.27 after large-scale synthetic augmentation. These full-sentence edit scores are not numerically comparable to our token-level exact-match measure. Their systems also differ in architecture, training scale, and data source, so they do not provide a matched contrast between natural, synthetic, and mixed data.

More broadly, synthetic GEC data can be valuable but is not automatically equivalent to human-annotated supervision. Prior work identifies mismatched error distributions and noisy labels as obstacles to using synthetic data in data-limited fine-tuning, and improves results through contextual generation and relabeling (Wang et al., 2024). This motivates testing composition directly rather than treating all training pairs as exchangeable.

Arabic grammar benchmarks and synthetic data. Nahw evaluates grammar understanding, detection, correction, and explanation and finds substantial weaknesses across current models (Mubarak et al., 2026). The 511-error Nahw-Passage component supports highlighted-token correction, which is narrower than full-sentence QALB correction but permits tightly aligned paired evaluation.

Tibyan is a released synthetic Arabic GEC corpus created with ChatGPT-assisted generation and iterative linguistic review (Alrehili and Alhothali, 2024). Its release contains sentence pairs but no official split, stable identifiers, or record-level error labels. ARETA provides automatic Arabic error categories (Belkebir and Habash, 2021), but automatic labels are not expert gold. We therefore derive a deterministic highlighted-token view without assigning linguistic error types.

3 Datasets and Experimental Controls

Table 1 summarizes the source, released size, and experimental role of every dataset used in this study.

3.1 Task and final evaluation

Each example supplies an MSA passage and one highlighted erroneous token. The model must return only the corrected token. This interface resembles a proofreading assistant asked to replace a word already flagged by a writer or another system. It does not test error detection, whole-sentence editing, or effectiveness in a deployed writing tool. We use all 511 ordered Nahw-Passage correction records as a test-only set.

Illustrative task example (author-created; not corpus text). Passage: هؤلاء الطالب يكتبون الدرس. Highlighted token: الطالب. Required output: الطالب. English: “These students write the lesson.” The model receives the passage and highlighted token but must output only the replacement token, not a rewritten sentence or explanation.

Dataset	Source	Released size	Use in this study
QALB 2014 L1	Mohit et al. (2014)	19,411 train; 1,017 dev; 968 test	2,000 filtered train records for F1; the nested first 1,000 for F3; 975 filtered dev records for checkpoint selection. Test is unused.
Tibyan	Alrehili and Alhothali (2024)	6,192 sentence pairs	Deterministic highlighted-token conversion; 2,000 project-train records for F2; the nested first 1,000 for F3.
Nahw-Passage	Mubarak et al. (2026)	100 passages, 511 correction records	All 511 records form the single frozen final test; never used for training, prompting, or selection.
QALB 2015 L2 dev	Rozovskaya et al. (2015)	154 records	Reconstructed corrected targets for the matched already-correct-token diagnostic.
ArabicMMLU	Koto et al. (2024)	14,575 questions, 40 tasks	Fixed balanced 1,000-question auxiliary capability diagnostic (25 per task).

Table 1: Datasets, release sizes, and experimental roles. Restricted corpus text is not reproduced.

Our primary metric is *exact-match accuracy* (*EM*): the percentage of records for which the frozen parser’s predicted token exactly equals the single reference after permitted outer-format and whitespace removal. We apply no Arabic-letter, hamza, diacritic, or spelling normalization. Thus a linguistically plausible alternative receives no credit if it differs from the reference. We use “exact-match accuracy” consistently for this task; ArabicMMLU uses conventional multiple-choice accuracy.

Nahw-Passage is never used for training, prompt design, checkpoint selection, or development generation. The frozen prepared input identity is recorded in the public audit; all evaluated systems use the same ordered records.

3.2 Natural supervision

The natural view is derived from QALB 2014 L1 training data (Mohit et al., 2014). We retain annotator-0 actions that replace exactly one source token with one nonempty correction token, require the erroneous token to occur once, and retain at most one eligible action per document. A content-neutral SHA-256 ordering with seed 3407 selects 2,000 records. QALB development and test records are excluded. We use a separately selected 975-record QALB 2014 L1 development view only for assistant-token loss and technical validation. QALB test is not used for training, development, prompting, or evaluation.

3.3 Synthetic and mixed supervision

Tibyan releases 6,192 erroneous–corrected sentence pairs under CC BY 4.0 (Alrehili and Alhothali, 2024). Because a strict whole-pair single-substitution filter retains only three pairs, we compute minimum-edit alignments and retain a substitution only when it is forced across every optimal alignment. We require one occurrence of the erroneous token, create connected groups for repeated

exact sides, and retain at most one seeded representative per group. This yields 6,165 candidate pairs in 6,156 groups.

Tibyan has no official train/development split. We assign groups to an 80:20 project split by seeded SHA-256 buckets. Before splitting, exact UTF-8 hashes of both sides are compared with all registered QALB splits and with Nahw fields; any matching group would be excluded. This check finds no exact full-side overlap, although it cannot detect paraphrases or pretraining contamination.

The synthetic arm (**F2**) takes the first 2,000 eligible project-training representatives under a fixed hash order. The mixed arm (**F3**) combines the first 1,000 records in the frozen natural order with the first 1,000 in the frozen synthetic order. Thus both halves are nested subsets of their single-source arms rather than new samples. The natural arm (**F1**), F2, and F3 each contain exactly 2,000 records. We chose 50:50 as a simple balanced condition before final evaluation, not as an optimized ratio. It allows a size-matched contrast against either single-source view. We do not test other ratios or a 4,000-record union; the latter would also change training-set size.

4 Systems and Training

4.1 Model choice

We use Gemma 3 4B because it is an open-weight, multilingual, instruction-tuned LLM that supports the Arabic task interface while remaining small enough for 4-bit QLoRA on a free 16 GB P100. Fixing one public revision also makes supervision composition the principal controlled variable and keeps all three trainable arms computationally feasible. This practical choice is not evidence that Gemma is uniquely suitable for Arabic GEC; dependence on one model and scale limits external validity.

4.2 Prompt-only baselines

Both baselines use the same 4-bit Gemma 3 4B instruction-tuned LLM at the same fixed revision. **B1** prepends five QALB-training demonstrations selected by a content-neutral hash rule. Its Arabic query then says, in English translation: “Correct the specified erroneous word in the following text. Return only the corrected word, without explanation or quotation marks.” The exact query is `صح الكلمة الخاطئة المحددة في النص التالي. أعد الكلمة المصححة فقط`

دون شرح أو علامات اقتباس.

B2 uses no demonstrations and asks an Arabic-correction tool to determine the correct form from context, leave the rest unchanged, and return only that word: `أنت أداة متخصصة في تصحيح العربية الفصحى. راجع سياق النص لتحديد الصيغة الصحيحة للكلمة المحددة، مع إبقاء بقية النص دون تغيير. أعد الكلمة المصححة فقط دون شرح أو تعليق أو علامات اقتباس.`

The exact Arabic templates are public in the code; restricted demonstration and passage text is not reproduced. “Expert-style” names the prompt design, not expert validation.

4.3 QLoRA fine-tuning

We train fresh unmerged adapters with QLoRA (Dettmers et al., 2023). The 4-bit base, processor, Arabic task instruction, response-only objective, and maximum formatted length of 1,024 tokens are shared. LoRA uses rank 16, alpha 32, zero dropout, no bias, and the query, key, value, output, gate, up, and down projections. The adapters contain 32,788,480 trainable parameters over a 4.33-billion-parameter base.

Every arm trains for two epochs and 250 optimizer steps with per-device batch size 1, gradient accumulation 16, AdamW 8-bit, learning rate 2×10^{-4} , a linear schedule, 3% warmup, weight decay 0.01, and seed 3407. Packing is disabled. Using the pinned Gemma tokenizer on the complete formatted conversations, the per-epoch views contain 312,644 tokens for F1, 522,756 for F2, and 416,746 for F3 (156.32, 261.38, and 208.37 tokens per record). F2 has 25.4% more than F3; record matching therefore does not match token budgets, and greater exposure cannot explain F3’s advantage.

Figure 1 shows the development losses used for checkpoint selection. We evaluate checkpoints after each epoch on the same frozen 975-record natural development view. The rule chooses the lowest

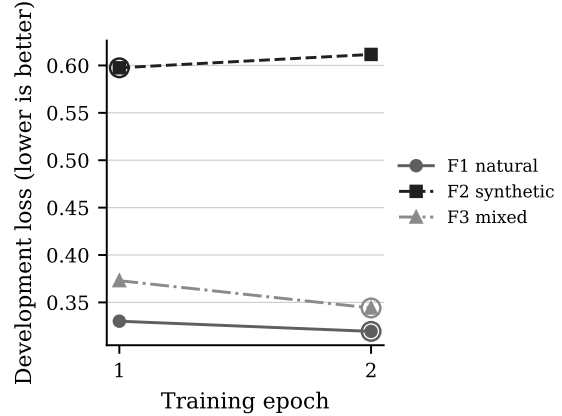


Figure 1: Development assistant-token cross-entropy across the two original training epochs. Lower is better; an open ring marks the checkpoint selected by the frozen common-development rule. These are selection losses, not final test scores.

mean assistant-token cross-entropy, breaking differences within 10^{-6} in favor of epoch 1. It selects epoch 2 for F1 and F3 but epoch 1 for F2: F2’s loss increases from 0.5975 to 0.6116, whereas F3’s falls from 0.3730 to 0.3441. The differing epochs thus follow one rule; neither Nahw nor generated-output exact-match accuracy influences selection.

4.4 Implementation and hardware

All primary F1–F3 training and B1/B2/F1–F3 final inference use Tesla P100 GPUs with the recorded 4-bit runtime. The separate post-hoc seed cohort trains on A100s and evaluates uniformly on RTX 3090s; the matched behavioral diagnostics use a P100. Exact package, device, revision, and artifact identities are recorded once in the public audit and repository rather than repeated throughout the paper.

5 Evaluation and Statistical Analysis

All systems use greedy decoding, no temperature argument, at most 32 new tokens, and seed 3407. We retain raw responses before applying a conservative parser. Empty outputs, parsing failures, and suspicious multi-token outputs are recorded rather than repaired.

For the pre-specified F3–F2 comparison, we report the difference in exact-match accuracy, an exact two-sided McNemar test, and a deterministic 10,000-sample paired bootstrap percentile interval with seed 3407. Both systems are aligned by the 511 frozen record identifiers. The interval is con-

System	Setting	Correct	EM (%)
B1	five-shot	89	17.42
B2	expert-style	108	21.14
F1	natural-only	145	28.38
F2	synthetic-only	105	20.55
F3	50:50 mixed	162	31.70

Table 2: Exact-match accuracy (EM) on 511 highlighted-token Nahw-Passage records. Prompt baselines are untouched; F1–F3 are equal-record QLoRA arms.

ditional on the single training seed and excludes re-training variance. Comparisons involving B1, B2, or F1 were executed or consolidated at different stages and are secondary. Their intervals and tests are descriptive rather than a simultaneous five-arm family.

This staging reflects the actual experimental timeline. F1 was trained and evaluated before the Tibyan-derived F2/F3 comparison was designed; the frozen one-access policy ruled out rerunning it merely to make a simultaneous table. F2 and F3 were then evaluated together as the primary matched contrast. B1 and B2 were added later to answer a separate prompt-only question and were frozen before their own test access. Consequently, only F3–F2 is treated as the primary simultaneous comparison.

We additionally report a post-hoc training-seed robustness cohort. Five independently trained F2/F3 replicas use seeds 3407–3411, the frozen training recipe, and the same checkpoint rule. All ten selected adapters are evaluated from record zero on one uniform hardware path. The seed-3407 replica does not reuse either original primary adapter. This cohort was run after the primary result and is therefore robustness evidence rather than a second pre-specified analysis.

Both epoch checkpoints are retained for every replica. In a further post-hoc sensitivity, we reuse the selected-checkpoint predictions and evaluate only the other checkpoint, then compare all arms at fixed epoch 1, fixed epoch 2, and the frozen dev-selected policy. This isolates whether selection on the common natural development view determines the direction of the result.

6 Results

Table 2 and Figure 2 show the final results. F3 obtains 162/511 exact matches (31.70%), compared with 105/511 (20.55%) for F2. F3 corrects 89

Comparison	Difference	95% interval
F3 – F2	+11.15	[7.05, 15.26]
F1 – F2	+7.83	[3.13, 12.52]
F3 – F1	+3.33	[−0.39, 7.05]
F3 – B2	+10.57	[6.65, 14.48]
F1 – B2	+7.24	[3.33, 10.96]

Table 3: Paired differences in exact-match accuracy, in percentage points. Only F3–F2 is the pre-specified primary comparison; the remaining rows are staged secondary comparisons.

records missed by F2, whereas F2 corrects 32 missed by F3. The primary paired difference is 11.15 points, with a 95% interval of 7.05–15.26 points and exact McNemar $p = 2.15 \times 10^{-7}$. This supports the specific mixed view over the specific synthetic view under the frozen development-selected protocol. The checkpoint policy contributes to the size of the gap, as Table 5 shows. Table 3 reports this primary comparison together with the staged secondary paired differences.

F1 reaches 28.38%, while F2 reaches 20.55%. In the staged comparison, F1 exceeds F2 by 7.83 points (interval 3.13–12.52). F3 is 3.33 points above F1, but its interval (−0.39 to 7.05) includes zero; we therefore do not claim that mixed supervision is better than natural-only supervision.

Prompt design matters within the displayed untouched-model baselines: B2 exceeds B1 by 3.72 points (interval 0.78–6.65). F1 and F3 exceed B2 in staged comparisons. F2 and B2 differ by only −0.59 points, with an interval spanning zero. All arms produce 511 aligned parsed predictions with no empty output or parsing failure. Suspicious or multi-token outputs number 2 for F1, 20 for F2, 2 for F3, and 6 for B2; they are retained unchanged when scoring. A post-hoc sensitivity takes the first whitespace-delimited token only for already flagged outputs and applies the rule symmetrically. It rescues 0/20 F2 outputs and 0/2 F3 outputs, leaving both scores and the 11.15-point gap unchanged. Format compliance therefore explains none of the primary difference under this sensitivity; the frozen parser remains primary.

Table 4 directly addresses training-seed sensitivity. F3 exceeds F2 in all five seeds; the paired difference ranges from 8.61 to 12.52 points. The mean difference of 10.29 points (SD 1.45) is close to the 11.15-point primary estimate. Because this cohort is post-hoc and uses a different but internally uniform hardware path, it supports robust-

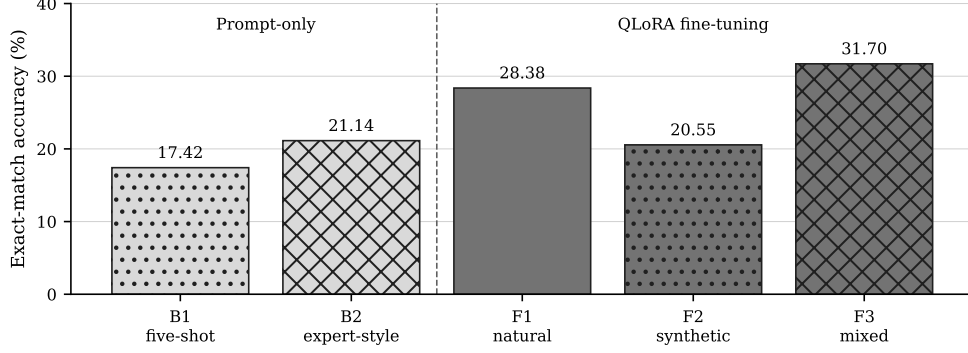


Figure 2: Final exact-match accuracy. Bar patterns distinguish protocols, not uncertainty. Confirmatory uncertainty is reported for the paired F3–F2 difference rather than as independent binomial error bars.

Training seed	F2	F3	F3 – F2
3407-A100	20.55	33.07	+12.52
3408	21.92	31.51	+9.59
3409	21.72	30.33	+8.61
3410	22.50	33.07	+10.57
3411	21.72	31.90	+10.18
Mean (SD)	21.68 (0.71)	31.98 (1.15)	+10.29 (1.45)

Table 4: Post-hoc five-seed robustness cohort. Values are exact-match accuracy percentages; SD is the sample standard deviation across seeds. All rows are independent retrains evaluated on one uniform hardware path; “3407-A100” is not the original primary run.

Checkpoint policy	F2	F3	F3 – F2
Fixed epoch 1	21.68	27.79	+6.11
Fixed epoch 2	25.40	31.98	+6.58
Dev-selected	21.68	31.98	+10.29

Table 5: Post-hoc five-seed checkpoint sensitivity. Values are mean exact-match percentages. F3 exceeds F2 in every seed under both fixed policies.

ness of the direction and magnitude without replacing the primary analysis. On seed 3407, the independent A100/RTX-3090 path gives F3 169/511 (33.07%), seven more exact matches than the original P100 run’s 162/511 (31.70%); F2 happens to give 105/511 on both paths. We verified these counts against the frozen aggregate audits. Cross-platform bitwise determinism was neither required nor achieved, so this row is a replication rather than a reproduction of the original adapter.

Table 5 shows that the mixed-over-synthetic direction is not created by the natural-development selection rule: the mean gap remains 6.11 points at fixed epoch 1 and 6.58 points at fixed epoch 2, with F3 higher in every seed under both policies. The rule selects epoch 1 for every F2 replica and epoch 2 for every F3 replica, expanding the dev-selected

gap to 10.29 points. The fixed-epoch differences are therefore the more conservative measure of this comparison; development selection affects its estimated magnitude. These post-hoc checks do not replace the frozen primary analysis.

7 Auxiliary Behavioral Diagnostics

Before the F1 diagnostic was run, we froze two matched checks and regenerated a fresh untouched-base control (**B0**) beside F1 within one run. This is distinct from the excluded historical T4 Nahw evaluation. First, on 154 corrected QALB-2015 L2 development sentences, each record designates one Arabic-script token already present in the corrected target. F1 leaves the designated token unchanged in 78/154 cases (50.65%), compared with 43/154 (27.92%) for B0. The paired difference is 22.73 points (95% interval 12.34–33.12). This indicates less overcorrection on this selected development diagnostic, not on arbitrary correct Arabic.

Second, we evaluate a balanced 1,000-question subset spanning all 40 ArabicMMLU tasks (Koto et al., 2024). B0 scores 53.7% and F1 53.1%; the -0.6 -point difference has a stratified paired-bootstrap interval of -3.2 to $+1.9$ points. This does not establish non-inferiority or general capability retention.

We subsequently apply the same frozen diagnostics to the original F2 and F3 adapters in one matched run. F2 leaves the designated correct token unchanged in 97/154 cases (62.99%), whereas F3 does so in 35/154 (22.73%). The primary F3–F2 difference is -40.26 points (95% interval -49.35 to -31.17 ; exact McNemar $p < 10^{-13}$). On ArabicMMLU, F2 scores 48.7% and F3 54.0%, a primary difference of $+5.30$ points (40-task stratified interval $+3.10$ to $+7.60$; $p < 10^{-5}$). Thus

synthetic-only supervision resists overcorrection better on this narrow diagnostic, while mixed supervision retains substantially more measured capability. B0/F1 comparisons remain staged references, and we make no broad safety claim.

8 Discussion

The primary result shows that the selected Tibyan-only view and the selected QALB–Tibyan view perform differently at fixed record count. The latter scores higher in all five post-hoc seeds, including under both fixed-epoch policies. QALB and Tibyan also differ in record length, token count, domain, and error distribution; their relative correction quality was not separately measured. The study cannot tell whether the gap arises from the natural origin of QALB examples, their closer match to the test task, data quality, a complementary mixture, or several of these. We have no expert error-type labels for the private outputs with which to identify a grammatical mechanism.

The F1–F2 secondary result also favors natural supervision, while the F3–F1 interval includes zero. The appropriate conclusion is therefore not that mixing is universally optimal. Rather, mixing substantially repairs the deficit of the synthetic-only view and reaches performance statistically compatible with the natural-only arm under this test. The interval does not show that synthetic additions improve on natural-only supervision. Other mixture ratios, a larger union, additional models, and expert-reviewed output samples would help test how broadly this pattern holds; none was evaluated here.

The prompt baselines sharpen this interpretation. A more explicit expert-style instruction improves the untouched model, but F1 and F3 remain well above it in secondary comparisons. Conversely, F2 is statistically compatible with B2. Fine-tuning alone is therefore not sufficient; supervision source remains consequential.

The auxiliary diagnostics add an important qualification: the mixed arm’s GEC advantage does not imply uniformly better behavior. Relative to F2, F3 trades substantially higher selected-token overcorrection for higher ArabicMMLU accuracy. The two measures cover different behaviors, so neither arm dominates the other outside the frozen correction test.

9 Reproducibility and Artifact Boundaries

We record the immutable base revision, selected-checkpoint identities, data selection hashes, exact prompt and parser versions, decoding configuration, package versions, hardware, private prediction hashes, and statistical seed. Every final aggregate is independently recomputed from aligned private predictions. Timeout-safe evaluation persists each row and resumes only after verifying the exact prefix, avoiding regeneration of completed predictions.

Code, protocols, corpus-free summaries, aggregate statistics, and hashes are available in the public research repository. QALB-derived records, Nahw text, record-level predictions, raw model responses, logs, checkpoints, and adapters are not redistributed under the applicable data and artifact restrictions. The published code lets a licensed custodian recompute every count and paired statistic from aligned private rows; published hashes establish which exact inputs, predictions, and checkpoints were audited. Hashes do not reveal the rows or permit an unlicensed reviewer to rescore them, so fully independent end-to-end replication remains limited. The upstream Tibyan release remains available under its own license. The public repository is <https://github.com/ALBA700TH-Research-Lab/Musahhih>.

10 Conclusion

In this equal-record Gemma 3 4B comparison, the selected 50:50 QALB–Tibyan view outperforms the selected Tibyan-only view on highlighted-token GEC. A post-hoc five-seed cohort preserves that direction at both matched epochs, where the mean gap is about six points; development selection expands it to 10.29 points. The mixed view is not established as better than natural-only supervision. The corpora’s many co-varying properties prevent attributing the result to supervision origin or mixing alone.

11 Limitations

Our evidence is limited to one 4B model, one equal-record training size, one 50:50 mixture, and one principal test set. Record matching does not equalize tokens, domains, error categories, or potential differences in label quality. The primary interval is conditional on seed 3407; the post-hoc cohort finds a positive paired difference in every seed but

uses a different, internally uniform hardware path. Although F2 contains more tokens than F3, token totals do not control length, domain, or source quality. We do not run QALB test, so our exact-match accuracy is not comparable to full-sentence M2 scores.

Every trainable arm selects between epoch 1 and epoch 2 using loss on the same natural QALB development view. The rule was frozen before test access, but a natural development distribution could still favor F3 over F2. In the post-hoc fixed-epoch sensitivity, F3 remains higher under both matched policies, but the rule selects F2 epoch 1 and F3 epoch 2 for all five replicas and expands the mean gap from roughly 6 points to 10.29. Thus check-point selection does not determine the sign here, but it materially affects the estimated magnitude.

Exact single-reference matching is intentionally strict. A linguistically valid alternative receives no credit unless it equals the supplied token. No qualified linguist reviewed the private predictions, so we provide no error-type or expert-quality claims. ARETA-derived aggregate evidence in the source literature does not substitute for expert labels on our outputs.

F1, F2/F3, and the prompt baselines were executed at different stages, so only the simultaneously executed F3–F2 contrast is primary. The earlier F1 and later prompt-only systems were frozen before their respective test access, but they cannot retrospectively become one simultaneous experiment. We exclude the historical T4 Nahw baseline and all comparisons derived from it. Hardware for retained primary, staged, robustness, and diagnostic runs is consolidated in the experimental setup rather than treated as an experimental variable.

Our exact-hash overlap checks do not detect paraphrases, shared upstream sources, or unknown pretraining contamination. The overcorrection diagnostic uses one designated token from corrected QALB development text and does not measure sentence-level preservation or all valid alternatives. ArabicMMLU is a selected balanced subset rather than a formal non-inferiority design. Finally, restricted artifacts prevent public release of the full training and prediction package.

12 Ethical Considerations

The work concerns text correction and does not involve participant recruitment or new human annotation. We preserve official dataset roles, keep

Nahw-Passage test-only, and do not use QALB test for modeling or evaluation. QALB-derived material is kept private under its research conditions; Tibyan attribution and CC BY 4.0 terms are preserved; ArabicMMLU questions remain private under CC BY-NC 4.0.

Automated correction can encode prescriptive language norms and should not be used to stigmatize dialectal Arabic or writers. Our task and claims concern MSA highlighted-token correction only. “Expert-style” refers solely to a prompt design, and we make no claim of expert linguistic validation. Adapters trained on restricted natural data remain private pending legal, institutional, and model-license review.

References

- Bashar Alhafni and Nizar Habash. 2025. [Enhancing text editing for grammatical error correction: Arabic as a case study](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17892–17914, Vienna, Austria. Association for Computational Linguistics.
- Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. [Advancements in Arabic grammatical error detection and correction: An empirical investigation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.
- Ahlan Alrehili and Areej Alhothali. 2024. [Tibyan corpus: Balanced and comprehensive error coverage corpus using ChatGPT for Arabic grammatical error correction](#). *arXiv preprint arXiv:2411.04588*.
- Riadh Belkebir and Nizar Habash. 2021. [Automatic error type annotation for Arabic](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Gemma Team. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

- Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [Beyond English: Evaluating LLMs for Arabic grammatical error correction](#). In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.
- Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. [The first QALB shared task on automatic text correction for Arabic](#). In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.
- Hamdy Mubarak, Majd Hawasly, and Abubakr Mohamed. 2026. [Nahw: A comprehensive benchmark of Arabic grammar understanding, error detection, correction, and explanation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6310–6328, Rabat, Morocco. Association for Computational Linguistics.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. [AraT5: Text-to-text transformers for Arabic language generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.
- Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. [The second QALB shared task on automatic text correction for Arabic](#). In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.
- Yixuan Wang, Baoxin Wang, Yijun Liu, Qingfu Zhu, Dayong Wu, and Wanxiang Che. 2024. [Improving grammatical error correction via contextual data augmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10898–10910, Bangkok, Thailand. Association for Computational Linguistics.