CHIENGMIXBENCH: WHEN "MODEL PREFERENCE" DEPENDS ON HOW YOU SCORE IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Technical concepts can often be expressed either with English terms or their Chinese equivalents. A common way to ask which form a language model favors is to compare the likelihoods the model assigns to the two forms. Yet such "preference" is not directly observed from the model; it is induced by the scoring rule used to measure it. Whether we score the whole sentence or the tokens overlapping the changed text, and whether we sum or average token-level scores, can change the resulting judgment. We introduce ChiEngMixBench, a controlled testbed for studying this issue through paired Chinese–English terminology substitutions, and compare several natural likelihood-based definitions of preference. We find an easily overlooked pattern: scoring definitions can agree on the majority direction while disagreeing substantially about the winners for individual pairs. Thus, agreement on the majority direction alone does not establish which individual items or subgroups behave consistently across scoring definitions; robustness must be checked at the level of the claim.

1 INTRODUCTION

The same technical concept can often be expressed in more than one reasonable way. In Chinese technical writing, a concept may keep its English term or use a Chinese counterpart. This raises a simple question: if a model is given two candidate forms for expressing the same concept, which one does it support more? We often summarize the answer as "the model prefers A over B." The model does not report such a preference directly; researchers infer it by comparing support for the two candidates, following a common minimal-pair and log-probability evaluation interface (Linzen et al., 2016; Warstadt et al., 2020; Kauf et al., 2024). A seemingly direct language choice therefore depends on how it is measured.

There is more than one natural way to make that measurement. We can score the complete sentences containing the two terms or focus only on the changed span. We can add the token-level log-likelihoods into a cumulative score or average them over the number of tokens. These choices interact with tokenization and length normalization in established language-model evaluation settings (Ueda et al., 2024; Cao & Rimell, 2021; Chirkova et al., 2023). If different choices produced opposite aggregate conclusions, the instability would be easy to notice. The more subtle case is when several choices tell the same overall story. Their agreement may then appear robust. But if they all say that the model favors Chinese overall, do they also identify the same individual examples as Chinese-favored?

Chinese technical terminology provides a particularly clean setting for this question. We can keep the rest of a sentence closely matched while changing the terminology form. The same local language choice can then be compared under different scoring definitions. ChiEngMixBench contains 1,167 strict single-replacement pairs covering 934 normalized terms. Each pair is constructed to keep the surrounding context, technical concept, and syntactic position as similar as possible while changing the English term to its Chinese counterpart. We therefore need not decide in advance which form is "correct"; we can ask whether different scoring definitions make the same judgment about the same choice.

At the aggregate level, the majority direction is stable across the four scoring definitions. On Qwen2.5-7B, all four scoring definitions give the Chinese form the higher score on 81.80%–97.34% of pairs. On Llama 3.1 8B, all four give the same overall direction, ranging from 90.57% to 94.94%.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

If the claim is only that the model more often assigns the higher score to one form overall, the conclusion looks robust. But the picture changes when we compare pairs one by one. The largest disagreement is 19.83% for Qwen2.5-7B and 13.30% for Llama 3.1 8B. For Qwen2.5-7B, the two most discrepant scoring definitions both treat English-favored cases as a minority, yet their English-preferred sets have a Jaccard similarity of only 0.083. The scoring definitions agree on the overall story but not on which specific examples make up its exceptions.

This combination of a stable overall direction and substantial item-level disagreement also appears in Qwen2.5-0.5B, TinyLlama-1.1B, Qwen2.5-7B, and Llama 3.1 8B. This distinction matters because agreement on the majority direction alone cannot establish consistent judgments for particular items or term types. Our contributions are therefore threefold: we introduce ChiEngMixBench as a controlled testbed for studying preference measurement; we show that natural likelihood scoring definitions can agree on the aggregate direction while disagreeing materially on item-level judgments; and we formulate a simple evaluation principle—robustness should be checked at the same granularity as the scientific claim.

2 Related Work

2.1 Likelihood-based contrastive evaluation

Likelihood comparisons over controlled alternatives are a standard interface for probing language-model behavior. Targeted syntactic evaluation and BLiMP use minimally different sentence pairs and record whether the model assigns higher probability to the linguistically preferred member (Linzen et al., 2016; Warstadt et al., 2020). Log probabilities have also been used as evidence for semantic plausibility in minimal-pair and context-dependent settings (Kauf et al., 2024). These studies establish the usefulness of pairwise likelihood judgments, but they typically fix a scoring convention and treat the resulting ranking as the evaluation outcome.

Contrastive evaluation also has a well-defined behavioral interface: it ranks predefined alternatives rather than directly observing a model's generation. Vamvas & Sennrich (2021) show that conclusions from minimal pairs can depend on the relation between the contrastive data and deployment-time behavior. Our setting shares the controlled pairwise interface, while asking a different question: when several reasonable scoring definitions are applied to the same pairs, does agreement on the aggregate direction extend to the same item-level winners?

2.2 Sensitivity of probability-based scoring

Sentence probability is sensitive to how strings are tokenized and how token-level values are aggregated. Ueda et al. (2024) show that matching word counts does not remove subword token-length bias in minimal-pair evaluation. Cao & Rimell (2021) and Chirkova et al. (2023) study the gap between canonical-token and marginal likelihood over alternative tokenizations. Lesci et al. (2025) further quantify causal effects of tokenizer vocabulary choices on string probabilities. In Chinese minimal-pair evaluation, Liu et al. (2026) propose a length-normalized log-probability metric to mitigate unequal-length bias.

These studies motivate making scoring choices explicit. ChiEngMixBench keeps each model and its tokenizer fixed while varying scoring scope (whole sentence versus changed span) and token aggregation (cumulative versus mean).

2.3 Chinese–English terminology as a controlled setting

Recent multilingual evaluation work uses linguistically controlled code-switched variants to isolate the effects of language choice in model comprehension and reasoning (Mohamed et al., 2026). ChiEngMixBench uses a narrower but complementary setting: closely matched Chinese technical sentences that differ in whether a concept is realized with an English term or its Chinese counterpart. This construction makes terminology choice the controlled contrast for comparing alternative preference measurements. The dataset construction and pair-fidelity criteria are described in the testbed section and Appendix A.

3 CHIENGMIXBENCH AS A CONTROLLED PREFERENCE TESTBED

The measurement setup is summarized in Figure 1. A fixed terminology pair is passed through four natural scoring definitions, and the resulting judgments are examined at both aggregate and item levels.

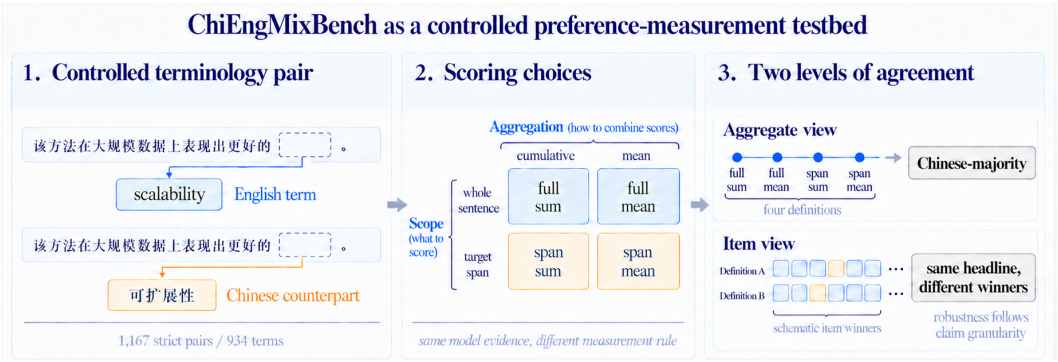


Figure 1: Overview of ChiEngMixBench as a controlled preference-measurement testbed. Each pair varies terminology realization within a closely matched context. Four natural scoring definitions arise by crossing scoring scope with token aggregation. The same definitions can agree on the aggregate Chinese-majority direction while assigning different item-level winners, motivating robustness checks at the granularity of the claim.

ChiEngMixBench contains 1,167 strict single-replacement pairs covering 934 normalized terms. The four scoring definitions are applied to this fixed set, although two pairs lack valid target-span scores (Section 3.2). Appendix A describes dataset construction and filtering.

3.1 TWO CHOICES BEHIND A LIKELIHOOD-BASED PREFERENCE

For each terminology pair, we want to answer a seemingly simple question: given an English term and its Chinese counterpart in the same context, which form does the model support more? Turning "support" into a likelihood-based preference requires two choices.

The first choice is where to score. One option is to score all eligible tokens in the complete sentences. This captures both the changed term and shifts in the conditional probabilities of following tokens, even when the following text is identical. The other option is to score only the target span that actually changed, focusing the measurement on the terminology form itself. The former is closer to preference between complete realizations; the latter is closer to a local terminology choice. The two scopes therefore answer different questions.

The second choice is how to aggregate token-level evidence. A cumulative log-likelihood retains the joint-probability form of the scored region. Averaging by the number of tokens reduces the effect of different sequence lengths, while changing the probabilistic interpretation of the score. Crossing these choices yields the 2×2 design shown in Table 1.

Table 1: The four core scoring definitions. All four use the same causal next-token log probabilities; they differ only in scoring scope and aggregation.

Scoring scope	Cumulative evidence	Mean evidence per token
Whole sentence	Whole-sentence sum	Whole-sentence mean
Changed target span	Target-span sum	Target-span mean

3.2 FORMAL SCORING DEFINITIONS

For pair j and terminology form $v \in \{E, C\}$, let the model tokenize the realization as

$$s_v^{(j)} = \left(x_{v,1}^{(j)}, \ldots, x_{v,L_v^{(j)}}^{(j)}\right),$$

where E and C denote the English and Chinese realizations. Under the standard causal next-token convention used by the scorer, the next-token log-likelihood for token i is

$$\ell_{v,i}^{(j)} = \log p_\theta\left(x_{v,i}^{(j)} \mid x_{v,<i}^{(j)}\right).$$

Let $I_v^{(j)}$ be the set of token positions with a valid next-token log-likelihood. For a sequence with raw token length $L_v^{(j)}$, the scorer uses $I_v^{(j)} = \{2, \ldots, L_v^{(j)}\}$: the first token is not scored because no preceding position predicts it. Let $M_v^{(j)} \subseteq I_v^{(j)}$ be the changed-span positions identified on the original complete-sentence tokenization. The four scores are

$$S_{\text{full,sum}}(s_v^{(j)}) = \sum_{i \in I_v^{(j)}} \ell_{v,i}^{(j)},$$

$$S_{\text{full,mean}}(s_v^{(j)}) = \frac{1}{|I_v^{(j)}|} \sum_{i \in I_v^{(j)}} \ell_{v,i}^{(j)},$$

$$S_{\text{span,sum}}(s_v^{(j)}) = \sum_{i \in M_v^{(j)}} \ell_{v,i}^{(j)},$$

$$S_{\text{span,mean}}(s_v^{(j)}) = \frac{1}{|M_v^{(j)}|} \sum_{i \in M_v^{(j)}} \ell_{v,i}^{(j)}.$$

For each model and sentence pair, all four definitions use the same next-token log probabilities; only the scored token positions and aggregation vary. Target-span scoring uses the changed-span mask on the original complete-sentence tokenization rather than re-tokenizing or re-encoding the target string separately. Whole-sentence scores are available for all 1,167 strict pairs. Two pairs do not yield a valid changed span under the fixed span-construction rule, so span-based analyses use the remaining 1,165 pairs; span exclusions and validation details are given in Appendix C.

3.3 FROM SCORES TO ITEM-LEVEL PREFERENCE

Let m denote one of the four scoring definitions. For pair j, define the score difference between the English and Chinese realizations as

$$\Delta_m^{(j)} = S_m\left(s_E^{(j)}\right) - S_m\left(s_C^{(j)}\right).$$

If $\Delta_m^{(j)} > 0$, the pair is English-preferred under m; if $\Delta_m^{(j)} < 0$, it is Chinese-preferred. Let V_m denote the set of pairs with a finite, nonzero score difference under m.

The English-preference rate for scoring definition m is

$$R_m^E = \frac{1}{|V_m|} \sum_{j \in V_m} \mathbf{1}\left[\Delta_m^{(j)} > 0\right],$$

with the Chinese-preference rate defined analogously. This aggregate quantity asks which form receives the higher score more often overall.

Aggregate rates summarize the majority direction. To test whether that agreement extends to individual decisions, we additionally compare item-level winners. For definitions m_a and m_b, let $V_{ab} = V_{m_a} \cap V_{m_b}$. Their item-level disagreement rate is

$$D(m_a, m_b) = \frac{1}{|V_{ab}|} \sum_{j \in V_{ab}} \mathbf{1}\left[\text{sign}\left(\Delta_{m_a}^{(j)}\right) \neq \text{sign}\left(\Delta_{m_b}^{(j)}\right)\right].$$

The corresponding agreement rate is $1 - D(m_a, m_b)$. We also define the English-preferred item set

$$E_m = \left\{ j \in V_m : \Delta_m^{(j)} > 0 \right\}$$

and measure the overlap between two such sets with Jaccard similarity when their union is nonempty:

$$J(m_a, m_b) = \frac{|E_{m_a} \cap E_{m_b}|}{|E_{m_a} \cup E_{m_b}|}.$$

3.4 Comparing preference judgments across scoring definitions

We compare the four scoring definitions at two levels: whether they agree on which form wins on most pairs, and whether they choose the same winner for each individual pair. Section 4 presents both comparisons.

4 Aggregate Agreement, Item-Level Disagreement

4.1 Same Direction, Different Preference Rates

All four scoring definitions yield a Chinese-majority aggregate result for both larger models. For Qwen2.5-7B, the Chinese-preference rates are 86.12%, 94.77%, 81.80%, and 97.34% for whole-sentence mean, whole-sentence sum, target-span mean, and target-span sum, respectively. The majority direction remains unchanged, but the Chinese-preference rate varies by 15.54 percentage points across scoring definitions. Equivalently, the English-preference rate ranges from 2.66% to 18.20% for the same model—a nearly sevenfold difference. For Llama 3.1 8B, the corresponding Chinese-preference rates are 92.12%, 90.57%, 93.05%, and 94.94%, a narrower 4.37-point range.

Table 2: Aggregate preference rates for the four core scoring definitions. The valid N is 1,167 for whole-sentence scores and 1,165 for target-span scores.

Model	Scoring definition	N	English count	English rate	Chinese count	Chinese rate
Qwen2.5-7B	Full mean	1167	162	13.88%	1005	86.12%
Qwen2.5-7B	Full sum	1167	61	5.23%	1106	94.77%
Qwen2.5-7B	Span mean	1165	212	18.20%	953	81.80%
Qwen2.5-7B	Span sum	1165	31	2.66%	1134	97.34%
Llama 3.1 8B	Full mean	1167	92	7.88%	1075	92.12%
Llama 3.1 8B	Full sum	1167	110	9.43%	1057	90.57%
Llama 3.1 8B	Span mean	1165	81	6.95%	1084	93.05%
Llama 3.1 8B	Span sum	1165	59	5.06%	1106	94.94%

4.2 Item-level disagreement is widespread and structured

Aggregate rates do not tell us whether the four definitions judge the same pairs. Across the six pairwise comparisons of scoring definitions, Qwen2.5-7B has disagreement rates from 4.12% to 19.83%, with five of the six comparisons above 10%. Llama 3.1 8B ranges from 6.61% to 13.30%, with several comparisons close to or above 10%.

For all four models, the largest pairwise disagreement occurs between whole-sentence sum and target-span mean.

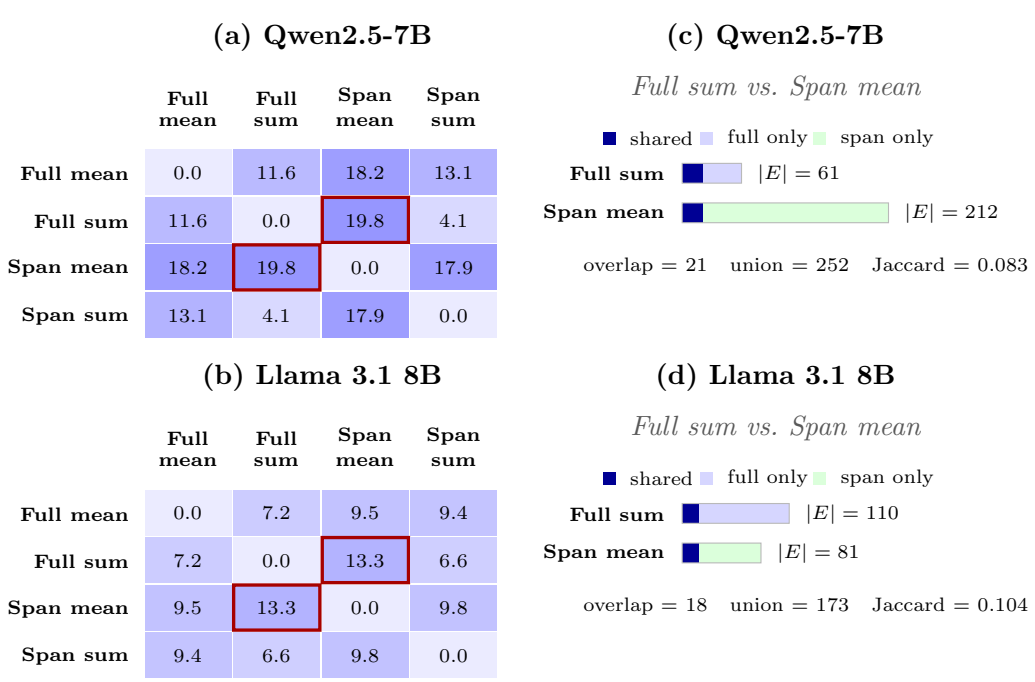


Figure 2: Pairwise disagreement despite an unchanged Chinese-majority direction. Panels (a) and (b) show disagreement across scoring definitions for Qwen2.5-7B and Llama 3.1 8B. Panels (c) and (d) compare the English-preferred sets under whole-sentence sum and target-span mean, highlighting their small overlap.

4.3 Scoring definitions disagree on which exceptions exist

The distinction is clearest when comparing whole-sentence sum with target-span mean. For Qwen2.5-7B, whole-sentence sum marks 61 items as English-preferred, whereas target-span mean marks 212. Only 21 items overlap: 40 are selected only by whole-sentence sum and 191 only by target-span mean. The two sets have Jaccard similarity 0.083.

Llama 3.1 8B shows a similar pattern: whole-sentence sum and target-span mean select 110 and 81 English-preferred items, respectively, with only 18 in common out of a 173-item union (Jaccard similarity 0.104).

Target-span mean identifies more English-preferred items than whole-sentence sum on Qwen2.5-7B, but fewer on Llama 3.1 8B. Neither definition consistently identifies more English-preferred exceptions across these two models, and their selected item sets differ as well.

4.4 Disagreement persists across evaluated checkpoints

The same phenomenon appears in the two smaller models. The maximum disagreement across the four core definitions is 17.85% for Qwen2.5-0.5B, 13.82% for TinyLlama-1.1B, 19.83% for Qwen2.5-7B, and 13.30% for Llama 3.1 8B.

This pattern appears in all four evaluated checkpoints; the comparison is descriptive, not a controlled estimate of scaling effects.

Table 3: Maximum pairwise disagreement across the four core scoring definitions.

Model	Scale	Family	Max. disagreement	Scoring definitions
Qwen2.5-0.5B	0.5B	Qwen	17.85%	Full sum vs. Span mean
TinyLlama-1.1B	1.1B	Llama-derived	13.82%	Full sum vs. Span mean
Qwen2.5-7B	7B	Qwen	19.83%	Full sum vs. Span mean
Llama 3.1 8B	8B	Llama	13.30%	Full sum vs. Span mean

5 Discussion and Implications

The four scoring definitions agree on the overall direction: across the tested models, Chinese realizations receive the higher score on most pairs. This aggregate result supports an aggregate claim about the majority direction. The same headline, however, can be produced by different item-level decisions. Agreement on the overall direction therefore does not by itself identify which terminology items are preferred or whether a subgroup follows the same pattern.

The resulting evaluation principle is to match robustness checks to the claim. Aggregate preference rates summarize the majority direction, whereas claims about individual exceptions need pairwise agreement or disagreement rates and, where relevant, overlap between the selected item sets. Claims about term types or other subgroups require checks within those groups rather than extrapolation from the full dataset.

Reports of likelihood-based preference should therefore name the scoring scope and aggregation rule alongside the conclusion. Otherwise, an unchanged aggregate direction may conceal substantial differences in which items receive the higher score.

6 Limitations

Our evidence is scoped to English–Chinese terminology alternatives in Chinese AI/CS technical discourse. The strict pairs provide closely matched local contrasts, but they do not cover every multilingual community, domain, or communicative goal. A higher score for one realization is a likelihood-based model judgment, not a human preference label, a generation-frequency estimate, or a claim that the realization is universally better.

The four core definitions are natural combinations of scoring scope and token aggregation, not an exhaustive catalogue of possible likelihood protocols. Other choices, including alternative tokenization treatments or context windows, may produce additional variation. We therefore interpret the reported rates as evidence about measurement stability under the definitions tested here.

The two larger-model evaluations use local Qwen2.5-7B and Llama 3.1 8B Ollama GGUF Q4_K_M models. These runs broaden the comparison beyond the smaller checkpoints but are not exact full-precision replications. The four-model comparison is descriptive rather than a controlled scaling study.

Finally, the benchmark construction aims to match local context and isolate terminology realization, but source-derived pairs can still contain residual linguistic differences. The appendix documents quality labels, span exclusions, protocol variants, and model provenance to support replication and follow-up analyses.

7 Conclusion

ChiEngMixBench treats English-versus-Chinese terminology realization as a controlled preference-measurement problem. Four natural likelihood scoring definitions agree on the aggregate direction for the evaluated models, but they can disagree substantially on which form wins for individual pairs. The same aggregate–item discrepancy appears across all four evaluated checkpoints.

Agreement on the majority direction does not establish item-level or subgroup consistency. Preference studies should match robustness checks to the claims they make, including item-level agreement or Jaccard similarity when interpreting individual exceptions.

ETHICS STATEMENT

ChiEngMixBench is derived from publicly accessible Chinese AI/CS discussions on CSDN and Zhihu, documented as collected between May 2021 and November 2025. The study concerns terminology usage rather than user attributes. To reduce privacy and redistribution risks during review, the anonymous supplementary artifact excludes exact source-derived sentence text, source URLs, account metadata, raw webpages, and other source-linking metadata, and contains only text-free alignment metadata, derived numerical outputs, reproduction code, and related validation and provenance records. Public accessibility does not itself establish redistribution rights or eliminate re-identification risk; no claim is made here about source-author consent, IRB or exemption status, platform redistribution permission, or complete legal or Terms-of-Service verification.

AI USE STATEMENT

The research motivation, core scientific questions, methodological design, and experimental plan of this study were proposed and determined by the human authors. Decisions about how experimental results were interpreted and used as evidence, as well as the final scientific conclusions, were also made by the authors. On this basis, generative AI tools were used primarily to assist with conceptual organization and methodology discussion, code implementation and inspection of experimental outputs, figure creation and document preparation, and manuscript drafting, restructuring, translation, and language polishing. They were also used to assist with literature search and summarization and with parts of the analysis and interpretation of experimental results. All AI-assisted code, numerical results, citations, claims, figures, and manuscript text were reviewed and verified by the authors. The authors take responsibility for the research design, experimental results, scientific argumentation, and the final content of the manuscript.

REPRODUCIBILITY STATEMENT

The anonymous supplementary materials provide text-free alignment metadata, sanitized item-level numerical outputs for all four evaluated models, the analysis and scoring implementation, numerical validation records, sanitized run manifests, and deterministic scripts for regenerating Tables 2–3, the numeric appendix tables, and Figure 2. These materials support independent recomputation of the reported aggregate preference rates, pairwise disagreement rates, Jaccard similarities, and subgroup summaries. Exact source-derived sentence text and model binaries are not redistributed in the review artifact; therefore, a full end-to-end inference rerun additionally requires separately available paired input text and model artifacts.

REFERENCES

Kris Cao and Laura Rimell. You should evaluate your language model on marginal likelihood over tokenisations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2104–2114. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.161. URL https://aclanthology.org/2021.emnlp-main.161/.

Nadezhda Chirkova, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Should you marginalize over possible tokenizations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–12. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.1. URL https://aclanthology.org/2023.acl-short.1/.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 263–277. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.18/.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1374/.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. doi: 10.1162/tacl_a_00115. URL https://aclanthology.org/Q16-1037/.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. A systematic assessment of language models with linguistic minimal pairs in Chinese. *Transactions of the Association for Computational Linguistics*, 14:755–771, 2026. doi: 10.1162/tacl.a.648. URL https://aclanthology.org/2026.tacl-1.34/.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. Lost in the mix: Evaluating LLM understanding of code-switched text. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.2080/.

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. Token-length bias in minimal-pair paradigm datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 16224–16236. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.1410/.

Jannis Vamvas and Rico Sennrich. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 58–68. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.blackboxnlp-1.5/.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. URL https://aclanthology.org/2020.tacl-1.25/.

A Dataset Construction and Filtering

Candidate sentences were drawn from publicly accessible Chinese AI and computer-science discussions on CSDN and Zhihu between May 2021 and November 2025. Exact deduplication, length and topic filters, terminology-density checks, and manual inspection produced a working corpus from which technical terms and sentence contexts were extracted. A bilingual terminology table supplied Chinese counterparts, and each candidate pair placed the English term and its counterpart into matched versions of the same Chinese-matrix sentence.

The candidate collection contains 1,706 source-derived pairs. Target terms are normalized by lowercasing and removing characters outside `[a-z0-9]`, yielding 1,344 non-empty normalized terms. A pair receives the strict single-replacement label when it contains one changed block that covers the target, the English target is absent from the Chinese realization, and the changed block contains at most two off-target characters on the English side. This rule identifies 1,168 pairs before duplicate removal; removing one duplicate gives the 1,167 analysis pairs covering 934 normalized terms.

Each strict pair changes one terminology realization while keeping the surrounding technical sentence as closely matched as the source construction permits. The filter excludes pairs with unresolved alignment, duplicated alternatives, or other quality-control failures. The released metadata expose pair identifiers, subgroup and span-validity flags, and aggregate quality-label counts. Source sentences and term-level normalization records are not redistributed; rerunning the original text-level filtering requires the separately held source records.

Table 4: Deterministic labels assigned to the 1,706 candidate pairs before duplicate removal.

Pair-quality label	Count
Strict single replacement	1,168
Localized replacement with minor edits	122
Target retained on Chinese side	294
Non-local or multi-edit pair	112
Missing target metadata	3
Identical pair	7
Candidate total	1,706

These labels encode reproducible string-level filtering criteria; linguistic equivalence remains a separate human judgment.

B Extended Scoring Protocols

The main text compares four core definitions formed by crossing scoring scope (whole sentence versus changed span) with token aggregation (cumulative versus mean). Additional diagnostic protocols include segmented target scoring, target-plus-suffix scoring, and alternative normalization conventions. They are outside the fixed four-definition comparison in the main text and are not included in the released item-level numerical outputs.

Table 5 lists the core scoring scopes alongside examples of additional diagnostic regions.

Table 5: Selected scoring regions used in the core comparison and additional diagnostics.

Scoring region	Scored evidence
Complete sentence	All valid causal next-token log probabilities
Original changed span	Tokens overlapping the changed span in full-sentence tokenization
Segmented target	Target encoded after a shared prefix
Target plus suffix	Target and the following retained suffix

C Tokenization, Implementation, and Run Provenance

Target-span scores use the changed-span mask identified on the original full-sentence tokenization. The target strings are not independently re-tokenized. Whole-sentence scores are defined for all 1,167 strict pairs; two pairs (IDs 1222 and 1446) do not yield a valid changed span under the fixed construction rule, so span-based analyses use 1,165 pairs.

The Qwen2.5-7B and Llama 3.1 8B evaluations use local Ollama GGUF Q4_K_M assets. These runs are reported as quantized local evaluations, not as exact reproductions of the corresponding full-precision Hugging Face checkpoints. Run manifests record the model tag and file identity, tokenizer and backend settings, dataset identifiers, release-code hashes, and available software-version details. Exact versions that could not be recovered are explicitly marked as such.

The backend validation checked causal next-token alignment against the local server and verified token-byte reconstruction. Recomputing the full-sentence and target-span sum and mean scores from the exported per-token log probabilities produced no discrepancies. The alignment and reconstruction checks passed for all 2,334 sentence realizations for each larger-model evaluation.

Table 6: Implementation checks for the two larger-model evaluations.

Check	Qwen2.5-7B	Llama 3.1 8B
Scored sentence realizations	2,334	2,334
Tokenizer/text reconstruction failures	0	0
Max. reported–recomputed score difference	0.0	0.0
Server alignment absolute difference	2.05×10^{-5}	4.23×10^{-5}

Table 7: Run provenance for the two larger-model evaluations used in the main results.

Field	Recorded value
Qwen run	`qwen2.5:7b`, GGUF Q4_K_M
Llama run	`llama3.1:8b`, GGUF Q4_K_M
Dataset	1,167 strict pairs
Inference	Local llama.cpp/Ollama backend
Output basis	Item-level CSV outputs and run manifests

D SPECIALIZED AND GENERAL TERMS

The strict set contains 1,079 specialized-term pairs and 88 general-term pairs. We retain this split as a descriptive secondary analysis. It helps characterize where the four scoring definitions differ across the analysis set, but it is not used to define the main claim.

The subgroup direction and magnitude vary across models and scoring definitions. The two pairs without valid target spans are both specialized, giving valid counts of 1,079/1,077 for specialized whole-sentence/span scores and 88/88 for general scores.

Table 8: English-preference rates by subgroup for the two larger-model evaluations. N reports the valid whole-sentence/span counts.

Model/group	N (whole/span)	Full mean	Full sum	Span mean	Span sum
Qwen specialized	1,079/1,077	13.99%	5.19%	18.11%	2.60%
Qwen general	88/88	12.50%	5.68%	19.32%	3.41%
Llama specialized	1,079/1,077	8.16%	9.45%	6.41%	4.83%
Llama general	88/88	4.55%	9.09%	13.64%	7.95%