ThaiEditBench: A Mechanism-Typed, Length-Stratified Benchmark for Thai Orthographic Error Correction

Nutchanon Yongsatianchot **Piyalitt Ittichaiwong** **Kanyakorn Veerakanjana**
Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
{nutchanon.yon, piyalitt.itt, kanyakorn.vee}@mahidol.ac.th

Abstract

Thai is written without spaces between words, and its spelling depends on tone marks, silent consonants, and the transliteration of many loanwords. Each of these is a common source of misspellings. We present **ThaiEditBench**, a benchmark for correcting Thai spelling errors at the character level. It has three components: (i) a **six-class taxonomy** that types each error by the part of the spelling that goes wrong, plus an etymology tag; (ii) **three length tiers** (sentence, paragraph, and page) that test whether a model stays accurate over long, mostly correct text; and (iii) a **contamination-robust construction**, in which naturally attested misspellings are *injected* into clean text, so the corrupted passages do not appear verbatim in the source corpora. Because the errors are injected, the gold corrections are exact, and a blind two-annotator study measures how reliably the taxonomy can be applied. Grading is at the level of Thai character clusters, with span-based matching adapted from ERRANT. We evaluate **16 models** and find that (1) under the benchmark's conditions, a model that locates an error almost always fixes it correctly, so what separates models is *detection*; (2) the evaluated Thai-adapted models do not lead the strongest general models; and (3) in long, mostly correct passages, weaker models accumulate over-edits that pull their scores far below their sentence-level results. On real errors mined from Wikipedia edit history, scores are somewhat lower, but the model ranking largely holds.

1 Introduction

Thai spelling goes wrong in ways that are specific to its script. Writers choose the wrong tone mark, add or drop the mark that silences a consonant, confuse consonants or vowel forms that sound alike, reorder the letters of a consonant cluster, and transliterate loanwords inconsistently. For instance, สังเกต ('to observe') is often written สังเกตุ, with an extra vowel mark. These are **character-level** errors, and because Thai has no spaces between words, they are hard to locate and score with word-based tools. However, the resources to study them are missing. Grammatical error correction (GEC) benchmarks are largely English (Bryant et al., 2017, 2019; Napoles et al., 2017), and the closest character-level task is Chinese spelling check (CSC) (Tseng et al., 2015; Liu et al., 2011). Thai error-annotated resources cover social-media, informal, or learner text (Lertpiya et al., 2020; Nakwijit and Purver, 2022). Document-level GEC studies correction over long context (Yuan and Bryant, 2021), but no benchmark tests how Thai spelling correction holds up over long passages that contain few errors.

This work addresses this gap with a formal-register benchmark that types each error by how it arises and that limits contamination. We score at the level of the Thai Character Cluster (TCC) (Theeramunkong et al., 2000), a group of characters, such as a consonant with its vowel and tone marks, that Thai writing rules do not allow to be split. Our work contributes the following. (1) **ThaiEditBench**, a character-level benchmark for Thai spelling correction. It types each error with one of six *mechanism* classes, the part of the spelling that goes wrong, plus an etymology tag, and it has **three length tiers** (sentence, paragraph, and page). It is **contamination-robust by construction**: we inject naturally attested misspellings into clean text, so the corrupted passages are new and the gold corrections are exact. A blind annotation study measures how reliably the taxonomy can be applied (Krippendorff's $\alpha = 0.75$). (2) An **ERRANT-style edit-typer for Thai**. It segments the difference between an input and its correction into TCC-level edits and assigns each edit one of the six classes where one applies. This lets us report **span-F0.5** separately for each error type, as ChERRANT does for Chinese (Zhang et al., 2022). We use it to score and diagnose models on this benchmark, not as a general Thai grammatical error classifier. (3) A

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 211–225
November 6-10, 2026 ©2026 Association for Computational Linguistics

16-model evaluation. Models that find an error almost always fix it, so detection separates them. The evaluated Thai-adapted models do not lead, weaker models accumulate over-edits in long text, and the ranking largely holds on naturally occurring errors (Spearman $\rho = 0.81$). We release the code, data, grader, edit-typer, and variant map at `https://github.com/yongsa-nut/ThaiEditBench`.

2 Related Work

GEC and CSC. English grammatical error correction is evaluated with span-based metrics. These metrics represent a correction as a set of edits, each a span of the input and its replacement, and credit a model for each gold edit it reproduces with the same replacement. The M^2 scorer matches a model's edits against gold edits (Dahlmeier and Ng, 2012). ERRANT (the ERRor ANnotation Toolkit) extracts the edits automatically by aligning the input with each correction, assigns each edit an error type, and reports span-F0.5, an F-score over matched edits that weights precision twice as much as recall (Bryant et al., 2017). GLEU instead measures fluency (Napoles et al., 2015). We adapt ERRANT-style matching to TCC units (Section 4.2). Character-level Chinese spelling check is the closest task, scoring position-based detection and correction F1 without word boundaries (Tseng et al., 2015; Wu et al., 2013; Yu et al., 2014), with confusion-set construction (Liu et al., 2011) and the MuCGEC/ChERRANT dual-granularity scorer (Zhang et al., 2022). ERRANT has been ported to Arabic (Belkebir and Habash, 2021) and Korean (Yoon et al., 2023), but not Thai.

Thai NLP. We build on PyThaiNLP for tokenization and dictionary tooling (Phatthiyaphaibun et al., 2023) and on TCC segmentation (Theeramunkong et al., 2000). Prior Thai spelling correction targets social text (Lertpiya et al., 2020, 2023), including the VISTEC-TP-TH-2021 Twitter corpus with misspelling and word-segmentation annotation (Limkonchotiwat et al., 2021), and Nakwijit and Purver (2022) study misspelling semantics.

3 The ThaiEditBench Benchmark

3.1 Task

Given a Thai passage containing character-level orthographic errors, return the passage with those errors corrected and **nothing else changed**: no rewriting, re-spacing, numeral, or punctuation edits. We score under a strict **minimal-edit contract**, where any change to a correct span is an *over-edit* and is penalized.

3.2 Mechanism taxonomy

We define six orthographic mechanism classes, each carrying an **etymology tag**: (1) tone mark, (2) consonant-pair, (3) silent-consonant, (4) vowel, (5) cluster or metathesis, and (6) loanword transliteration. Worked examples per class are in Appendix A. We check canonical forms against the Royal Institute Dictionary (Royal Institute, 2013) (Appendix C). Class 6 is governed by a **loanword-absorption rule**: only *contemporary* foreign loans are class 6, while long-integrated borrowings (Tamil, Khmer, Persian, and Sanskrit–Pali doublets) count as native.

3.3 Construction

- **(a) Confusion-set injection.** The confusion set holds 1,044 wrong→correct pairs across the six classes. Every wrong form is a **naturally attested misspelling**, harvested from the Thai Wiktionary appendix of commonly misspelled words and from fixes made by a Thai Wikipedia spelling-cleanup bot; a rule-based classifier assigns each pair a mechanism class (83 of 86 hand-labeled pairs, 96.5%; Appendix C). Only the placement is synthetic: we inject pairs into *new clean text* by matching **whole tokens**, which protects short valid forms, and record each error's TCC span, class, and etymology. Applying the gold edits restores the clean original for 100% of items, so the gold is exact.

- **(b) Cleanup-bot mining.** Real errors in their original sentences are mined from the same bot's edit history, keeping only fixes whose corrected word is in the Royal Institute Dictionary or PyThaiNLP word list and whose wrong and correct forms align cleanly, following the revision-mining tradition (Tanaka et al., 2020; Grundkiewicz and Junczys-Dowmunt, 2014), with the bot's fix as gold. These items form a separate natural-error evaluation set (Section 5.5).

Registers. The clean text of the evaluated tiers spans news (ThaiSum), encyclopedic text (Thai Wikipedia), and Thai government press releases, all permissively redistributable.

3.4 Contamination control

Models are evaluated on **corrupted passages that do not appear verbatim in the source corpora**. A residual risk remains: a model that memorized a clean source passage could reproduce it instead of finding and fixing the injected errors. We therefore probe memorization directly. Given the first half of 30 sentences per source, Typhoon 2.5 (30B) continues them verbatim for only 0.3% to 1.7% of the true second half on average, and never for half of it. We fill a separate held-out test split first from the least-memorized sources (details in Appendix C).

3.5 Length tiers

The same pipeline produces three tiers that vary passage length and error sparsity. **T1** (sentence) has 774 items with a mean length of 100 characters (95% of items fall between 43 and 174); **T2** (paragraph) has 224 items with a mean of 425 (270 to 577); and **T3** (page) has 122 items with a mean of 1,242 (963 to 1,530). They contain 949, 278, and 193 gold edits respectively. Injection is stratified by class quota rather than uniform: the class mix follows the skew of the attested pairs (the per-class share of T1 gold edits correlates with that of the attested pairs at Spearman $\rho = 0.83$) and is partially flattened (after the audit, every class has at least 112 T1 gold edits; Table 3). The error budget is **dense at the sentence level**, with roughly one error per 100 characters, and deliberately **sparse at paragraph and page length**, where most sentences are clean distractors. T2 and T3 therefore measure **precision under long correct context** (over-editing).

3.6 Quality control

Construction makes the correction of each injected error exact, but it cannot rule out two other problems: typos already present in the source text, and words with two accepted spellings. We add one check for each.

- **Cross-model consensus audit.** A typo left in the source text stays in the gold. An item is flagged when, among the models' full outputs, the most common one other than the gold comes from at least three models and from at least as many as return the gold. A flagged item is kept if that output is the unchanged input or makes only some of the gold corrections (the models missed an error), or if it differs from the gold only in punctuation or spacing. Every other flagged item is dropped. The same rule is applied to all three tiers and removed 45 of 819 (T1), 26 of 250 (T2), and 28 of 150 (T3) items. They include source typos and non-standard spellings, but also items where the models agree on a register change or on a wrong form (Appendix B).
- **Variant reconciliation.** Before scoring, we map words with two accepted spellings (four groups, eight word forms; Appendix B) to the gold's form. No injected wrong form belongs to the map, so the map cannot turn a missed error into a fix; it changes 2.35% of non-empty model outputs.

3.7 Human validation of the labels

The mechanism classes are assigned automatically. To measure how reliably the taxonomy can be applied, two native Thai speakers independently labeled 60 items (40 curated mined items and 20 synthetic ones), blind to the automatic label. Both judged all 60 marked spans to be genuine errors, and their class agreement is Krippendorff's $\alpha = 0.752$ (Krippendorff, 2011). Where the raters agree (48 items), their class matches the automatic label on 35; most disagreements fall at the loanword-absorption and class-2 conventions (Appendix E).

4 Evaluation Protocol

4.1 Metrics

We grade at the TCC level with ERRANT-style span matching (Bryant et al., 2017). The **primary metric is correction F1**. We derive minimal TCC edits by aligning the wrong text with *both* the gold and the model output under the **same** Levenshtein alignment, so an output identical to the gold scores exactly 1.0, as in SIGHAN CSC scoring (Tseng et al., 2015). **Detection F1** credits the correct error site and ignores the replacement. **Clean-sentence fidelity** is the fraction of error-free sentences left unchanged, a precision signal for T2 and T3. The edit-typer (Section 4.2) also gives **span-F0.5**, overall and per mechanism. We report 95% item-bootstrap confidence intervals (Appendix D) and test differences between models with a **paired item-level bootstrap** (Koehn, 2004) over 10,000 resamples.

4.2 ERRANT-style edit-typer

We adapt the ERRANT/ChERRANT design (Bryant et al., 2017; Zhang et al., 2022) to Thai orthography. Given a (wrong, correct) pair, the typer works in three steps. First, it extracts edits at

the **TCC level**, with the same alignment the grader uses. Second, an edit often covers only part of a word, such as a single tone mark, so the typer widens each edit to the whole word that contains it, found with the dictionary-based word segmenter of PyThaiNLP (`newmm`; Phatthiyaphaibun et al., 2023). Third, a rule-based classifier compares the wrong and correct forms of that word character by character and assigns the class from what changed; for example, a changed tone mark gives class 1. If the corrected word is in a loan lexicon, the edit is class 6 (Appendix A gives the full rules). The same classifier labels the confusion set, and the typer also tags the edit operation (substitution, insertion, deletion, transposition, or mixed). Because it types both the gold and the model edits, it gives **span-F0.5 per mechanism** on any (wrong, correct) pair, without construction labels. On the dev split, it recovers 98.9% of the gold edit spans and agrees with the construction class on 98.1% of them. Because the typer reuses the construction-time classifier, we treat it as a **scoring instrument** for this benchmark, not a general Thai grammatical error classifier.

4.3 Models and inference

We evaluate 16 models, each in a single fixed configuration, at temperature 0 where supported, with one fixed Thai instruction prompt. The **frontier proprietary** models are GPT-5.5 (medium and high reasoning), Gemini 3.5 Flash, Claude Opus 4.7 (non-thinking), Claude Sonnet 4.6 (non-thinking), and GPT-5.4-mini. The **open-weight** models are Gemma-4-31B, DeepSeek V4 Pro and Flash, GLM-5.1, MiniMax-M2.7, and Qwen3.6-35B-A3B. The **Thai-adapted** models, which adapt general base models to Thai, are Typhoon 2.5 (30B) and three ThaiLLM 8B-instruct models served through the national ThaiLLM gateway, namely OpenThaiGPT (v7.2), Typhoon-S, and THaLLE (Pipatanakul et al., 2024; Pipatanakul and Taveekitworachai, 2026). Appendix G gives the prompts and a prompt-sensitivity ablation.

5 Results and Analysis

5.1 Leaderboard

Table 1 reports correction F1 on the three tiers and on the natural-error set (post-audit $n = 774/224/122/86$); the full table is in Appendix D. The frontier clusters at 0.96 to 0.98 on T1, and the confidence intervals separate three strata: the frontier and the strongest open-weight models, then GLM-5.1, then the Thai-adapted models, MiniMax, and Qwen. Within the frontier, T1 differences are mostly not significant: GPT-5.5 (high) is tied with GPT-5.5 (med) and Claude Opus 4.7.

	correction F1				fidelity
model	T1	T2	T3	Nat.	T3
GPT-5.5 (high)	**0.981**	**0.966**	0.855	0.862	0.971
GPT-5.5 (med)	0.978	0.953	0.857	0.862	0.973
Claude Opus 4.7	0.975	0.948	0.853	0.885	0.976
Gemini 3.5 Flash	0.975	0.960	0.825	0.871	0.963
Gemma-4-31B	0.961	0.954	**0.898**	**0.925**	0.984
GPT-5.4-mini	0.959	0.915	0.865	0.924	**0.990**
Claude Sonnet 4.6	0.957	0.905	0.825	0.898	0.974
DeepSeek V4 Flash	0.956	0.945	0.840	0.854	0.971
DeepSeek V4 Pro	0.939	0.926	0.754	0.850	0.958
GLM-5.1	0.856	0.811	0.657	0.795	0.954
Typhoon-S 8B†	0.792	0.702	0.617	0.708	0.964
THaLLE 8B†	0.760	0.594	0.507	0.744	0.814
Typhoon 2.5 30B†	0.754	0.632	0.476	0.753	0.899
OpenThaiGPT v7.2†	0.753	0.476	0.463	0.722	0.901
MiniMax-M2.7	0.749	0.720	0.471	0.725	0.970
Qwen3.6-35B-A3B	0.709	0.384	0.310	0.589	0.838

Table 1: Correction F1 by tier and on the natural-error set (Nat.), and T3 clean-sentence fidelity (the fraction of error-free sentences left unchanged). **Bold** = column best; underlined = correction F1 tied with the best (paired bootstrap, $p \geq 0.05$); because the test is paired, a lower score can be tied while a higher one is not. † = Thai-adapted.

General frontier models dominate, while the Thai-adapted models trail (0.75 to 0.79 on T1 and 0.46 to 0.62 on T3). The strongest **open-weight** model is a general 31B model, Gemma-4. The frontier leads it on T1 ($\Delta = 0.020$ against GPT-5.5 high, $p < 0.001$), and on T3 it has the highest point estimate (0.898) but is statistically tied with the top frontier models (against GPT-5.5 med, $\Delta = +0.041$, 95% CI $[-0.004, 0.085]$). Under this setup the evaluated Thai-adapted models do not lead, and the gap widens with length (Gemma-4 against Typhoon 2.5 on T3: $\Delta = 0.422$, $p < 0.001$). This comparison is confounded by base model, scale, serving stack, and inference settings, so it is not evidence about Thai adaptation in general.

5.2 Where models fail: finding versus fixing

We decompose every gold edit into **fixed**, **miscorrected** (the model edits exactly the gold span but with a different replacement), and **missed** (no model edit matches the gold span). Pooled over all 16 models, 0.873 of T1 gold edits are fixed, only 0.006 are miscorrected, and 0.122 are missed. The shares are 0.826, 0.005, and 0.168 at T3 and 0.820, 0.006, and

0.175 on natural errors (Section 5.5), and no single model's miscorrected share exceeds 0.032 on T1 (Table 6, Appendix D). The pattern holds within every mechanism class, including the hardest (loanword: miscorrected 0.003 against missed 0.186). A model that rewrites a wider region around an error is scored as missed, but such edits are only 1.2 to 1.8% of gold edits per tier, so untouched errors still outnumber wrong fixes at least five to one (four to one without empty replies; Limitations). Under this benchmark's conditions (formal register, single-reference gold, and one to three errors per passage), models rarely fail by choosing the wrong spelling once they locate an error, so what the leaderboard separates is **detection**. Spelling knowledge still matters, since recognizing a misspelling requires knowing the correct form; whether the finding holds for dense, interacting errors is open (Limitations).

5.3 Length robustness

Sentence-tier scores largely predict the page-tier ranking (Spearman $\rho = 0.90$), but the **drop from T1 to T3** differs widely between models (Figure 2a). The frontier loses only 0.09 to 0.15 correction F1 (GPT-5.5-med $0.978 \rightarrow 0.857$) and Gemma-4 only 0.06 ($0.961 \rightarrow 0.898$), while Typhoon 2.5 and Qwen lose 0.28 and 0.40 ($0.754 \rightarrow 0.476$ and $0.709 \rightarrow 0.310$). Measured per 10,000 clean TCCs, the over-edit rate is lower at T3 than at T1 for 15 of the 16 models (Figure 2b), although it rises from T2 to T3 for 11 of them, so over-edits per passage grow mainly with the amount of correct text. What separates models is the *level* of that rate: Typhoon 2.5, OpenThaiGPT, and Qwen over-edit at 4 to 30 times the rate of GPT-5.5 (med) at every tier (Appendix D). Clean-sentence **fidelity** agrees: the frontier holds at or above 0.95, while the heaviest over-editors sit at 0.77 to 0.90. These results suggest that **over-editing and missing reflect the same weak discrimination**. At T3, even the frontier models are limited more by over-editing than by detection: they fix 0.95 to 0.98 of gold edits, but their precision is 0.71 to 0.76 (GPT-5.4-mini, with recall 0.85 and precision 0.88, is the exception).

5.4 Per-mechanism difficulty

The edit-typer types the model's edits as well as the gold's, so it gives span recall, precision, and F0.5 for every mechanism and model (F0.5 in Figure 1; recall and precision in Table 8). At the sentence level most cells are above 0.9, and the weak cells belong to particular models. The four Thai-adapted

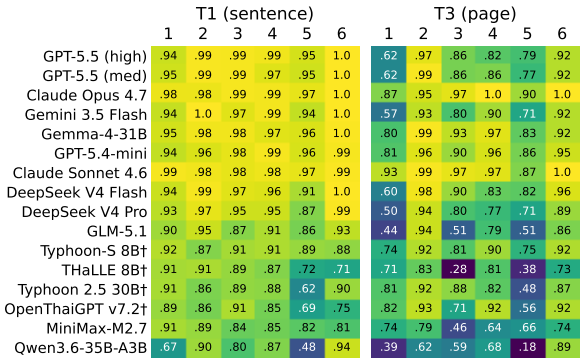


Figure 1: Span-F0.5 per mechanism class at T1 (sentence) and T3 (page); rows follow Table 1. Classes: 1 tone mark, 2 consonant-pair, 3 silent-consonant, 4 vowel, 5 cluster, 6 loanword. † = Thai-adapted.

models and MiniMax find only 0.33 to 0.71 of loanword errors, against 0.95 to 1.00 for the nine models above GLM-5.1 in Table 1. Cluster edits (class 5: reordered, doubled, or inserted letters, and other consonant changes) have the lowest precision for 12 of the 16 models. At page length the weakness is shared: tone marks (class 1) or clusters are the weakest class for 14 of the 16 models. Their median precision falls from 0.93 and 0.91 at T1 to 0.73 and 0.73 at T3, while median recall falls only from 0.93 and 0.95 to 0.88 and 0.89: in long text, models mostly change tone marks and consonants in correct words rather than miss errors of these types. For GPT-5.5 (high), most such edits respell loanwords to the Royal Institute standard, formalize informal words, or fix typos left in the source (Appendix D).

5.5 Natural errors

We evaluate all 16 models on the mined cleanup-bot items, each a real error in its original encyclopedic sentence. The consensus audit of Section 3.6 removes 17 of the 103 items, where the agreeing models correct more than the recorded bot fix does (Appendix F). On the remaining 86 items, mean correction F1 is 0.81 against 0.88 on T1, and the model ranking transfers (Spearman $\rho = 0.81$ with T1; Table 1).

6 Conclusion

We presented ThaiEditBench, a mechanism-typed, length-stratified benchmark for Thai spelling correction. Across 16 models, **detection** rather than the choice of spelling separates models, and weaker models accumulate over-edits in long text.

Limitations

Sparse error density (primary). The long tiers (T2 and T3) are deliberately sparse, with 1 to 3 errors scattered across otherwise-clean passages, to isolate precision under long correct context. We therefore do **not** test **high error-density** regimes such as heavy-typo input, degraded OCR, or L2 writing, where errors co-occur, interact, mask one another, and correction order matters. Extending ThaiEditBench to dense multi-error passages, and testing whether the detection finding survives when errors crowd each other, is the most important next step.

A fixed inventory of attested errors. Every injected error comes from 1,044 attested pairs, each mapping one wrong form to one correct form. The benchmark therefore does not cover misspellings outside the inventory, and a model that consulted the released inventory would be answering a different, tool-augmented question; all models are evaluated zero-shot without it. The class mix follows the distribution of attested misspelling types, not the frequency of errors in running text.

Scope of the natural-error check. The audited natural set has 86 items from one cleanup bot in one register, with only 6 loanword and 4 class-5 edits, so it supports the ranking and the decomposition but not per-class conclusions; loanword behaviour is measured mainly on injected errors. The audit relies on cross-model consensus, and the bot's fix is taken as gold, so the natural gold is less certain than the construction gold of the tiers. A larger, naturally distributed test set, which the released mining pipeline supports, is the main way to strengthen this check.

Source spellings in the gold. The gold keeps the clean source as written, including common spellings outside the Royal Institute standard, such as many loanwords. A model that respells these words, or fixes a source typo the consensus audit missed, is scored as over-editing. The audit needs several models to return the same whole passage, so it catches fewer source typos in long passages; for GPT-5.5 (high), most of its unsupported tone and cluster edits at page length are of these kinds (Appendix D). Conversely, the audit's automatic rule drops some items where the models agree on a register change or on a wrong form (Appendix B). Precision at T2 and T3, and that of the strongest models in particular, should be read with this in mind.

Label reliability. The annotation study covers 60 items and two raters. Per-class agreement on ten items per class is noisy (α from 0.49 to 1.00) and lowest for class 6, whose boundary depends on the absorption convention, and the automatic class matches the rater consensus on 35 of 48 items. Per-mechanism results near the class 6 and class 2 boundaries should be read with this uncertainty.

Configured-model comparison. Each model is evaluated under one fixed configuration and one fixed Thai prompt. Reasoning modes, token budgets, provider defaults, serving stacks, and model scale are not controlled, so the leaderboard reflects configured-model behaviour rather than an intrinsic ranking of base-model capability. A prompt ablation on two models shifts correction F1 by at most 0.07 without reordering them (Appendix G), but we do not sweep prompts across all models. A reply that exhausts its token budget is empty and scored as the unchanged input, so its errors count as missed. This affects 110 of 774 T1 items for GLM-5.1, 91 for MiniMax, and 55 for Qwen, and 51 of 122 T3 items for MiniMax; empty replies account for 18% of missed edits at T1 and 27% at T3. For these models, a low score partly reflects failure to deliver an answer rather than failure to detect errors, and excluding empty replies, untouched errors outnumber wrong fixes four to one rather than five to one. Budgets also differ: GPT-5.5 runs with 3k or 6k tokens against 24k for the open reasoning models; its 22 replies that exhausted that budget (11 of them on T3) were rerun with 24k tokens. Appendix G covers server errors and THaLLE's partial replies.

Acknowledgments

We thank Pachaya Sailamul for help with the annotation study.

References

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

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In Adam Przepiórkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing – Lecture Notes in Computer Science, Vol. 8686 (PolTAL 2014)*, pages 478–490. Springer.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. Departmental Papers (ASC) 43, Annenberg School for Communication, University of Pennsylvania.

Anuruth Lertpiya, Tawunrat Chalothorn, and Pakpoom Buabthong. 2023. How to progressively build Thai spelling correction systems? *IEEE Access*, 11:72704–72716.

Anuruth Lertpiya, Tawunrat Chalothorn, and Ekapol Chuangsuwanich. 2020. Thai spelling correction and word normalization on social text using a two-stage pipeline with neural contextual attention. *IEEE Access*, 8:133403–133419.

Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Handling cross- and out-of-domain samples in Thai word segmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1003–1016, Online. Association for Computational Linguistics.

Chao-Lin Liu, Min-Hua Lai, Kan-Wen Tien, Yi-Hsuan Chuang, Shih-Hung Wu, and Chia-Ying Lee. 2011. Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing*, 10(2):10:1–10:39.

Pakawat Nakwijit and Matthew Purver. 2022. Misspelling semantics in Thai. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 227–236, Marseille, France. European Language Resources Association.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

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

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Singapore. Association for Computational Linguistics.

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. 2024. Typhoon 2: A family of open text and multimodal Thai large language models. *Preprint*, arXiv:2412.13702.

Kunat Pipatanakul and Pittawat Taveekitworachai. 2026. Typhoon-S: Minimal open post-training for sovereign large language models. *Preprint*, arXiv:2601.18129.

Royal Institute. 2013. *Photchananukrom Chabap Ratchabandittayasathan Pho. So. 2554 [The Royal Institute Dictionary, B.E. 2554]*. Nanmeebooks Publications, Bangkok. In Thai. ISBN 9786167073569.

Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2020. Building a Japanese typo dataset from Wikipedia's revision history. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 230–236, Online. Association for Computational Linguistics.

Thanaruk Theeramunkong, Virach Sornlertlamvanich, Thanasan Tanhermhong, and Wirat Chinnan. 2000. Character cluster based Thai information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages (IRAL 2000)*, pages 75–80, Hong Kong, China. Association for Computing Machinery.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings*

of the Eighth SIGHAN Workshop on Chinese Language Processing, pages 32–37, Beijing, China. Association for Computational Linguistics.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42, Nagoya, Japan. Asian Federation of Natural Language Processing.

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean grammatical error correction: Datasets and annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132, Wuhan, China. Association for Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Online. Association for Computational Linguistics.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

A Mechanism taxonomy in detail

Each class is defined by the orthographic *mechanism* of the fault, not its surface character. Multiple worked examples per class:

- **1 tone mark:** ค๊ะ→คะ, น๊ะ→นะ, จ๊กจั่น→จักจั่น, กระทั้ง→กระทั่ง
- **2 consonant-pair:** กฎหมาย→กฎหมาย, กบฏ→กบฏ, อากาส→อากาศ
- **3 silent-consonant:** กลยุทธ→กลยุทธ์, ขาดดุลย์→ขาดดุล, เท่ห์→เท่
- **4 vowel:** สังเกตุ→สังเกต, มาตราฐาน→มาตรฐาน, ฉะเพาะ→เฉพาะ
- **5 cluster/metathesis:** กระทันหัน→กะทันหัน, กระบังลม→กะบังลม
- **6 loanword:** กราฟฟิค→กราฟิก, กงศุล→กงสุล, ลิ๊งค์→ลิงก์

Class 2 scope. Class 2 covers swaps within a homophone group that contains an Indic-origin letter, for example ศ/ษ/ส, ท/ธ, and น/ณ, plus the final-stop pair บ/ป. Swaps between letters outside these groups, such as ฝ/ฟ, fall in class 5 together with other structural edits.

Loanword-absorption rule. Only *contemporary* foreign loans, chiefly modern English technical and commercial terms (กราฟิก, ดิจิทัล, ลิงก์), are class 6. Long-integrated borrowings are treated as **native** even though etymologically foreign: Portuguese (กะละแม), Burmese (กะปิ), Mon (กะพง), Khmer (กบฏ), and Tamil, Persian, and Sanskrit–Pali doublets. Borderline cases (e.g. กงสุล, *consul*) are resolved by the Royal Institute Dictionary listing + Wiktionary borrowed-term categories, and the tag is recorded per item so the rule is auditable.

Automatic class assignment. The rule-based classifier used to label the confusion set and inside the edit-typer (Section 4.2) compares the wrong and correct forms of a word character by character. It checks, in order, for a swap within one of the homophone consonant groups above (class 2), an added or removed silent-consonant mark์ or a change in the consonant it silences (class 3), reordered, doubled, or inserted letters (class 5), a changed tone mark (class 1), and a changed vowel sign (class 4), and it assigns the first class that applies. Any other consonant change falls in class 5. If the corrected word is in the loan lexicon, which holds the correct forms of the class-6 confusion pairs and a few proper nouns, the class is 6 whatever the character change.

B Gold quality control and the consensus audit

Why injected items still need an audit. Injection guarantees that applying the gold edits recovers the clean original (verified 100%), so the *injected* error's correction is exact. It does **not** guarantee that the **source passage was typo-free**: the permissive corpora contain a small rate of residual orthographic errors. When an item's source carries such a residual typo, the gold inherits it, so a model that fixes *that* genuinely wrong span is unfairly scored as a miss or an over-edit. We find and remove such items with a cross-model consensus audit.

Cross-model consensus audit. For every item we pool the models' full outputs (Unicode-

normalized and variant-reconciled) and look for a *consensus* output: the most common output other than the gold, compared as the whole passage (an empty reply counts as the unchanged input), if it comes from at least k=3 models and from at least as many models as reproduce the gold. Each flagged item falls into one of three cases. (i) If the consensus equals the corrupted input, or makes some of the gold corrections and no other change, this is a *shared miss*: the gold is correct, and we **keep** the item. (ii) If the consensus differs from the gold only in punctuation or whitespace, this is a systematic *over-edit*, which the benchmark should penalize, and we **keep** it. (iii) Otherwise we **drop** the item. This case targets residual *source* typos, such as ราชกิจจาบุเบกษา→ราชกิจจานุเบกษา, คววม→ความ, and โดบ→โดย, and non-standard spellings in the source, such as the loanword อัพเดต for อัปเดต. The test of whether the consensus equals the input keeps shared misses from being taken for gold defects. The same automatic rule is applied to all three tiers and removed 45 (T1), 26 (T2), and 28 (T3) items.

What the rule also drops. The rule cannot tell a source typo from an accepted variant or a register change outside the variant map, so it also drops items where the models agree on such a change, for example วิกฤติ→วิกฤต, the informal word เท่าไหร่, and informal-to-formal rewordings such as มั้ย→ไหม and วันนึง→วันหนึ่ง. It also drops six T1 items whose gold is correct, where the models agree on a wrong fix of the injected error (เคารวะ→เคารพ for คารวะ twice, ขวาญ→ขวัญ for ขวาน, and มั๊ย→มั้ย for ไหม twice) or change นะคะ to the common misspelling นะค่ะ; dropping them slightly lowers the measured miscorrection rate. For 24 of the 45 dropped T1 items, only nine models have outputs (GPT-5.5 med and high, GPT-5.4-mini, Gemini 3.5 Flash, Claude Sonnet 4.6, DeepSeek V4 Pro and Flash, GLM-5.1, and Typhoon 2.5), plus MiniMax for 17 of them; applied to these outputs, the rule drops all 24, including one whose source sentence itself lacks a syllable (ความใจ for ความเข้าใจ). One T1 item and one T2 item contain an error that no model fixes; both inject แน่นหนา for หนาแน่น, two valid words, so only the context marks the error. We release the audit reports, which list the gold-versus-consensus difference for every dropped item.

Variant map. The four equivalence groups are an orthographic doublet (สิทธิ/สิทธิ์), two loanword transliterations (ดิจิตอล/ดิจิทัล and ออร์แกนิค/ออร์แกนิก), and an informal-to-formal pronoun pair (เค้า/เขา). Words in which the final mark is mandatory, such as ลิขสิทธิ์, are protected from reconciliation. Beyond these groups, orthographic correction is essentially single-reference, since a word has one correct spelling, so we do not adopt multi-reference GEC scoring.

C Construction details

Confusion set. Wrong→correct pairs were harvested from the Thai Wiktionary appendix of commonly misspelled words and from the fix history of a Thai Wikipedia spelling-cleanup bot, then deduplicated and cleaned (fragments, multi-token and spacing-only pairs, and context-dependent real-word pairs removed; one such pair, แน่นหนา/หนาแน่น, was not caught), giving 1,044 canonical pairs. A rule-based classifier assigns the mechanism class from the character difference; it agrees with the hand labels on 83 of the 86 bot-fix pairs that have a class (96.5%). Etymology is assigned from the Wiktionary borrowed-term categories under the absorption rule, which moves contemporary loans into class 6. We check canonical forms against the Royal Institute Dictionary word list distributed with PyThaiNLP: 87.5% are listed, and the unlisted forms include modern loanwords and common words that the list omits, such as นะคะ. Each pair is also tagged with its edit operation. Table 3 gives the per-class sizes.

Injection and mining. Injection uses **whole-token-run matching** so the 51 valid short forms ($\leq$ 3 characters) that coincide with a confusion target are never corrupted. Cleanup-bot mining reads the bot's contribution history on Thai Wikipedia and keeps a fix only when the corrected word is in the Royal Institute Dictionary word list or PyThaiNLP's general Thai word list and the wrong and correct forms align without shifting neighbouring characters; the mined items are kept out of the three tiers. The curated subset used in the annotation study keeps at most two contexts per distinct error and drops fragments of two characters or fewer.

Memorization probe. For each of the four sources we sample 30 clean sentences of 90 to 180 characters, give Typhoon 2.5 (30B) the first half at temperature 0 with the instruction เติมข้อความต่อไปนี้ให้สมบูรณ์โดยใช้ถ้อยคำเดิมที่ถูกต้องที่สุด ตอบเฉพาะส่วนที่เติมต่อ ห้ามอธิบาย ('Complete the following text with

tier	unit	items	chars	err/it	news/gov/enc
T1 sent.	1 sent.	774	100	1.23	264/363/147
T2 para.	≈5 sent.	224	425	1.24	80/79/65
T3 page	≈15 sent.	122	1,242	1.58	38/43/41

Table 2: Per-tier statistics (post-audit; Section 3.6): mean characters per item and gold edits per item. The tiers are 100% confusion-set injection; the cleanup-bot items form the separate natural-error set.

class	pairs	T1	T2	T3
1 tone mark	43	155	42	32
2 consonant-pair	168	135	37	26
3 silent-consonant	177	166	54	26
4 vowel	189	177	60	38
5 cluster/metathesis	338	204	55	54
6 loanword	129	112	30	17
total	1,044	949	278	193

Table 3: Confusion-set pairs and injected gold edits per class and tier; the grader scores the TCC-aligned edits, which number 957, 278, and 195 at T1–T3. The class shares of T1 gold edits correlate with those of the confusion-set pairs at Spearman $\rho = 0.83$ (Pearson $r = 0.67$); cluster/metathesis is the largest class in both.

its exact original wording. Reply with the continuation only, without explanation.'), and allow up to 120 output tokens. The score is the longest common prefix between the continuation and the true second half, ignoring whitespace, divided by the length of the true second half. The mean score is 0.003 (news), 0.004 (government press releases), 0.012 (Wikipedia), and 0.017 (Thai Constitution text, used only in the train split), and no sentence scores 0.5 or more. The probe uses a single model, so it does not measure memorization by the other models directly. Items that share a clean sentence are then kept together as one group, and groups are assigned to splits class by class (by the class of the group's first item): within each class, groups are shuffled with a fixed seed and ordered by their source's mean score, and the lowest-scoring 34% (mined groups first) form the held-out test split, followed by the public test (32%), dev (24%), and train splits.

Release and license. The injected items are released as train (249), dev (606), and public-test (819 before the audit) splits, plus a held-out test split (868) drawn from the least-memorized sources; the results in this paper use the audited public-test split as T1. We release the benchmark for research use under CC BY-SA (data) and the MIT license (code), consistent with the permissive redistribution terms of the source corpora.

D Full results

Per-model results by class. Table 8 gives the edit-typer's span recall and precision per class for every model (Figure 1 shows span-F0.5), and Table 7 gives the fraction of gold edits fixed per construction class, pooled over models. A model edit that matches no gold edit lowers the precision of the class the typer assigns to it; edits that fit no class, such as spacing changes, enter no class, while many word-level rewordings fall in class 5, which collects every consonant change outside the class-2 groups. Loanword precision is high by construction: the typer assigns class 6 only when the output word is in its loan lexicon, so a change that turns a correct loanword into another form is usually charged to another class.

What GPT-5.5 (high) gets wrong. GPT-5.5 (high), the best model on T1 and T2, leaves few errors unfixed: it misses 3 of the 195 T3 gold edits, and two of those it fixes inside a wider edit. Most of its unsupported T3 edits are tone-mark (23) and cluster (18) edits, and on manual inspection only 3 of these 41 clearly damage the text, by changing a proper name or a correct form. Of the rest, 25 respell a loanword or foreign name toward the standard of the Royal Society (formerly the Royal Institute), which the system prompt names, 5 fix a typo left in the source that the consensus audit did not catch, and 8 replace an informal form with the dictionary form (นึง→หนึ่ง, เท่าไหร่→เท่าไร). Because the gold keeps the source's spelling, its T3 precision partly reflects the single-reference gold rather than over-editing (Table 9).

Pooled per-class outcome split (T1, all 16 models).

class	edits×16	fixed	miscorr.	missed
1 tone	2,480	0.906	0.008	0.085
2 consonant	2,160	0.879	0.006	0.115
3 silent-consonant	2,640	0.887	0.010	0.103
4 vowel	2,816	0.863	0.002	0.134
5 cluster	3,280	0.876	0.006	0.118
6 loanword	1,920	0.810	0.003	0.186

Over-edits per 100 items across tiers (absolute accumulation).

model	T1 det	T1 corr	T1 span-F0.5	T2 corr	T2 fidelity	T3 corr	T3 fidelity
GPT-5.5 (high)	**0.981**	**0.981**	**0.972**	**0.966**	**0.988**	0.855	0.971
GPT-5.5 (med)	0.978	0.978	0.969	0.953	0.982	0.857	0.973
Claude Opus 4.7	0.977	0.975	0.964	0.948	0.986	0.853	0.976
Gemini 3.5 Flash	0.977	0.975	0.965	0.960	0.983	0.825	0.963
Gemma-4-31B	0.964	0.961	0.960	0.954	**0.988**	**0.898**	0.984
GPT-5.4-mini	0.967	0.959	0.963	0.915	0.986	0.865	**0.990**
Claude Sonnet 4.6	0.959	0.957	0.950	0.905	0.977	0.825	0.974
DeepSeek V4 Flash	0.959	0.956	0.950	0.945	0.986	0.840	0.971
DeepSeek V4 Pro	0.940	0.939	0.927	0.926	0.970	0.754	0.958
GLM-5.1	0.859	0.856	0.892	0.811	0.961	0.657	0.954
Typhoon-S 8B†	0.798	0.792	0.850	0.702	0.961	0.617	0.964
THaLLE 8B†	0.766	0.760	0.801	0.594	0.774	0.507	0.814
Typhoon 2.5 30B†	0.785	0.754	0.740	0.632	0.881	0.476	0.899
OpenThaiGPT v7.2†	0.761	0.753	0.787	0.476	0.887	0.463	0.901
MiniMax-M2.7	0.753	0.749	0.814	0.720	0.974	0.471	0.970
Qwen3.6-35B-A3B	0.725	0.709	0.656	0.384	0.776	0.310	0.838

Table 4: Full leaderboard: detection and correction F1, span-F0.5 from the edit-typer, and clean-sentence fidelity across the three length tiers. **Bold** = column best; † = Thai-adapted. Table 1 marks which models are statistically tied with the best correction F1 in each tier. Table 5 gives the confidence intervals, and Table 8 the per-class results.

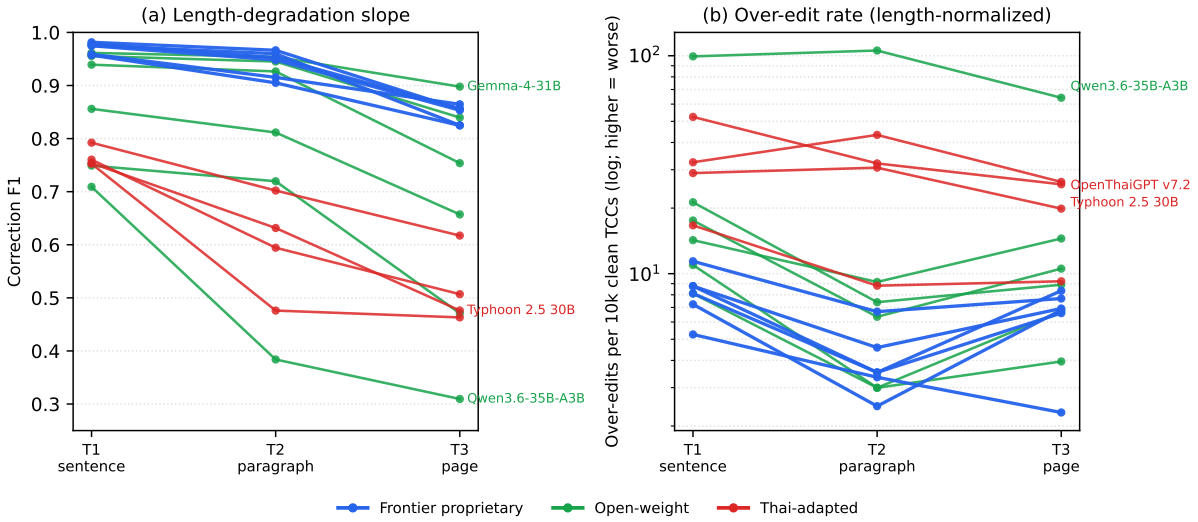


Figure 2: Length robustness. *(a)* Correction-F1 slope from T1 to T3. The frontier (blue) and the strongest open-weight model (Gemma-4, green) stay high while weak models slope off. *(b)* Over-edit rate per 10,000 clean TCCs on a log scale, where higher is worse. The rate is lower at T3 than at T1 for every model except GLM-5.1, which is about level, so the per-passage accumulation (the absolute counts in this appendix) mainly reflects the amount of correct text. Weak models over-edit at many times the frontier's per-unit rate at every tier.

model	T1	T2	T3
GPT-5.5 (med)	5	9	49
Gemma-4-31B	5	8	30
GLM-5.1	8	23	108
Typhoon 2.5 30B	31	81	192
OpenThaiGPT v7.2	19	110	197
Qwen3.6-35B-A3B	59	268	479

model	T1	T2	T3
GPT-5.5 (med)	8.1	3.5	6.6
Gemma-4-31B	8.1	3.0	4.0
GLM-5.1	14.3	9.2	14.5
Typhoon 2.5 30B	52.4	32.1	25.7
OpenThaiGPT v7.2	32.5	43.3	26.4
Qwen3.6-35B-A3B	99.3	105.7	64.2

Over-edit rate per 10,000 clean TCCs (length-normalized; the Figure 2b data). The per-unit rate is lower at T3 than at T1 for every model except GLM-5.1, which is about level (though higher at T3 than at T2 for most), so the per-item growth above mainly reflects the amount of correct text.

Register effect (pooled correction recall, T1/T2/T3). encyclopedic 0.82/0.79/0.77; news 0.88/0.85/0.83; government 0.89/0.90/0.88. Per-model correction recall by construction class is released with the data.

model	T1	T2	T3
GPT-5.5 (high)	[0.975, 0.988]	[0.947, 0.982]	[0.803, 0.904]
GPT-5.5 (med)	[0.972, 0.985]	[0.930, 0.972]	[0.804, 0.907]
Claude Opus 4.7	[0.966, 0.983]	[0.924, 0.970]	[0.783, 0.918]
Gemini 3.5 Flash	[0.966, 0.984]	[0.942, 0.977]	[0.764, 0.883]
Gemma-4-31B	[0.951, 0.971]	[0.931, 0.974]	[0.859, 0.934]
GPT-5.4-mini	[0.948, 0.969]	[0.887, 0.942]	[0.822, 0.902]
Claude Sonnet 4.6	[0.945, 0.968]	[0.873, 0.935]	[0.754, 0.891]
DeepSeek V4 Flash	[0.944, 0.966]	[0.919, 0.968]	[0.788, 0.889]
DeepSeek V4 Pro	[0.926, 0.952]	[0.902, 0.950]	[0.692, 0.813]
GLM-5.1	[0.834, 0.877]	[0.767, 0.854]	[0.537, 0.773]
Typhoon-S 8B†	[0.770, 0.814]	[0.652, 0.750]	[0.549, 0.682]
THaLLE 8B†	[0.725, 0.790]	[0.540, 0.649]	[0.426, 0.589]
Typhoon 2.5 30B†	[0.727, 0.779]	[0.581, 0.684]	[0.406, 0.550]
OpenThaiGPT v7.2†	[0.710, 0.788]	[0.385, 0.569]	[0.386, 0.544]
MiniMax-M2.7	[0.723, 0.774]	[0.672, 0.765]	[0.391, 0.550]
Qwen3.6-35B-A3B	[0.677, 0.741]	[0.312, 0.472]	[0.227, 0.421]

Table 5: 95% item-bootstrap confidence intervals on correction F1 (10,000 resamples). † = Thai-adapted.

model	fixed	miscorrected	missed
GPT-5.5 (high)	0.997	0.000	0.003
Gemma-4-31B	0.963	0.002	0.034
Qwen3.6-35B-A3B	0.819	0.018	0.163
Typhoon 2.5 30B	0.775	0.032	0.192
All 16, T1	0.873	0.006	0.122
All 16, T3	0.826	0.005	0.168
All 16, natural	0.820	0.006	0.175

Table 6: Gold-edit outcomes (T1 unless stated). The miscorrected share stays near zero throughout.

class	T1	T2	T3
1 tone mark	0.906	0.871	0.809
2 consonant-pair	0.879	0.887	0.822
3 silent-consonant	0.887	0.865	0.805
4 vowel	0.863	0.870	0.831
5 cluster/metathesis	0.876	0.809	0.845
6 loanword	0.810	0.790	0.829
all classes	0.873	0.850	0.826
spread (max−min)	0.096	0.097	0.040

Table 7: Fraction of gold edits fixed, by mechanism class and tier, pooled over all 16 models (construction labels). Table 3 gives the injected errors per class; the scored TCC edit spans per item can differ slightly after alignment.

E Annotation study

Design. The 60 items comprise the 40 curated mined items (Appendix C) and 20 synthetic items chosen so that each automatic class has ten items. Each item shows one marked span with its proposed correction. Raters answer two questions per item: whether the marked span is a genuine orthographic error, and which of the six classes it belongs to, following a written guideline with the class definitions and the two conventions of Appendix A. No automatic label is shown. The two raters, one of them an author, worked independently. Etymology and edit-operation tags are not annotated, since they are a dictionary lookup and a deterministic function of the edit.

Results. Error validity: both raters accepted all 60 spans. Class agreement: Krippendorff's $\alpha = 0.752$ (nominal, $n = 60$). Table 10 gives one-vs-rest α and raw agreement per automatic class; with ten items per class these values are noisy.

Where labels differ. The raters split on 12 items and agree with each other against the automatic label on 13. Of these 25, 17 involve the loanword-absorption rule in one of two directions: raters labeled the surface mechanism of a contemporary loan (for example, a consonant swap in เวียตนาม→เวียดนาม rated class 2 rather than 6), or recognized a loan that the loan lexicon missed (เยอรมันนี→เยอรมนี, typed class 5 but rated 6); old borrowings such as เวทมนตร์ split the raters between class 3 and 6. Four more are homophone swaps outside the Indic-origin groups (คน→คน three times, and กบถ→กบฏ), which the taxonomy places in class 5 and both raters placed in class 2. The remaining four concern whether a stray or misplaced diacritic counts as a tone-mark, vowel, or silent-consonant error. The disagreements therefore sit at the class boundaries fixed by convention rather than at error perception.

F Natural-error evaluation

The natural set comes from the edit history of a Thai Wikipedia spelling-cleanup bot: 103 items, each a real error in its original encyclopedic sentence,

	T1 (sentence)												T3 (page)											
	1		2		3		4		5		6		1		2		3		4		5		6	
model	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
GPT-5.5 (high)	.99	.93	1.0	.99	1.0	.99	.99	.98	1.0	.94	1.0	1.0	.97	.57	1.0	.96	.96	.83	.97	.79	1.0	.75	1.0	.90
GPT-5.5 (med)	.99	.94	.99	.99	1.0	.99	.99	.96	1.0	.94	1.0	1.0	.97	.57	.96	1.0	.96	.83	.97	.84	1.0	.73	1.0	.90
Claude Opus 4.7	1.0	.97	.99	.98	1.0	.99	1.0	.98	.98	.97	.99	1.0	1.0	.84	.92	.96	1.0	.96	1.0	1.0	.98	.88	1.0	1.0
Gemini 3.5 Flash	.99	.93	.99	1.0	.99	.96	1.0	.98	.99	.92	1.0	1.0	.97	.52	.96	.93	.92	.77	1.0	.88	1.0	.67	1.0	.90
Gemma-4-31B	.95	.95	.93	1.0	.97	.99	.98	.97	.97	.96	.98	1.0	.94	.77	.96	1.0	.96	.93	.95	.97	1.0	.79	1.0	.90
GPT-5.4-mini	.92	.94	.96	.96	.95	.98	.98	.99	.95	.96	.95	1.0	.75	.83	.85	1.0	.92	.89	.82	1.0	.87	.85	.95	.95
Claude Sonnet 4.6	.98	.99	.98	.99	.95	.98	.98	.98	.96	.98	.98	1.0	1.0	.91	.96	1.0	.88	1.0	.95	.97	.94	.85	1.0	1.0
DeepSeek V4 Flash	.93	.95	.98	.99	.96	.97	.99	.96	.95	.90	.99	1.0	.91	.56	.92	1.0	.92	.89	.97	.80	.98	.79	1.0	.95
DeepSeek V4 Pro	.95	.92	.93	.98	.97	.94	.98	.94	.96	.85	.98	1.0	.88	.45	.88	.96	.92	.77	.92	.74	.91	.67	.89	.89
GLM-5.1	.83	.92	.87	.97	.79	.90	.84	.94	.75	.89	.74	1.0	.69	.41	.88	.96	.81	.47	.92	.76	.83	.47	.79	.88
Typhoon-S 8B†	.78	.97	.66	.95	.72	.97	.69	.99	.75	.93	.62	.99	.50	.84	.69	1.0	.54	.93	.71	.96	.69	.77	.68	1.0
THaLLE 8B†	.90	.91	.69	.99	.81	.91	.61	.97	.77	.71	.33	1.0	.88	.68	.50	1.0	.65	.25	.61	.88	.72	.34	.42	.89
Typhoon 2.5 30B†	.84	.93	.79	.88	.86	.89	.68	.95	.77	.60	.71	.96	.75	.83	.69	1.0	.69	.95	.68	.87	.70	.44	.58	1.0
OpenThaiGPT v7.2†	.87	.90	.70	.90	.81	.94	.62	.94	.75	.68	.38	1.0	.59	.90	.85	.96	.58	.75	.71	1.0	.67	.54	.68	1.0
MiniMax-M2.7	.84	.93	.73	.94	.60	.93	.64	.92	.67	.87	.47	.98	.50	.84	.50	.93	.35	.50	.37	.78	.48	.72	.37	1.0
Qwen3.6-35B-A3B	.76	.65	.88	.90	.79	.80	.84	.88	.82	.43	.84	.97	.66	.36	.62	.62	.81	.55	.74	.67	.74	.15	.89	.89

Table 8: Span recall (R) and precision (P) per mechanism class for every model, from the edit-typer (Figure 1 shows span-F0.5). Classes: 1 tone mark, 2 consonant-pair, 3 silent-consonant, 4 vowel, 5 cluster, 6 loanword. † = Thai-adapted.

tier	cl.	input	GPT-5.5 (high)	note
T1	4	ธนาบัตร	ธนาบัตร	missed: ธนบัตร
T1	5	แน่นหนา	แน่นหนา	missed: หนาแน่น
T3	4	สารสำคัญ	สารสำคัญ	missed: สาระสำคัญ
T3	1	เอมพิวี	เอ็มพิวี	standard loanword
T3	5	สตาร์ท	สตาร์ต	standard loanword
T3	1	ปีนึง	ปีหนึ่ง	formalized
T3	5	ย่ามรุ่ง	ย่ำรุ่ง	source typo
T3	1	ริเวอร์รี่	ริเวอร์รี	name changed

Table 9: Examples of GPT-5.5 (high) edits scored as errors. The first three are gold errors it leaves unchanged (gold form after the colon); the rest are unsupported edits to words the gold leaves as in the input, labelled by the edit-typer's class (cl.).

automatic class	α (one-vs-rest)	agreement
1 tone mark	0.793	1.00
2 consonant-pair	0.921	0.90
3 silent-consonant	0.611	0.60
4 vowel	0.773	0.70
5 cluster/metathesis	1.000	1.00
6 loanword	0.485	0.60
all (multiclass)	0.752	0.80

Table 10: Inter-rater agreement per automatic class (ten items each).

with the bot's fix as gold. We apply the consensus audit of Section 3.6 with the same rule as the tiers: at least $k = 3$ models, and at least as many as reproduce the gold, must agree on an output that differs from the gold. The audit drops such a flagged item unless that output equals the input, differs only in whitespace or punctuation, or is a partial fix on the way to the gold. It removes 17 items. In some, the consensus also fixes a second, unmarked error in the sentence (for example เพีย→เพีย); in others, the recorded fix is itself incomplete and the consensus completes it (ตำ่ recorded as ตำ, น้ำ as น้า, and สราง้ as สราง, where the intended words are ต่ำ, น้ำ, and สร้าง). Without the condition that the agreeing models at least match those reproducing the gold, the audit would remove 25 items (78 kept; mean correction F1 0.81, $\rho = 0.82$), including one where the models change a correctly spelled film title. The audited set has 86 items with 86 recorded fixes (87 aligned gold edits). Classified without the loanword rule, the fixes fall in classes 1: 26, 2: 21, 3: 14, 4: 20, and 5: 5; the edit-typer, which applies the rule, types 6 of the 87 gold edits as loanwords and 4 as class 5. Table 11 gives per-model results on the audited set, with correction F1 and the miscorrected share on all 103 items for comparison. On the audited set the ranking correlates with the T1 ranking at Spearman $\rho = 0.81$ (Pearson $r = 0.93$); on all 103 items, where mean correction F1 is 0.71, mostly because of the 17 dropped items, at $\rho = 0.80$ ($r = 0.92$).

G Prompt and inference

Prompt. All models receive the same Thai system message and instruction. The system message is คุณเป็นเครื่องมือตรวจแก้การสะกดคำภาษาไทย ตามอักขรวิธีมาตรฐานของราชบัณฑิตยสภา ('You

	audited set ($n = 86$)						all items ($n = 103$)	
model	corr F1	95% CI	det F1	fixed	miscorrected	missed	corr F1	miscorrected
Gemma-4-31B	**0.925**	[0.876, 0.966]	0.936	0.919	0.011	0.069	0.819	0.058
GPT-5.4-mini	0.924	[0.871, 0.971]	0.936	0.908	0.011	0.081	0.811	0.058
Claude Sonnet 4.6	0.898	[0.847, 0.942]	0.898	0.862	0.000	0.138	0.771	0.058
Claude Opus 4.7	0.885	[0.790, 0.955]	0.885	0.931	0.000	0.069	0.791	0.038
Gemini 3.5 Flash	0.871	[0.768, 0.955]	0.871	0.966	0.000	0.035	0.757	0.038
GPT-5.5 (high)	0.862	[0.768, 0.939]	0.862	0.966	0.000	0.035	0.751	0.048
GPT-5.5 (med)	0.862	[0.768, 0.941]	0.862	0.966	0.000	0.035	0.760	0.048
DeepSeek V4 Flash	0.854	[0.766, 0.929]	0.854	0.839	0.000	0.161	0.757	0.048
DeepSeek V4 Pro	0.850	[0.778, 0.917]	0.850	0.908	0.000	0.092	0.729	0.038
GLM-5.1	0.795	[0.708, 0.876]	0.795	0.759	0.000	0.241	0.689	0.048
Typhoon 2.5 30B†	0.753	[0.663, 0.837]	0.763	0.805	0.011	0.184	0.655	0.048
THaLLE 8B†	0.744	[0.657, 0.820]	0.744	0.667	0.000	0.333	0.670	0.048
MiniMax-M2.7	0.725	[0.638, 0.802]	0.725	0.621	0.000	0.379	0.663	0.038
OpenThaiGPT v7.2†	0.722	[0.632, 0.805]	0.722	0.655	0.000	0.345	0.630	0.048
Typhoon-S 8B†	0.708	[0.618, 0.789]	0.722	0.586	0.011	0.402	0.626	0.038
Qwen3.6-35B-A3B	0.589	[0.492, 0.685]	0.625	0.759	0.046	0.195	0.544	0.067

Table 11: Natural-error evaluation. **Bold** = best on the audited set; underlined = not separable from the best by a paired bootstrap test ($p \geq 0.05$); † = Thai-adapted.

are a tool that checks and corrects Thai spelling according to the standard orthography of the Royal Society'). The T1 instruction, with the input in place of {text}, is:

> จงแก้คำที่สะกดผิดในประโยคต่อไปนี้ให้ถูกต้องตามอักขรวิธี แก้เฉพาะตัวพยัญชนะ/สระ/วรรณยุกต์/ตัวการันต์ที่สะกดผิดเท่านั้น ห้ามเปลี่ยนถ้อยคำ ห้ามเรียบเรียงใหม่ ห้ามเพิ่มหรือลบเนื้อความ ห้ามเปลี่ยนการเว้นวรรค ตัวเลข หรือเครื่องหมายวรรคตอน ถ้าไม่มีคำผิดให้ตอบประโยคเดิมทุกตัวอักษร ตอบกลับเป็นประโยคที่แก้ไขแล้วบรรทัดเดียวเท่านั้น ไม่ต้องอธิบาย
>
> ประโยค: {text}
> ประโยคที่แก้ไขแล้ว:

For T2 and T3, ประโยค ('sentence') becomes ข้อความ ('text'), and the single-line requirement is dropped. The English prompt of the sensitivity test below translates the T1 instruction:

> Correct the misspelled Thai words in the following sentence according to standard Thai orthography. Fix only misspelled consonants, vowels, tone marks, or silent-consonant (karan) marks. Do not change the wording, do not rephrase, do not add or remove content, and do not change spacing, numerals, or punctuation. If there are no misspellings, return the sentence exactly as given. Reply with only the corrected sentence on a single line, with no explanation.
>
> Sentence: {text}
> Corrected sentence:

The few-shot prompt keeps the Thai instruction and inserts ตัวอย่าง ('Examples') followed by two worked examples from the train split, in the same ประโยค / ประโยคที่แก้ไขแล้ว format, before the input. The examples contain a tone-mark error (ก้อ→ก็) and a consonant-pair error (ฐาณ→ฐาน):

> ประโยค: แต่เพลงนี้ก้อยังไม่ได้มีการประกาศอย่างเป็นทางการว่าเป็นเพลงชาติ
> ประโยคที่แก้ไขแล้ว: แต่เพลงนี้ก็ยังไม่ได้มีการประกาศอย่างเป็นทางการว่าเป็นเพลงชาติ
>
> ประโยค: โดยถ้าลอการิทึมในสมการข้างต้น เป็นฐาณสอง เอนโทรปีที่วัดจะอยู่ในหน่วยบิตเช่นกัน
> ประโยคที่แก้ไขแล้ว: โดยถ้าลอการิทึมในสมการข้างต้น เป็นฐานสอง เอนโทรปีที่วัดจะอยู่ในหน่วยบิตเช่นกัน

Both variants keep the Thai system message.

Inference. Temperature is 0 where supported. Claude models use the Anthropic Message Batches API in non-thinking mode. GPT-5.x models use the Responses API, GPT-5.5 at medium or high reasoning effort with 3k and 6k output-token budgets and GPT-5.4-mini at its default effort with a 2k budget; these budgets include the reasoning, and the 22 GPT-5.5 replies that exhausted them were rerun with a 24k budget. Gemini 3.5 Flash and the open reasoning models run in native thinking mode with 24k-token budgets (8k for OpenThaiGPT and THaLLE); the other models run without thinking. Token-runout counts as a failure. Replies are cleaned by removing reasoning traces and echoed labels, keeping the last non-empty line for T1 or joining all lines for T2 and T3, and stripping quotation marks at both ends. For the 114 of 17,920 scored replies (0.6%) whose passage itself begins or ends with a quotation mark, this cleaning also removes a correct quotation mark, which counts as one error. Scoring these replies without the stripping changes any model's correction F1 by at most

model	Thai (default)	English	Thai + 2 examples
Gemma-4-31B	0.961	0.933	0.893
Typhoon-S 8B	0.792	0.808	0.844

Table 12: T1 correction F1 under three prompts ($n = 774$).

0.005 and leaves the order of models unchanged in every tier.

Failed and partial replies. A request that still fails with a server error after four attempts is scored like an empty reply, as the unchanged input; this affects four T1 items for OpenThaiGPT and one T3 item for THaLLE. THaLLE also often returns only part of a long passage, a short stretch from its middle or end, even though the reply ends well within its token budget. This happens on 44 of 224 T2 items and 19 of 122 T3 items. These replies are scored as written, so the omitted text counts as deleted, and THaLLE's T2 and T3 scores reflect these incomplete answers as well as its corrections.

Prompt sensitivity. We rerun the full T1 set for one strong general model (Gemma-4-31B) and one Thai 8B model (Typhoon-S) under two alternative prompts: the same instruction in English, and the Thai instruction with two worked examples from the train split, which is disjoint from T1 (Table 12). Prompt choice shifts correction F1 by at most 0.07, and Gemma-4 stays ahead of Typhoon-S under every prompt. The two models move in opposite directions: the worked examples help the Thai 8B model and hurt the general model.