G-IdiomAlign: A Gloss-Pivoted Benchmark for Cross-Lingual Idiom Alignment

Fengying Ye[1,*] **Yanming Sun**[1,*] **Runzhe Zhan**[1]
Lidia S. Chao[1] **Zheqi Zhang**[2] **Derek F. Wong**[1,†]
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2]Faculty of Arts and Humanities, University of Macau
nlp2ct.{fengying, yanming, runzhe}@gmail.com,
{lidiasc, derekfw}@um.edu.mo

Abstract

Idioms are difficult to transfer across languages due to their non-compositionality and weak surface-form grounding, making literal mappings unreliable. We present **G-IdiomAlign**, a gloss-pivoted benchmark where each idiom is anchored by an English gloss from Wiktionary. We further construct a high-confidence reference alignment set for reproducible evaluation. G-IdiomAlign supports two protocols: (1) a controlled Multiple-Choice Idiom Equivalence with typed distractors for error attribution; and (2) a Gloss-Contrastive Generation contrasting *No-gloss* and *With-gloss* inputs to isolate the effect of an explicit semantic pivot. Across diverse LLMs, a bias to literal translation is a dominant failure mode, especially when the target is a low-resource language. Glosses consistently improve Gloss-Contrastive Generation under an embedding-based semantic proxy, but performance remains modest, indicating substantial headroom in the open output space. Subsequent analysis on Qwen3-8B further suggests that cross-condition differences are concentrated more in attention heads than in layers, while better *With-gloss* generations coincide with stronger gloss anchoring[1].

1 Introduction

Idioms pose a persistent challenge for cross-lingual meaning transfer because their figurative meanings are non-compositional and culturally grounded, making word-by-word composition unreliable (He et al., 2025). Recent evidence suggests that large language models (LLMs) often over-index on surface statistical cues (such as collocational frequency or sentence probability) rather than recovering figurative intent from context (Mi et al., 2025; Yang et al., 2025c). Accurate cross-lingual idiom alignment is crucial not only for cross-cultural communication and machine-assisted localization, but also as a litmus test for whether LLMs genuinely grasp culturally embedded semantics beyond surface patterns, motivating evaluation protocols that target idiom-to-idiom semantic equivalence over mere lexical overlap.

I_{source}	I_{target}	G_{source}	G_{target}
一鋪清袋	lose one's shirt	lose all money at once	lose all of the money
mouton de Panurge	follow ... off a cliff	blindly follow others	follow a leader blindly
守口如瓶	tenir sa langue	keep one's mouth shut	hold one's tongue

Table 1: Example idiom pairs from **G-IdiomAlign** across different language pairs. I and G denote idioms and their English glosses; subscripts indicate languages.

However, existing resources offer limited support for controlled and diagnostic evaluation of cross-lingual idiom-to-idiom equivalence. Strong systems frequently produce literal, partial, or missing idiom renderings (Yang et al., 2025c), yet current datasets lack unified benchmarks with standardized protocols for systematic error attribution. To enable explicit and comparable semantic grounding across languages, we adopt English glosses as a shared *semantic pivot*, modeling a resource-augmented setting where models can leverage external semantic support (e.g., lexicons or knowledge bases). We introduce a contrastive setup between *No-gloss* and *With-gloss* inputs to isolate the effect of this explicit semantic signal.

In this work, we introduce **G-IdiomAlign**, a gloss-pivoted idiom alignment benchmark across nine core languages, with coverage of four languages from underrepresented families, where each idiom is linked to a meaning-equivalent English gloss from Wiktionary (examples are shown in Table 1). We construct a high-confidence refer-

[*]Equal Contribution.
[†]Corresponding Author.
[1]Dataset: `https://github.com/NLP2CT/G-IdiomAlign`

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38720–38739

ence set via a precision-first pipeline that combines distribution-aware filtering with bidirectional one-to-one constraints. On top of this benchmark, we provide two complementary evaluation settings: a Multiple-Choice Idiom Equivalence task with typed distractors, and a Gloss-Contrastive Generation under *No-gloss* and *With-gloss* inputs.

Across diverse LLMs, models show a pervasive bias to literal translations. Although adding glosses yields consistent but limited improvements under an embedding-based semantic proxy, this underscores the difficulty of producing canonical, meaning-equivalent idioms in an unconstrained space. Subsequent attention-based correlational analyses on Qwen3-8B further suggest that improved *With-gloss* generations align with stronger gloss anchoring, with cross-condition differences concentrating mainly at the level of attention heads rather than broad layer-level shifts. Our analysis positions G-IdiomAlign as a foundation for future work on robust cross-lingual idiom modeling.

Our contributions are as follows: (1) We release **G-IdiomAlign**, a gloss-pivoted dataset covering 36 language pairs across nine core languages (18,785 idiom pairs), filtered via a bidirectional pipeline to ensure high semantic equivalence. (2) We establish two diagnostic protocols for cross-lingual idiom alignment: Multiple-Choice Idiom Equivalence with typed distractors, and Gloss-Contrastive Generation that contrasts *No-gloss* and *With-gloss* inputs to test semantic grounding. (3) We reveal a widespread bias towards literal translation in LLMs and provide attention-based evidence linking gloss usage to improved semantic anchoring.

2 Related Work

2.1 Idiom Benchmarks

Existing idiom benchmarks support two core tasks: detection, which identifies whether a phrase is used idiomatically, and disambiguation, which resolves whether an expression should be interpreted literally or figuratively. Detection datasets test whether a model distinguishes idiomatic from literal usages, ranging from linguist-curated contrastive sets (Mi et al., 2025) to English test suites such as IdioTS (De Luca Fornaciari et al., 2024), multilingual benchmarks ID10M (Tedeschi et al., 2022) and CLCL framework (Zhou et al., 2023). Disambiguation datasets including EPIE (Saxena and Paul, 2020), MAGPIE (Haagsma et al., 2020), and MultiCoPIE (Sentsova et al., 2025) label potential idiomatic expressions with literal versus idiomatic readings, supporting contextual sense selection (Fakharian and Cook, 2021; Zhou et al., 2021). Complementary resources broaden coverage further: LIdioms (Moussallem et al., 2018) links idioms across languages as linked data, and Fu et al. (2025) evaluate Chinese idioms across multiple competencies. While these efforts are valuable for idiom recognition and interpretation, they do not directly target *idiom-to-idiom* meaning-equivalence alignment within a unified cross-lingual evaluation.

2.2 Idiom Alignment

Cross-lingual idiom alignment remains challenging because figurative meanings often diverge from literal forms and are shaped by language- and culture-specific conventions (Moussallem et al., 2018; Donthi et al., 2025). Recent evaluations confirm persistent failures in both NMT systems and LLMs (Yang et al., 2025c; Sun et al., 2026), prompting approaches that decompose translation into semantic analysis and candidate selection (Qian, 2024) or inject external signals, such as retrieval-augmented MT with loss weighting (Liu et al., 2023) or multilingual idiom knowledge bases (Li et al., 2024). However, these methods often rely on surface-level cues: Sentsova et al. (2025) report substantially higher performance on idioms with direct English lexical counterparts, and cross-lingual evaluations note strong prompt sensitivity and performance gaps in low-overlap language pairs (Khoshtab et al., 2025). This reliance is exacerbated in retrieval-based alignment frameworks like bilingual lexicon induction (BLI), which formulate cross-lingual matching as nearest-neighbor search over candidate sets (Li et al., 2023). Such approaches are prone to false positives (Ding et al., 2024), unless constrained by precision-oriented criteria like bidirectional agreement. To address these limitations, we move beyond surface-driven retrieval by anchoring alignment in meaning-equivalent English glosses and adopt bidirectional constraints to ensure high-precision idiom-to-idiom pairing, thus enabling controlled evaluation that isolates semantic equivalence from lexical shortcuts.

3 G-IdiomAlign

We introduce **G-IdiomAlign**, a gloss-pivoted benchmark for cross-lingual idiom alignment across nine core languages, with coverage of four languages from low-resource language families.

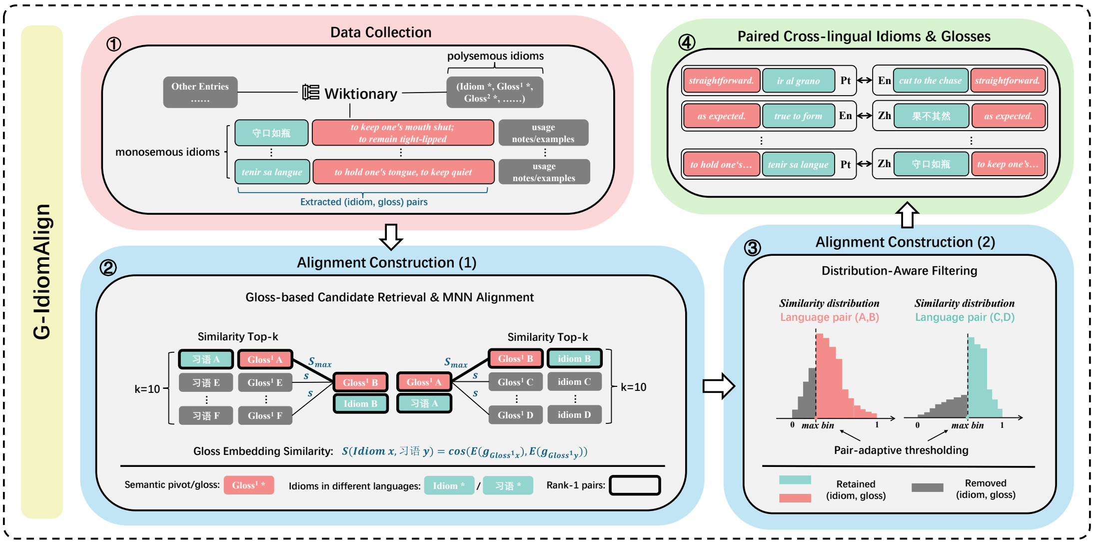

Figure 1: Overview of the G-IdiomAlign construction pipeline. Using English glosses as a shared semantic pivot, we extract idiom entries and core glosses from Wiktionary, retrieve top-k candidates in a gloss-embedding space, keep MNN pairs, and apply a pair-specific distribution-aware filter, yielding the final G-IdiomAlign benchmark.

English glosses from Wiktionary[2] serve as a shared semantic pivot, supporting cross-lingual meaning comparison while mitigating shortcuts based on surface lexical overlap. We construct G-IdiomAlign with a precision-first, staged construction pipeline: (i) extract idiom entries and core glosses (excluding usage notes/examples), (ii) retrieve top-k candidates in a shared gloss-embedding space, (iii) retain mutual nearest neighbor (MNN) pairs via bidirectional agreement, and (iv) apply distribution-aware filtering to produce a high-confidence alignment set. The resulting benchmark is intended for evaluation and diagnostic analysis. Figure 1 summarizes the construction pipeline.

3.1 Language Coverage & Data Collection

Language Scope. G-IdiomAlign covers nine core languages: De, En, Es, Fi, Fr, Ja, Pl, Zh, and Pt, for which our extraction pipeline yields sufficient high-quality aligned pairs. To broaden language coverage, we further include four languages (Arabic, Korean, Thai, and Vietnamese) and report their results separately in Appendix A.

Collection Pipeline. We collect idiom entries from language-specific Wiktionary category pages and extract a cleaned core gloss from each entry's sense definition, excluding auxiliary material such as usage notes and examples; see Appendix B for implementation details. These glosses provide a consistent meaning description and function as the semantic pivot throughout construction.

Single-Sense Filtering. To preserve interpretability of the reference, we keep idioms with a single Wiktionary sense (one gloss) and remove polysemous entries. This avoids one-to-many sense correspondences that would make idiom-to-idiom equivalence ambiguous at construction time. The resulting reference set is smaller but cleaner, supporting more controlled evaluation and diagnosis.

Gloss-based Candidate Retrieval. For each directed language pair $A \to B$, we embed the glosses associated with idioms in both languages using the OpenAI text-embedding-3-large (OpenAI, 2024). For a source idiom x with gloss g_x and a candidate idiom y with gloss g_y, we define gloss similarity as

$$s(x, y) = \cos\big(E(g_x), E(g_y)\big),$$

where $E(\cdot)$ denotes the embedding function. For each x, we retrieve the top-k candidates in B by $s(x, y)$ with $k = 10$, producing a candidate set for subsequent bidirectional filtering. Although final alignments are determined by rank-1 agreement (see below), using $k > 1$ improves candidate coverage and robustness to embedding noise before enforcing one-to-one constraints.

MNN Alignment. To obtain unambiguous evaluation pairs, we enforce a one-to-one matching constraint via mutual nearest neighbors (MNN). We retain a pair (x, y) if and only if x and y are rank-1 nearest neighbors of each other under both directions ($A \to B$ and $B \to A$). This bidirectional

[2] https://www.wiktionary.org/

Pair	N	%	Pair	N	%	Pair	N	%	Pair	N	%
De–En	520	2.77	En–Fi	693	3.69	Es–Pl	888	4.73	Fr–Pl	329	1.75
De–Es	456	2.43	En–Fr	376	2.00	Es–Pt	429	2.28	Fr–Pt	209	1.11
De–Fi	353	1.88	En–Ja	336	1.79	Es–Zh	1028	5.47	Fr–Zh	343	1.83
De–Fr	238	1.27	En–Pl	1182	6.29	Fi–Fr	267	1.42	Ja–Pl	345	1.84
De–Ja	229	1.22	En–Pt	534	2.84	Fi–Ja	251	1.34	Ja–Pt	206	1.10
De–Pl	447	2.38	En–Zh	1782	9.49	Fi–Pl	589	3.14	Ja–Zh	416	2.21
De–Pt	280	1.49	Es–Fi	569	3.03	Fi–Pt	333	1.77	Pl–Pt	432	2.30
De–Zh	509	2.71	Es–Fr	336	1.79	Fi–Zh	655	3.49	Pl–Zh	1156	6.16
En–Es	1114	5.93	Es–Ja	306	1.63	Fr–Ja	186	0.99	Pt–Zh	463	2.46

Table 2: G-IdiomAlign language-pair composition. **N** denotes the count of aligned idiom pairs and **%** denotes the proportion of the dataset (out of all aligned pairs). We report each pair once using a canonical ordering.

criterion removes asymmetric or many-to-one associations that may arise from retrieval artifacts, ensuring that retained alignments reflect strong mutual semantic correspondence.

Distribution-Aware Filtering. Since similarity score scales differ substantially across language pairs, fixed global thresholds can be poorly calibrated. Moreover, even under MNN, nearest-neighbor retrieval always returns a best match within the dataset, which can force alignments even when no true equivalent exists, leading to spurious pairs. For example, a Chinese idiom "洞房花燭夜" (gloss: the wedding night) may be aligned with the English idiom "white marriage" (gloss: an unconsummated marriage): although their glosses share salient words (wedding and marriage), the underlying meanings are not equivalent. Accordingly, we apply a language-pair-specific, parameter-light cutoff to remove weak matches while preserving high-confidence alignments.

For each language pair, we collect the rank-1 similarity scores of MNN-confirmed pairs and discretize similarity scores within-pair range into 10 equal-width bins. Let b denote the modal bin. We retain pairs whose scores fall in bin b or higher, using the lower edge of the modal bin as a cutoff. Similarity scores are used here as a diagnostic signal for relative strength within each language pair, rather than as an absolute criterion of semantic correctness (details are shown in Appendix C).

To ensure deterministic reporting, we compute similarity scores using a fixed canonical direction for each unordered language pair, while the MNN criterion itself is always enforced bidirectionally.

3.2 Benchmark Statistics

G-IdiomAlign comprises 18,785 aligned idiom pairs across 36 *unordered* language pairs drawn from nine languages. Although alignments are reported without direction, each pair supports evaluation in either direction (e.g., Zh→En or En→Zh). For reporting and aggregation, each unordered language pair is listed once using a canonical ordering. Pair sizes range from 186 to 1,782, with a median of 422 (interquartile range: 323–574); the full breakdown is provided in Table 2.

3.3 Alignment Quality Evaluation

We assess the semantic alignment quality of G-IdiomAlign using both human evaluation and LLM-based majority voting. Sampled pairs are rated on a 3-point scale: 2 denotes equivalent meaning and interchangeability across contexts; 1 denotes partial equivalence, where meanings are close but differ in tone, intensity, or pragmatics; and 0 denotes non-equivalence, where lexical or topical relatedness does not imply semantic equivalence.

For Zh–En idiom pairs, we randomly sample 200 pairs and evaluate them with native speakers and senior Ph.D. students with expertise in relevant languages. For the remaining language pairs, we sample 50 pairs per language pair and score them independently with GPT-5.1 (OpenAI, 2026), Gemini-2.5-Pro (Gemini Team, 2025), and Claude-4.5-Haiku (Anthropic, 2025). LLM judges are prompted to follow the same annotation instructions as human annotators. We use majority voting as the final label; when all judges disagree, we assign score 1 to reflect partial equivalence.

We report strict accuracy (only score-2 pairs), and lenient accuracy (score-1 and score-2 pairs). Across non-Zh–En language pairs, the LLM-based evaluation yields a mean strict accuracy of 0.685 and a lenient accuracy of 0.923, where each language pair is treated as one observation. The corresponding 95% confidence intervals are computed using a t-interval over language pairs (strict: [0.645, 0.724]; lenient: [0.907, 0.940]). Details are provided in Appendix D.

In addition, performance varies substantially across language pairs. High-resource or closely related pairs such as En–De, En–Es, En–Pt, and Pt–Es achieve very high strict accuracy (up to 0.96), with most annotations assigned score 2. In contrast, more distant pairs such as De–Ja, Fr–Ja, and Zh–Es show lower accuracy and a larger proportion of score 1 and score 0 cases, reflecting the greater difficulty of establishing idiomatic equivalence across typologically distant languages.

3.4 Similarity Characterization

Complementing the alignment quality evaluation, we analyze gloss-based similarity scores to understand how embedding-space signals support and characterize the constructed alignments.

Overall distribution. Across 18,785 aligned pairs, construction-time similarity scores span a wide range (min $= 0.368$, max ≈ 1.0) with moderately high central tendency (mean $= 0.670$, median $= 0.651$). Near-saturation scores are rare and typically correspond to highly formulaic or nearly identical glosses (see Appendix E.1).

Threshold sensitivity. To assess the effect of encoder calibration on coverage, we sweep nine thresholds t over the shared overlap interval of the two encoders' score distributions and count pairs with $sim \geq t$ (see Appendix E.2). In Figure 2, both encoders produce monotonic retention curves but diverge substantially at the same absolute threshold, reflecting calibration and scaling differences rather than semantic disagreement. This supports treating similarity as a relative diagnostic signal and motivates distribution-aware filtering.

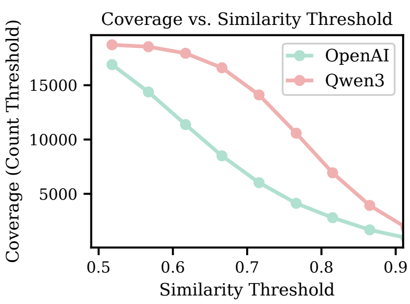

Figure 2: Retention curves for OpenAI text-embedding-3-large and Qwen3-Embedding-8B as the similarity threshold t varies, illustrating encoder-dependent calibration effects under fixed absolute thresholds.

Embedding consistency and calibration. Embedding consistency refers to the agreement in relative similarity structure across different embedding models. When we recompute similarities with independent multilingual encoder, Qwen3-Embedding-8B (Yang et al., 2025b), we observe a systematic increase in absolute cosine similarity while maintaining strong agreement in relative structure (Pearson $r = 0.807$, Spearman $\rho = 0.784$; see Appendix E.3). This suggests that absolute similarity values are encoder-dependent, while relative similarity structure is largely preserved.

4 Experiments

We evaluate cross-lingual idiom alignment under two complementary settings. Task 1 formulates alignment as a controlled multiple-choice problem, enabling fine-grained diagnosis of error patterns through typed distractors. Task 2 evaluates open-ended target-idiom generation in a large output space and contrasts *No-gloss* and *With-gloss* inputs to assess the effect of an explicit semantic pivot under semantic ambiguity and surface-form mismatch. These settings support reproducible quantitative comparison and reveal recurring failure modes in cross-lingual idiom alignment.

4.1 Task 1:Multiple-Choice Idiom Equivalence

Task formulation. Given a source-language idiom, the model selects the meaning-equivalent target-language option from a 4-way candidate set. Each instance contains one canonical target idiom from G-IdiomAlign and three typed distractors, enabling error analysis by distractor type in addition to accuracy. To reduce positional bias, we shuffle the option order per instance and store option-type labels (reference, LT, LC, CA; defined below). The model outputs a single choice (A-D) under greedy decoding (temperature $= 0$, top-$p = 1$). We evaluate 30 direction settings in total, including 16 directions with *high-resource* target languages (Chinese or English; 8 each) and 14 additional directions with other target languages. Here a *direction* is an ordered mapping from a source language to a target language, and we group directions by the target language (Zh-target, En-target, and other targets).

Candidate construction. The correct (reference) option is the canonical target idiom from G-IdiomAlign. We construct three types of target-language distractors: *Literal Translation Trap* (LT),

Model	Zh-target		En-target		Other targets		Overall	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
DeepSeek-V3.2 (NT)	56.84	56.96	57.84	56.16	47.39	47.53	52.70	53.65
DeepSeek-V3.2 (T)	69.37	69.18	66.70	65.38	53.92	54.09	61.45	63.02
Gemini-2.5-Pro	67.77	67.57	66.53	64.95	62.82	62.91	65.13	65.17
Claude-4.5-Haiku	56.80	57.30	55.54	53.31	44.03	43.44	50.50	51.47
Ministral-8B-Instruct	28.63	28.51	30.29	28.97	27.78	27.76	28.68	28.43
Qwen3-8B (NT)	32.69	33.72	33.95	33.62	24.02	23.99	28.98	30.56
Qwen3-8B (T)	41.20	41.89	39.57	39.10	31.29	31.39	36.14	37.55

Table 3: Multiple-choice accuracy aggregated by target-language groups. Micro is instance-weighted within each target group (Zh-/En-/Other-target), while Macro averages over directions. All numbers are percentages.

a word-for-word translation of the source idiom; *Lexical Cue Trap* (LC), a target-language idiom that shares a *partial* lexical cue with the literal translation (typically one salient content word) but conveys an unrelated meaning; and *Contextual Association Trap* (CA), a target-language idiom that is contextually plausible yet semantically opposite. Distractors are generated using Qwen-Max under a unified prompt with explicit type constraints; the details are provided in the Appendix F.1.

Validity checks for LLM-generated distractors. To assess whether LLM-generated distractors introduce superficial shortcuts, we conduct an option-only control and a manual validity check. The results show no significant preference for the gold option over chance and confirm 81.5% distractor-type validity. Details are provided in Appendix F.2.

4.2 Task 2: Gloss-Contrastive Generation

Task formulation. In the open-ended setting, the model is given a source-language idiom and must generate a meaning-equivalent idiom in the target language. We compare two input conditions: *No-gloss* (source idiom only) and *With-gloss* (source idiom plus its English gloss), where the gloss provides an explicit semantic pivot. We evaluate Task 2 on all 72 directions available in G-IdiomAlign, using greedy decoding (temperature $= 0$, top-p $= 1$). We require models to output exactly one target-language idiom with no additional explanation, enabling deterministic parsing and automatic scoring (see Appendix G).

Automatic evaluation. Because multiple outputs can be valid in open-ended generation, we use an embedding-based semantic matching proxy for coarse-grained scoring. We embed the model output and the canonical target idiom in G-IdiomAlign using Qwen3-Embedding-8B and compute cosine similarity. We report Acc@t, counting a prediction as correct if the similarity exceeds a threshold t in the same embedding space. We choose two operating points, $t = 0.70$ and $t = 0.80$: 0.70 is a more permissive threshold, while 0.80 is a stricter threshold close to the median similarity of canonical aligned pairs in G-IdiomAlign under Qwen3-Embedding-8B (median ≈ 0.78). This proxy supports aggregate comparison but is not a definitive correctness criterion; for example, it can undercount valid synonymous idioms that diverge from the canonical reference. We therefore treat embedding similarity as a high-confidence semantic indicator rather than a complete estimate of idiom-form correctness, and report a small human evaluation together with auxiliary surface-form metrics (EM/BLEU/ChrF) in Appendix J. We interpret Task 2 results primarily as comparative trends.

4.3 Models

We evaluate several open-source LLMs and proprietary, including DeepSeek-V3.2 (T/NT) (DeepSeek-AI et al., 2025), Gemini-2.5-Pro, Claude-4.5-Haiku, Ministral-8B-Instruct (Mistral AI team, 2024), and Qwen3-8B (T/NT) (Yang et al., 2025a). T denotes *Thinking mode* and NT is *No thinking mode*. During dataset preprocessing, we embed glosses with text-embedding-3-large to select high-confidence reference alignments. For Task 1, we generate typed distractors using Qwen-Max (Team, 2025). For Task 2 automatic scoring, we use Qwen3-Embedding-8B. We run open-source models on a NVIDIA A40 GPU, while proprietary models are accessed via official APIs.

5 Results and Analysis

5.1 Multiple-Choice Idiom Equivalence

5.1.1 Overall Accuracy Across Target Groups

Table 3 reports Task 1 accuracy by target-language group. Across models, performance is consistently

Model	No-gloss MeanSim	Acc@0.70	Acc@0.80	With-gloss MeanSim	Acc@0.70	Acc@0.80	ΔAcc@0.80
DeepSeek-V3.2 (NT)	68.30	41.99	19.90	70.94	48.35	23.57	3.66
DeepSeek-V3.2 (T)	72.31	49.01	24.66	**74.24**	54.90	**29.63**	4.97
Gemini-2.5-Pro	68.71	44.06	19.70	72.54	51.09	27.11	7.41
Claude-4.5-Haiku	70.90	45.47	20.47	72.94	52.16	25.66	5.19
Ministral-8B-Instruct	68.30	37.97	12.96	71.69	50.00	21.22	8.28
Qwen3-8B (NT)	67.54	36.29	11.59	70.69	47.09	18.80	7.22
Qwen3-8B (T)	68.53	39.03	14.04	71.40	49.16	20.86	6.82

Table 4: Gloss-Contrastive Generation results. MeanSim ($100 \times sim$) and Acc@t are direction-averaged (macro) over 72 language directions; Acc@t counts a prediction as correct if cosine $sim \geq t$. We report $t = 0.70$ and $t = 0.80$. ΔAcc@0.80 is the *With-gloss* minus *No-gloss* improvement. Bold indicates the best performance.

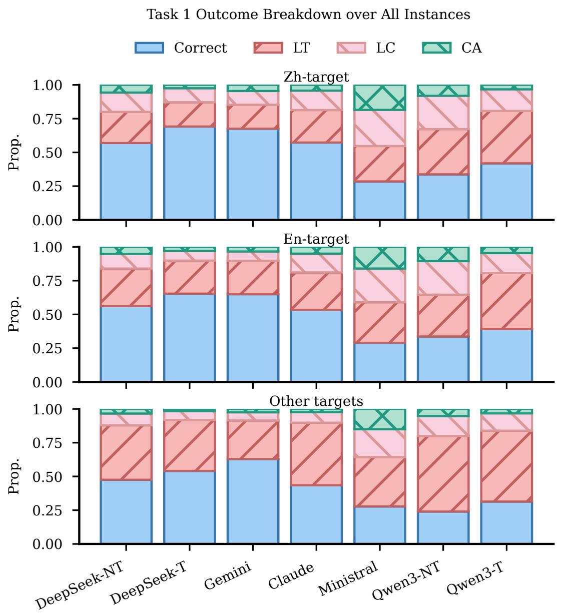

Figure 3: Task 1 outcomes by target-language regime. Stacked bars (normalized within each regime) decompose specific instances into Correct predictions and LT/LC/CA, highlighting differences between high-resource targets (Zh/En) and other target languages.

lower on Other targets directions than on Zh-target or En-target, suggesting that idiom alignment is more challenging when the target language is not a high-resource language such as Chinese or English. This gap is most pronounced for lower-performing models. Gemini-2.5-Pro achieves the strongest overall performance, particularly on Other targets. Enabling *thinking mode* yields consistent gains across models, with clear improvements for DeepSeek-V3.2 and Qwen3-8B.

5.1.2 Outcome Decomposition

Figure 3 decomposes Task 1 outcomes by target-language regime into *Correct* predictions and three distractor types (LT/LC/CA); proportions are in Appendix H.1. These patterns are further illustrated with representative examples and error cases in Appendix H.2, which provide concrete instances of LT-, LC-, and CA-driven errors. The Other targets regime has fewer correct predictions, consistent with lower accuracy.

Across models, *Literal Translation Trap* (LT) dominates errors, especially for lower-resource targets, indicating stronger literal-transfer attraction. In contrast, *Zh-target* directions show more *Lexical Cue* (LC) errors, suggesting partial lexical overlap misleads models when Chinese is the target. Lower-performing models also exhibit higher *Contextual Association* (CA) rates, particularly in Other targets.

Enabling *thinking mode* improves accuracy for DeepSeek-V3.2 (reducing LT and LC errors) and Qwen3-8B (mainly decreasing LC and CA, with LT still dominant). Overall, literal-translation attraction remains the primary challenge in cross-lingual idiom alignment.

5.2 Gloss-Contrastive Generation

Overall performance. Table 4 summarizes Task 2 results under the semantic-similarity proxy. Providing an English gloss (*With-gloss*) improves MeanSim and Acc@t for every model, with consistent gains at the stricter threshold Acc@0.80, suggesting that glosses help constrain generation toward the intended meaning. Despite this, Acc@0.80 remains modest even with gloss, highlighting the difficulty of producing meaning-equivalent idioms in an unconstrained output space.

Under *With-gloss*, DeepSeek-V3.2 (T) achieves the strongest overall performance. For models with both variants, enabling thinking yields consistent improvements, with the advantage more apparent at higher similarity thresholds. Results at additional thresholds are reported in Appendix I.1. Representative examples and error cases for open-ended generation are provided in Appendix I.2, illustrating both correct outputs and acceptable paraphrases that may be under- or over-estimated by the similarity-based metric.

5.3 Attention-based Diagnostics

We introduce attention-based diagnostics for Task 2 to characterize how the model allocates attention over *input spans* during decoding and how these allocations relate to output quality. Throughout, we treat attention strictly as a **correlational diagnostic signal**. All analyses are conducted on **Qwen3-8B** with greedy decoding (temperature $= 0$, top-$p = 1$). Post-softmax self-attention weights are extracted using `TransformerLens` hooks[3].

For each input condition, we annotate 200 generations (400 total) and exclude Type 0 invalid outputs (empty, garbled, or not interpretable as meaningful target-language text), yielding $n = 189$ valid instances in *With-gloss* and $n = 189$ in *No-gloss*. Table 5 reports the outcome breakdown by error type. Two collaborators independently label the outputs (Cohen's $\kappa = 0.81$) and resolve disagreements by discussion. We use the following error taxonomy: Type 2 (T2, literal word-by-word translation missing the idiom's figurative meaning), Type 3 (T3, meaning-correct but non-idiomatic), and Type 4 (T4, meaning-incorrect and not a word-by-word literal translation).

Condition	Valid	Correct	Wrong	T2 / T3 / T4
With-gloss	189	123	66	6 / 30 / 30
No-gloss	189	53	136	60 / 40 / 36

Table 5: Outcome composition for the annotated subset (Type 0 excluded). T2/T3/T4 denote the breakdown of error types within Wrong outputs.

Head/layer-level structural overlap. To assess whether cross-condition differences are driven more by head changes or layer shifts, we compare the cross-condition overlap of the salient heads and salient layers (Table 6; see Appendix L for details).

As shown in Table 6, head overlap is substantially lower than layer overlap overall ($J_{\text{heads}} = 0.32$ vs. $J_{\text{layers}} = 0.90$). This pattern also holds across subsets, although layer overlap is lower for Correct-only than for Wrong-only outputs. These results suggest that cross-condition differences are expressed more strongly through head-level reconfiguration than broad layer-level shifts: the two conditions largely recruit similar layers but differ in which heads within those layers are salient. Lower layer overlap for Correct-only outputs further suggests less shared layer-level structure across conditions for correct than for wrong cases.

[3] https://github.com/TransformerLensOrg/TransformerLens

Subset	J_{heads}	J_{layers}
All	0.32	0.90
Correct-only	0.32	0.76
Wrong-only	0.37	0.90

Table 6: Cross-condition overlap between *With-gloss* and *No-gloss* for salient head sets and salient layer sets (Jaccard; higher indicates more overlap).

Token-level diagnostics. Let $\bar{A}(k)$ denote the aggregated post-softmax attention mass assigned to input key token k, computed by averaging attention over layers and heads and over template-defined generation positions corresponding to the content-bearing segment. Here k ranges over key-token positions in the full prompt sequence. We interpret $\bar{A}(k)$ as the model's average attention mass assigned to token k during the generation of the content-bearing output (see Appendix M).

Using tokenizer offset mapping, we map the idiom span and (when available) the gloss span to token-index sets K_{idiom} and K_{gloss} in the full prompt sequence. We then define:

$$\mathrm{IAR} = \sum_{k \in K_{\text{idiom}}} \bar{A}(k)$$

IAR (Idiom Attention Ratio) is the total attention mass on the idiom span; larger values indicate stronger concentration on idiom tokens.

$$\mathrm{GAR} = \sum_{k \in K_{\text{gloss}}} \bar{A}(k) \quad (\textit{With-gloss})$$

GAR (Gloss Attention Ratio) is the total attention mass on the gloss span (defined only under *With-gloss*); larger values indicate stronger anchoring to the provided gloss.

$$\mathrm{DR} = 1 - \mathrm{IAR} - \mathrm{GAR}$$

DR (Diffuse Ratio) captures the residual attention mass outside the tracked spans; larger values indicate greater allocation to other context tokens. Under *No-gloss*, GAR is defined as 0, so $\mathrm{DR} = 1 - \mathrm{IAR}$.

Metric	Analysis	*With-gloss*				*No-gloss*			
		Correct	Wrong	δ	q	Correct	Wrong	δ	q
IAR	CvsW	0.03 (0.01)	0.03 (0.01)	-0.10	0.250	0.06 (0.02)	0.06 (0.02)	0.18	0.087
GAR	CvsW	0.07 (0.04)	0.06 (0.03)	0.24	0.026	–	–	–	–
DR	CvsW	0.90 (0.04)	0.91 (0.03)	-0.20	0.033	0.94 (0.02)	0.94 (0.02)	-0.18	0.087
OtherTop1	CvsW	0.57 (0.03)	0.58 (0.02)	-0.21	0.033	0.58 (0.02)	0.58 (0.02)	0.01	0.900

Table 7: Token-level attention diagnostics (Correct vs. Wrong). Each metric measures where the model attends during generation: IAR, GAR, DR, and OtherTop1. We report median (IQR) for each group, Cliff's δ, and q-values from two-sided Mann-Whitney U tests with BH-FDR correction within each condition.

Metric	Analysis	*With-gloss*				*No-gloss*			
		Type 2	Type 3	Type 4	q	Type 2	Type 3	Type 4	q
IAR	WrongType	0.03 (0.03)	0.03 (0.02)	0.03 (0.01)	0.878	0.05 (0.03)	0.06 (0.02)	0.05 (0.02)	< 0.001
GAR	WrongType	0.05 (0.01)	0.07 (0.03)	0.06 (0.04)	0.269	–	–	–	–
DR	WrongType	0.91 (0.03)	0.90 (0.03)	0.91 (0.03)	0.269	0.95 (0.03)	0.94 (0.02)	0.95 (0.02)	< 0.001
OtherTop1	WrongType	0.58 (0.02)	0.58 (0.02)	0.58 (0.02)	0.269	0.58 (0.02)	0.57 (0.03)	0.58 (0.02)	< 0.001

Table 8: Token-level attention diagnostics across error types (Type 2/3/4). We report median (IQR) per error type and q-values from Kruskal-Wallis tests with BH-FDR correction within each condition.

$$\mathrm{OtherTop1} = \max_{k \in K_{\mathrm{other}}} \frac{\bar{A}(k)}{\sum_{j \in K_{\mathrm{other}}} \bar{A}(j)}$$

$\mathrm{OtherTop1}$ is the maximum share among off-span tokens after renormalizing within the off-span set; higher values indicate a stronger off-span peak.

For Correct-vs.-Wrong comparisons, we use two-sided Mann-Whitney U tests, a non-parametric two-group comparison, and report Cliff's δ, whose sign indicates the direction of the difference and whose magnitude reflects its strength. For comparisons across Types 2/3/4 within wrong outputs, we use Kruskal-Wallis tests, which assess whether the error types differ overall without assuming normality. Within each condition, p-values are adjusted across metric-wise tests using BH-FDR. Table 7 and 8 summarize the result of token-level diagnostics.

With-gloss (Type 0 excluded; $n = 189$). Correct outputs exhibit stronger gloss anchoring and reduced off-span allocation: GAR increases, while both DR and OtherTop1 decrease (all $q < 0.05$). By contrast, idiom-span mass does not distinguish Correct from Wrong (IAR; $q = 0.250$). Within wrong outputs, cross-type differences are not robust after correction.

No-gloss (Type 0 excluded; $n = 189$; GAR not applicable). The Correct-vs.-Wrong contrast is weaker and does not survive correction for IAR or DR, and OtherTop1 shows no meaningful difference. Nevertheless, WrongType comparisons are strongly structured by error type across the available diagnostics (IAR/DR/OtherTop1; all $q < 0.001$), suggesting more heterogeneous failure modes in the absence of an explicit gloss anchor.

Overall, glosses consistently improve open-ended idiom generation, and attention diagnostics suggest that correctness under *With-gloss* aligns with stronger gloss anchoring, whereas *No-gloss* errors exhibit more heterogeneous attention patterns across error types.

6 Conclusion

We present **G-IdiomAlign**, a gloss-pivoted benchmark supporting Multiple-Choice Idiom Equivalence and Gloss-Contrastive Generation to diagnose literal-translation biases across LLMs. Our results, spanning diverse proprietary and open-source models, highlight literal-translation attraction as a persistent obstacle in cross-lingual idiom alignment. Attention-based diagnostics further suggest that successful *With-gloss* generations are associated with stronger anchoring to gloss information. While providing explicit glosses consistently improves open-ended generation under an embedding-based semantic proxy, performance remains far from saturated. This motivates further developments in robust idiom translation.

Limitations

G-IdiomAlign is designed as a precision-first benchmark for diagnosing cross-lingual idiom alignment rather than exhaustive idiomatic equivalence. This improves interpretability and reproducibility, but limits coverage and external validity.

English-pivot bias. We use English Wiktionary glosses as a single semantic pivot across nine languages. This improves consistency, but may introduce English-centric bias because glosses can compress pragmatic or culture-specific meaning and vary in style and granularity across editions.

Trade-offs in reference alignments. To ensure unambiguous supervision, we apply single-sense filtering and exclude polysemous idioms. This improves interpretability but removes sense selection and reduces coverage. We further impose a one-to-one constraint via MNN, which favors high-precision pairs but under-represents many-to-many relations such as synonym clusters.

Dependence on embeddings, proxies, and tools. The pipeline relies on embedding-based retrieval and filtering, as well as an LLM for distractor generation. As a result, benchmark construction is sensitive to embedding calibration, dataset size, and gloss noise, and may inherit model-specific biases. Task 2 further uses fixed-threshold embedding similarity as a scalable proxy, which may miss valid non-canonical generations and may not be fully comparable across languages, such dependence remains a limitation.

Limited generality. Our attention analyses are correlational rather than causal, and are based on a single model with a modest annotated sample under greedy decoding. The observed patterns may therefore not generalize across models, decoding settings, or language directions.

Ethical Considerations

This work involves several value-sensitive design choices. First, we use English glosses as a shared semantic pivot to enable controlled cross-lingual idiom alignment. While this results in high-confidence alignment, it may introduce English-centric bias and compress culture-specific pragmatic or stylistic distinctions encoded in non-English idioms. We treat this as a deliberate trade-off for diagnostic clarity, rather than as a claim of cultural neutrality.

Second, idioms are culturally grounded expressions, and operationalizing idiomatic equivalence through glosses and embedding-based similarity necessarily abstracts away contextual and socio-cultural nuance. Our benchmark is therefore intended to support analysis of model behavior under controlled conditions, not to define authoritative judgments of idiomatic correctness across cultures.

Finally, our evaluation metrics, especially the embedding-based proxy in Gloss-Contrastive Generation, are designed for consistent comparison rather than deployment. We caution against using benchmark scores as standalone indicators of translation quality or fairness in real-world applications.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

References

Anthropic. 2025. Claude Haiku 4.5.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *Preprint*, arXiv:2512.02556.

Qiuyu Ding, Hailong Cao, and Tiejun Zhao. 2024. Enhancing bilingual lexicon induction via bi-directional translation pair retrieving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17898–17906.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51:505–555.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18554–18563.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Mistral AI team. 2024. Un ministral, des ministraux.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

OpenAI. 2024. New embedding models and API updates.

OpenAI. 2026. GPT-5.1: A smarter, more conversational ChatGPT.

Ming Qian. 2024. Automating idiom translation with cross-lingual natural language generation grounded in semantic analyses using large language models. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 95–115, Chicago, USA. Association for Machine Translation in the Americas.

Prateek Saxena and Soma Paul. 2020. EPIE Dataset: A Corpus For Possible Idiomatic Expressions. *Preprint*, arXiv:2006.09479.

Uliana Sentsova, Debora Ciminari, Josef Van Genabith, and Cristina España-Bonet. 2025. MultiCoPIE: A multilingual corpus of potentially idiomatic expressions for cross-lingual PIE disambiguation. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 67–81, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Yanming Sun, Runzhe Zhan, Chi Seng Cheang, Han Wu, Xuebo Liu, Yuyao Niu, Fengying Ye, Kaixin Lan, Lidia S Chao, and Derek F Wong. 2026. Exposing the cracks: Vulnerabilities of retrieval-augmented llm-based machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33135–33143.

Qwen Team. 2025. Qwen3-Max: Just Scale it.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Junyang Lin, Jingren Zhou, and et al. 2025b. Qwen3-embedding-8B.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025c. Evaluating LLMs on chinese idiom translation. *ArXiv*, abs/2508.10421.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023. CLCL: Non-compositional expression detection with contrastive learning and curriculum learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–743, Toronto, Canada. Association for Computational Linguistics.

A Additional Language Pairs

In addition to the core G-IdiomAlign benchmark reported in Table 2, we construct a supplementary set of additional language pairs to broaden cross-lingual coverage in Table 9. This extension introduces four new languages: Arabic (Ar), Korean (Ko), Thai (Th), and Vietnamese (Vi), and matches them with both the original benchmark languages and one another using the same gloss-pivoted pipeline.

Overall, this supplementary set contains 42 language pairs and 1,014 bidirectional rank-1 aligned idiom pairs. As in the core benchmark, these pairs are obtained through gloss-based candidate retrieval, bidirectional agreement, and distribution-aware filtering. Because coverage remains limited after precision-oriented filtering, we do not include these pairs in the main evaluation; instead, we release them to support future research on broader cross-lingual idiom alignment.

B Wiktionary Idiom Harvesting and Gloss Cleaning

This appendix describes our cross-lingual harvesting framework for extracting idiom entries and their definition-based glosses from Wiktionary.

License. The data used in this section are derived from Wiktionary, a collaboratively constructed resource available under the Creative Commons Attribution-ShareAlike 3.0 License (CC BY-SA 3.0).

Harvesting. For each language, we enumerate idiom entry pages by traversing the corresponding idiom category on English Wiktionary and collecting linked entry pages across all "next page" partitions. We optionally apply conservative filters to remove obvious auxiliary pages introduced by category-page organization.

Gloss extraction. From each entry page, we extract an example-free definition string as the gloss representation. The extraction prioritizes sense-definition text and excludes usage examples and other non-definitional material.

Precision-oriented screening. We enforce a strict single-sense criterion: an entry is retained only if extraction yields exactly one candidate gloss string. We then apply lightweight normalization to remove leading labels and standardize whitespace.

C Distribution-Aware Filtering: Implementation Details

This appendix provides an implementation-level specification of the distribution-aware filtering step described in Section 3.1. The filtering procedure operates on the set of MNN-confirmed idiom pairs for a given language pair, each associated with its rank-1 gloss similarity score.

Input. For each *unordered* language pair, after enforcing mutual nearest neighbors (MNN), we obtain a set of candidate alignments $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{n}$, where each retained pair is associated with a rank-1 similarity score $s_i \in \mathbf{R}$, computed as cosine similarity between the corresponding English-gloss embeddings (as defined in Section 3.1). The MNN criterion itself is always enforced bidirectionally.

Equal-width binning. Let $s_{\min} = \min_i s_i$ and $s_{\max} = \max_i s_i$. We partition the interval $[s_{\min}, s_{\max}]$ into 10 equal-width bins using 11 bin edges:

$$e_j = s_{\min} + j \cdot \frac{s_{\max} - s_{\min}}{10}, \quad j = 0, 1, \ldots, 10.$$

Pair	N	Pair	N	Pair	N	Pair	N
Ar–De	5	Ar–Zh	2	Ja–Ko	9	Pt–Vi	42
Ar–En	3	De–Ko	6	Ja–Th	68	Th–Vi	58
Ar–Es	3	De–Th	52	Ja–Vi	52	Zh–Ko	2
Ar–Fi	4	De–Vi	55	Ko–Es	2	Zh–Th	28
Ar–Fr	13	En–Ko	1	Ko–Pl	3	Zh–Vi	18
Ar–Ja	9	En–Th	28	Ko–Pt	9		
Ar–Ko	13	En–Vi	25	Ko–Th	8		
Ar–Pl	2	Es–Th	54	Ko–Vi	8		
Ar–Pt	8	Es–Vi	38	Pl–Th	42		
Ar–Th	5	Fi–Ko	5	Pl–Vi	39		
Ar–Vi	11	Fi–Th	62	Pt–Th	60		

Table 9: Additional language-pair composition in G-IdiomAlign. **N** denotes the count of aligned idiom pairs. Each pair is reported once using a canonical ordering.

Modal-bin cutoff. Let c_j denote the number of MNN-confirmed pairs whose rank-1 scores fall into bin j. We define the modal bin as

$$b = \arg \max_{j \in \{1,\ldots,10\}} c_j.$$

When multiple bins tie for the maximum count, ties are resolved by selecting the first maximizer returned by the implementation.

Filtering rule. An MNN-confirmed pair (x_i, y_i) is retained if and only if its rank-1 similarity score falls in the modal bin or any higher bin:

$$(x_i, y_i) \text{ is kept } \iff \text{bin}(s_i) \geq b.$$

Equivalently, the lower edge of the modal bin acts as a language-pair-specific cutoff. As in the main text, similarity scores are treated as a relative diagnostic signal within each language pair, rather than as an absolute criterion of semantic correctness.

D Alignment Quality Evaluation Details

This appendix provides additional details on the annotation protocol, LLM prompting strategy, and statistical estimation used in Section 3.3.

Annotation protocol. All idiom pairs are evaluated using a 3-point semantic equivalence scale: 2 (fully equivalent), 1 (partially equivalent), and 0 (non-equivalent). A score of 2 indicates that two idioms express the same core meaning and can be reasonably substituted in similar contexts; a score of 1 indicates partial equivalence with differences in tone, intensity, or pragmatic usage; and a score of 0 indicates non-equivalence. Annotators are instructed to focus on semantic meaning rather than literal form. Both human annotators and LLM judges follow the same annotation instructions and scoring criteria.

LLM prompting strategy. We implement LLM-based evaluation by embedding the annotation rubric directly into structured prompts. Each prompt takes as input a pair of idioms and their English glosses and requires the model to output an equivalence label (0/1/2).

To improve robustness and reduce prompt sensitivity, we adopt a multi-view prompting strategy with complementary perspectives, including (i) direct semantic comparison, (ii) a substitutability-based test that evaluates whether the two idioms can be used interchangeably in similar contexts, and (iii) comparison of pragmatic function and strength. All prompts share the same core rules: prioritize English glosses over surface forms and avoid relying on literal similarity.

Prompts are shown below:

```
Task:
Given two expressions in different
languages and their English glosses,
assign one equivalence label:
- 2 = Correct Equivalent (same core
meaning; substitutable in similar
contexts)
- 1 = Partially Correct (overlapping
meaning but differences in tone or
usage)
- 0 = Incorrect (different core meaning
or function).

- Do not rely on literal similarity.

Input:
- Idiom A: <IDIOM_A>
- GLOSS A: <GLOSS_A_EN>
- Idiom B: <IDIOM_B>
- GLOSS B: <GLOSS_B_EN>


Output:
- Equivalence: <0/1/2>
```

```
Task:
Judge equivalence using a substitutable
test.
Step 1: Based on the English glosses,
imagine 2 short English contexts where
```

```
this meaning is used.
Step 2: Decide if A and B could reasonably substitute each other in those contexts.
- 2 = Correct Equivalent (same core meaning; substitutable in similar contexts)
- 1 = Partially Correct (overlapping meaning but differences in tone or usage)
- 0 = Incorrect (different core meaning or function).

- Do not rely on literal similarity.

Input:
- Idiom A: <IDIOM_A>
- GLOSS A: <GLOSS_A_EN>
- Idiom B: <IDIOM_B>
- GLOSS B: <GLOSS_B_EN>

Output:
- Equivalence: <0/1/2>

Task:
Judge equivalence by explicitly comparing pragmatic function and strength.
- 2 = Correct Equivalent (same core meaning; substitutable in similar contexts)
- 1 = Partially Correct (overlapping meaning but differences in tone or usage)
- 0 = Incorrect (different core meaning or function).

- Do not rely on literal similarity.

Input:
- Idiom A: <IDIOM_A>
- GLOSS A: <GLOSS_A_EN>
- Idiom B: <IDIOM_B>
- GLOSS B: <GLOSS_B_EN>

Output:
- Equivalence: <0/1/2>
```

For each idiom pair, we obtain independent judgments from three LLMs (GPT-5.1, Gemini-2.5-pro, and Claude-4.5-Haiku). The final label is determined via majority voting across models. When all three models disagree, we assign a score of 1 to avoid over-claiming full equivalence while retaining borderline cases instead of discarding them. This combination of multi-prompt design and multi-model aggregation improves the robustness and stability of LLM-based judgments.

Sampling and statistical estimation. For Zh–En, we randomly sample 200 idiom pairs and evaluate them using human annotators. For all other language pairs, we sample 50 idiom pairs per language pair and evaluate them using the LLM-based protocol described above.

For non-Zh–En language pairs, each language pair is treated as one observation. We compute the mean strict and lenient accuracy across language pairs and report 95% confidence intervals computed as t-intervals over language pairs.

For Zh–En, statistics are computed over individual samples ($n = 200$). The reported accuracies correspond to the proportion of samples satisfying each criterion, where strict accuracy counts only score-2 pairs and lenient accuracy counts score-1 and score-2 pairs.

Results summary. For non-Zh–En language pairs, the mean strict accuracy is 0.685 (95% CI [0.645, 0.724]) and the mean lenient accuracy is 0.923 (95% CI [0.907, 0.940]). These correspond to 68.5% fully equivalent pairs (score = 2) and 92.3% partially or fully equivalent pairs (score ≥ 1).

For Zh–En human evaluation, the strict accuracy is 0.655 (95% CI [0.589, 0.721]) and the lenient accuracy is 0.895 (95% CI [0.852, 0.938]). Despite being slightly more conservative, human evaluation yields results consistent with LLM-based estimates.

Overall, both LLM-based and human evaluations provide converging evidence that G-IdiomAlign achieves high semantic alignment quality, supporting the effectiveness of the proposed mining and filtering pipeline.

E Model Dependence of Similarity Scores

We assess the sensitivity of gloss-based similarity scores to encoders by recomputing scores for the same aligned idiom pairs using Qwen3-Embedding-8B, an alternative multilingual embedding model. We examine three aspects: global score calibration, threshold-based coverage, and cross-encoder consistency in relative score structure.

E.1 Global Statistics and Calibration Shift

For the same $N = 18{,}785$ aligned pairs, we recompute cosine similarities between source and target glosses using Qwen3-Embedding-8B and compare them with the construction-time scores obtained using OpenAI text-embedding-3-large. Both encoders use cosine similarity over L_2-normalized embeddings.

Table 10 summarizes the original scores s^{orig}, the recomputed scores s^{Qwen3}, and the per-instance difference $\Delta_i = s_i^{\text{Qwen3}} - s_i^{\text{orig}}$. Qwen3 produces systematically higher absolute similarity values than the original encoder (mean $0.67 \rightarrow 0.78$;

	mean	std	min	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}
Original s^{orig}	0.67	0.13	0.37	0.52	0.57	0.65	0.75	0.86
Qwen3 s^{Qwen3}	0.78	0.10	0.35	0.66	0.72	0.78	0.85	0.91
Shift Δ	0.11	0.08	-0.45	0.01	0.06	0.11	0.17	0.21

Table 10: Global similarity statistics for the OpenAI text-embedding-3-large and Qwen3-Embedding-8B, and the per-instance difference $\Delta_i = s_i^{\text{Qwen3}} - s_i^{\text{orig}}$ over $N = 18{,}785$ aligned pairs.

Threshold t	0.518	0.567	0.617	0.666	0.716	0.766	0.815	0.865	0.914
Original count	16898	14381	11378	8499	6027	4125	2803	1686	945
Qwen3 count	18721	18545	17942	16606	14108	10583	6934	3933	1895

Table 11: Coverage under fixed similarity thresholds: number of aligned pairs with $sim \geq t$ under each encoder.

median $0.65 \rightarrow 0.78$), with an average shift of $\Delta\mu = 0.11$. However, the shift is heterogeneous across instances ($p_{10} = 0.01$, $p_{90} = 0.21$) and includes negative values, indicating that the difference is not reducible to a simple global offset or rescaling.

E.2 Threshold Sensitivity Under Calibration Shift

To illustrate the practical effect of calibration differences, we count the number of aligned pairs satisfying $s \geq t$ under fixed absolute thresholds t. We sweep nine thresholds uniformly over the intersection of the two encoders' 10th–90th percentile score intervals.

Table 11 shows that the number of retained pairs differs substantially across encoders at the same threshold. This shows that absolute similarity thresholds are not directly comparable across embedding models in this setting, supporting our treatment of similarity as a relative diagnostic signal in the main text.

E.3 Embedding Consistency

Despite the shift in absolute similarity values, the relative score structure is largely preserved across encoders. Across all $N = 18{,}785$ aligned pairs, the two score sets are strongly correlated (Pearson $r = 0.807$) and show substantial rank agreement (Spearman $\rho = 0.784$). Thus, encoder choice has a larger effect on absolute calibration and threshold-based coverage than on the comparative ordering of aligned pairs.

F Task 1 Distractor Generation

This appendix reports the unified prompt template used to generate the three typed distractors for Task 1 (Multiple-Choice Idiom Equivalence). For each instance, the canonical target idiom from G-IdiomAlign is used as the reference option, while Qwen-Max is used *only* to generate the remaining three distractors under a single prompt: Literal Translation Trap (LT), Lexical Cue Trap (LC), and Contextual Association Trap (CA).

F.1 Prompt

Model and decoding. We generate distractors using `Qwen-Max` with greedy decoding (temperature $= 0$, top-$p = 1$).

Prompt template. The following prompt is used verbatim in our implementation, with placeholders instantiated per instance.

```
You are an expert linguist specializing
in cross-cultural idiom translation and
test design. Your task is to create a
multiple-choice question dataset to test
whether an AI model truly understands
idioms or just relies on literal
translation.

Input Data:
- Source Idiom: <SOURCE_IDIOM>
- Source Meaning: <SOURCE_MEANING>
- Source Language: <SOURCE_LANGUAGE>
- Target Language: <TARGET_LANGUAGE>

Task:
Generate 3 options for a multiple-choice
question.
1. Option (Literal Translation Trap): A
direct, word-for-word translation of the
source in <TARGET_LANGUAGE>.
2. Option (Lexical Cue Trap): A real
idiom in <TARGET_LANGUAGE> that shares
only part of a salient keyword from the
literal translation but has a completely
DIFFERENT meaning.
3. Option (Contextual Association Trap):
A real idiom in <TARGET_LANGUAGE> that
has a related context but opposite
meaning.

Constraints:
- The 'Lexical Cue Trap' should clearly
```

```
reflect a salient lexical cue from the
literal translation.
Output strictly in JSON format.
```

Distractor intent. The prompt defines three distractor types for diagnosis by error category: (i) **LT (Literal Translation Trap)** is a word-by-word rendering of the source idiom into the target language, designed to be surface-faithful but not meaning-equivalent; (ii) **LC (Lexical Cue Trap)** is a target-language idiom that overlaps with a salient lexical cue from the literal translation while conveying a different meaning; (iii) **CA (Contextual Association Trap)** is a target-language idiom that is contextually related yet semantically opposite to the intended meaning.

Instance assembly and shuffling. For each benchmark alignment pair (x, y) in G-IdiomAlign, we take the canonical reference target idiom y as the correct option and populate the remaining three options with the generated LT/LC/CA distractors. We then shuffle the option order per instance to reduce positional bias.

F.2 Validity of LLM-Generated Distractors

We conduct a control in which the model is shown only the answer options, without the question stem. To avoid a trivial signal, we exclude the literal-translation hard negatives in this control, since they are intentionally designed to be non-idiomatic. The goal is to test whether the model can systematically prefer the gold idiomatic option based on surface properties alone.

Across three runs, the average selection rate for the gold option type is 0.3567, compared with a random baseline of approximately 0.3333. A chi-square test does not show a significant deviation from a uniform distribution ($\chi^2 = 2.94$, $p = 0.23$). This suggests that there is no strong evidence that the model can reliably identify the gold option from stylistic cues alone. To further reduce superficial shortcuts, we also shuffle option order and normalize option formatting.

Manual verification of distractor-type validity. We additionally manually inspect 200 questions to verify whether the generated distractors match their intended categories (e.g., LT, LC, and CA). We adopt a strict per-question criterion: a question is counted as valid only if all four options match their intended types. Under this criterion, 163 out of 200 questions are valid, corresponding to an accuracy of 81.5%.

These results indicate that the generated distractors largely satisfy the intended hard-negative constraints and are not trivially distinguishable by superficial signals alone.

G Task 2 Generation Prompts and Output Constraints

This appendix reports the prompts used for Task 2 (Gloss-Contrastive Generation) under the two input conditions: *With-gloss* (source idiom plus its English gloss) and *No-gloss* (source idiom only).

Prompt template (With-gloss). In the *With-gloss* condition, we provide the source idiom together with its English gloss as an explicit semantic pivot, and ask the model to generate a meaning-equivalent idiom in the target language:

```
Output the <TARGET_LANGUAGE> idiom
corresponding to "<SOURCE_IDIOM>" with
meaning "<GLOSS>". Return only one idiom
and do not include any explanation or
additional text.
```

Prompt template (No-gloss). In the *No-gloss* condition, we provide only the source idiom and ask the model to generate the corresponding idiom in the target language:

```
Output the <TARGET_LANGUAGE> idiom
corresponding to "<SOURCE_IDIOM>".
Return only one idiom and do not include
any explanation or additional text.
```

H Task 1 Outcomes

H.1 Breakdown by Distractor Type

This appendix reports the numeric outcome proportions corresponding to Figure 3. For each model and target-language regime (Zh-target, En-target, and Other targets), we decompose outcomes into the *Correct* selection and three typed distractor selections: *Literal Translation Trap* (LT), *Lexical Cue Trap* (LC), and *Contextual Association Trap* (CA). All values in Table 12 are regime-level **micro** proportions (instance-weighted within each regime) computed over the Task 1 evaluation instances for that regime; within each model–regime block, the four percentages sum to 100% (up to rounding). These numeric breakdowns support the error-type comparisons discussed in Section 5.1.2.

H.2 Case Study: Task 1 (Multiple-choice)

Table 13 presents representative examples from Task 1 to illustrate how the multiple-choice design

Model	Zh-target				En-target				Other-targets			
	Corr	LT	LC	CA	Corr	LT	LC	CA	Corr	LT	LC	CA
DeepSeek-V3.2 (NT)	56.96	23.06	14.31	5.67	56.16	27.83	10.83	5.19	47.53	40.33	8.72	3.42
DeepSeek-V3.2 (T)	69.18	17.96	10.37	2.49	65.38	24.49	7.13	3.00	54.09	37.77	6.53	1.60
Gemini-2.5-Pro	67.57	17.74	10.17	4.52	64.95	24.81	6.84	3.40	62.91	28.54	6.02	2.52
Claude-4.5-Haiku	57.30	24.06	14.40	4.23	53.31	27.78	13.92	4.99	43.44	46.38	7.77	2.41
Ministral-8B-Instruct	28.51	26.18	26.78	18.53	28.97	30.03	24.97	16.03	27.76	36.55	20.64	15.06
Qwen3-8B (NT)	33.72	33.50	24.72	8.06	33.62	31.01	24.84	10.52	23.99	55.99	14.68	5.34
Qwen3-8B (T)	41.89	38.81	16.01	3.29	39.10	41.53	14.78	4.59	31.39	52.52	12.80	3.29

Table 12: Task 1 outcome proportions by target-language regime. Each row reports the **micro** fraction (%) of instances that are *Correct* or correspond to choosing one of the three typed distractors: Literal Translation Trap (LT), Lexical Cue Trap (LC), and Contextual Association Trap (CA). Values sum to 100% within each model–regime block (up to rounding).

Source Idiom	Options	Gold Target Idiom	Model Prediction
一丈差九尺	(A) wide of the mark (correct) (B) one zhang short by nine chi (literal trap) (C) measure twice, cut once (lexical cue trap) (D) hit the nail on the head (contextual association trap)	(A) wide of the mark	(A) wide of the mark ✓
殺人不眨眼	(A) fish-blooded (correct) (B) kill without blinking an eye (literal trap) (C) bat an eyelash (lexical cue trap) (D) have a heart of gold (contextual association trap)	(A) fish-blooded	(B) kill without blinking an eye (literal trap) ✗
仆心仆肺	(A) bend over backwards (correct) (B) servant heart, servant lungs (literal trap) (C) heart and soul (lexical cue trap) (D) pull your punches (contextual association trap)	(A) bend over backwards	(C) heart and soul (lexical cue trap) ✗
善罷甘休	(A) fold like a cheap suit (correct) (B) willingly stop and sweetly rest (literal trap) (C) sweet tooth (lexical cue trap) (D) hold a grudge (contextual association trap)	(A) fold like a cheap suit	(D) hold a grudge (related but not equivalent) ✗

Table 13: Representative examples for Task 1 (multiple-choice). Each instance contains one correct target idiom and three typed distractors. ✓ indicates a correct prediction and ✗ an incorrect one.

probes different types of distractors. Each instance includes one correct target idiom and three typed distractors: a Literal Translation Trap, a Lexical Cue Trap, and a Contextual Association Trap.

The examples show that models can succeed when they recover the intended figurative meaning (e.g., 一丈差九尺 $\rightarrow$ wide of the mark), but models may select (i) a literal translation that closely mirrors the source form (e.g., 殺人不眨眼), (ii) an option triggered by salient lexical cues (e.g., 仆心仆肺), or (iii) an expression that is semantically related but not equivalent (e.g., 善罷甘休).

These cases highlight that Task 1 not only evaluates overall accuracy, but also reveals distinct and interpretable failure modes through the use of typed distractors.

I Task 2 Outcomes

I.1 Threshold Sensitivity

We report Task 2 results under additional similarity thresholds $t \in \{0.65, 0.70, 0.75, 0.80\}$ in Table 14. MeanSim is reported as $100 \times sim$ and macro-averaged over directions. Acc@t is macro-averaged accuracy (%) at threshold t. For the *With-gloss* panel, we additionally report the per-threshold improvement over *No-gloss* in parentheses (percentage points).

I.2 Case Study: Task 2 (Open-ended Generation)

Table 15 presents representative examples from Task 2 to illustrate model behavior in open-ended idiom generation. Unlike Task 1, this setting does not constrain models to a fixed set of options, and

Model	MeanSim	Acc@0.65	Acc@0.70	Acc@0.75	Acc@0.80
No-gloss					
DeepSeek-V3.2 (NT)	68.30	58.02	41.99	28.63	19.90
DeepSeek-V3.2 (T)	72.31	67.64	49.01	34.10	24.66
Gemini-2.5-Pro	68.71	61.33	44.06	29.98	19.70
Claude-4.5-Haiku	70.90	65.47	45.47	30.25	20.47
Ministral-8B-Instruct	68.30	59.86	37.97	21.96	12.96
Qwen3-8B (NT)	67.54	57.22	36.29	20.78	11.59
Qwen3-8B (T)	68.53	60.26	39.03	23.42	14.04
With-gloss (absolute; Δ vs. No-gloss)					
DeepSeek-V3.2 (NT)	70.94	65.57 (+7.55)	48.35 (+6.36)	33.68 (+5.05)	23.57 (+3.67)
DeepSeek-V3.2 (T)	74.24	72.50 (+4.86)	54.90 (+5.89)	39.71 (+5.61)	29.63 (+4.97)
Gemini-2.5-Pro	72.54	68.22 (+6.89)	51.09 (+7.03)	36.86 (+6.88)	27.11 (+7.41)
Claude-4.5-Haiku	72.94	70.58 (+5.11)	52.16 (+6.69)	36.41 (+6.16)	25.66 (+5.19)
Ministral-8B-Instruct	71.69	69.78 (+9.92)	50.00 (+12.03)	32.61 (+10.65)	21.22 (+8.26)
Qwen3-8B (NT)	70.69	67.41 (+10.19)	47.09 (+10.80)	30.53 (+9.75)	18.80 (+7.21)
Qwen3-8B (T)	71.40	69.20 (+8.94)	49.16 (+10.13)	32.36 (+8.94)	20.86 (+6.82)

Table 14: Task 2 threshold sensitivity in a single table with two panels. MeanSim is $100 \times sim$ macro-averaged across directions. Acc@t is macro-averaged accuracy (%) at threshold t. In the *With-gloss* panel, each Acc@t cell reports the absolute accuracy followed by the per-threshold gain over *No-gloss* in parentheses (percentage points).

Source Idiom	Gold Target	Model	Output	Sim.
修身養性	build character	DeepSeek-V3.2 (T)	turn over a new leaf ✗	0.6750
		Claude-4.5-Haiku	cultivate one's moral character ✓	0.5561
		Qwen3-8B (T)	practice restraint ✗	0.4700
倒錢落海	fool away	DeepSeek-V3.2 (T)	throw money down the drain ✓	0.7733
		Claude-4.5-Haiku	pour money down the drain ✓ (paraphrase)	0.7617
		Qwen3-8B (T)	pour money down the drain ✓ (paraphrase)	0.7617
de plantilla	一技之長	DeepSeek-V3.2 (T)	千篇一律 ✗	0.5789
		Claude-4.5-Haiku	專業人士 ✗	0.5027
		Qwen3-8B (T)	行家 ✗	0.6918
ir al grano	cut to the chase	DeepSeek-V3.2 (T)	get to the point ✓ (paraphrase)	0.8811
		Claude-4.5-Haiku	get down to brass tacks ✓ (paraphrase)	0.8509
		Qwen3-8B (T)	cut to the chase ✓	1.0000

Table 15: Representative examples for Task 2 (open-ended generation) with embedding-based similarity scores. ✓ indicates correct or acceptable paraphrase; ✗ indicates incorrect predictions.

predictions are evaluated based on semantic equivalence or acceptable paraphrases of the target idiom.

The examples show that models can produce correct idiomatic expressions (e.g., ir al grano → cut to the chase) or acceptable paraphrases (e.g., 倒錢落海 → pour money down the drain), but errors often reflect difficulties in capturing the precise figurative meaning. In particular, models may generate expressions that are overly general (e.g., 修身養性 → practice restraint), semantically shifted (e.g., turn over a new leaf), or unrelated to the intended meaning (e.g., de plantilla).

J Task 2 Supplementary Evaluation

Human calibration. To provide a point of calibration for the embedding-based metric, we conduct a small human evaluation on Task 2 outputs. We manually annotate $N = 86$ examples for idiom-to-idiom correctness, obtaining an overall correct ratio of 0.407. Using these annotations as ground truth, embedding similarity yields an AUC of 0.77, with a 95% confidence interval of $[0.67, 0.86]$, computed by nonparametric bootstrap resampling over the 86 examples.

At a representative threshold $t = 0.75$, the proxy achieves precision 0.875 ($TP = 7$, $FP = 1$) with coverage 9.3%. These values indicate that the metric is most reliable for identifying a relatively high-confidence subset of semantically correct outputs, rather than for exhaustively capturing all acceptable generations.

Surface-form metrics. As a complement to semantic matching, we also report auxiliary surface-form metrics against the canonical gold reference in Table 16. Specifically, we compute Exact Match (EM), and additionally report BLEU and ChrF as reference-based overlap measures. These metrics

Model	EM			BLEU			ChrF		
	with	w/o	Δ	with	w/o	Δ	with	w/o	Δ
DeepSeek-V3.2 (NT)	9.36	8.62	0.74	16.82	15.45	1.37	23.90	22.60	1.30
DeepSeek-V3.2 (T)	11.61	9.24	2.37	21.96	18.54	3.43	27.54	24.43	3.11
Gemini-2.5-Pro	10.85	9.49	1.36	19.77	17.59	2.18	26.81	24.39	2.42
Claude-4.5-Haiku	9.06	6.74	2.31	19.17	15.75	3.42	24.62	21.60	3.02
Ministral-8B-Instruct	6.15	4.62	1.52	15.31	12.14	3.16	20.47	17.46	3.01
Qwen3-8B (NT)	4.08	4.05	0.03	13.22	11.65	1.57	18.79	17.46	1.34
Qwen3-8B (T)	4.76	2.90	1.86	14.24	10.92	3.32	19.68	16.78	2.90

Table 16: Auxiliary surface-form metrics for Task 2.

provide a view of whether a model output matches an attested or canonical idiom form.

As expected in open-ended generation, surface-form metrics should be interpreted with caution: semantically valid idiomatic paraphrases may still receive low scores if they differ from the canonical reference string. For this reason, we treat EM, BLEU, and ChrF as supplementary indicators. Although absolute EM values are low, the directional trends remain informative.

K Human Annotation Scheme for Gloss-Contrastive Generation

We categorize each idiom chosen from Gloss-Contrastive Generation into one of five mutually exclusive labels $\{0, 1, 2, 3, 4\}$. Labels are assigned following the decision procedure below.

Label definitions.

- **Label 0 (Invalid / generation failure).** The output is empty, gibberish, or otherwise not interpretable as meaningful text in the target language.
- **Label 1 (Correct idiomatic translation).** The output is a fluent and semantically correct translation that realizes an appropriate idiomatic expression in the target language, conveying the intended meaning of the source idiom.
- **Label 2 (Literal word-by-word translation).** The output is semantically linked to the idiom's surface form and appears to translate the idiom compositionally (word-by-word or phrase-by-phrase), rather than conveying the intended idiomatic meaning.
- **Label 3 (Meaning paraphrase, non-idiomatic form).** The output correctly conveys the intended meaning of the source idiom, but does so using a non-idiomatic paraphrase (i.e., not an idiom or conventional idiomatic expression in the target language).
- **Label 4 (Incorrect meaning).** The output is meaningful text in the target language but fails to convey the intended meaning of the source idiom, including cases of mistranslation, wrong sense, or unrelated content.

Decision procedure. Labels are assigned in the following order to ensure mutual exclusivity:

1. If the output is not interpretable as meaningful text in the target language, assign **Label 0**.
2. Otherwise, if the output is a fluent and correct idiomatic translation that appropriately realizes the source idiom in the target language, assign **Label 1**.
3. Otherwise, if the output translates the idiom literally based on its surface form without conveying the intended idiomatic meaning, assign **Label 2**.
4. Otherwise, if the output correctly conveys the intended meaning but does not use an idiomatic expression in the target language, assign **Label 3**.
5. All remaining meaningful but semantically incorrect outputs are assigned **Label 4**.

L Attention Diagnostics: Salient Heads and Layer Aggregation

This appendix defines the salient head and layer sets used in Table 6. All computations use post-softmax attention weights from Qwen3-8B using `TransformerLens` under greedy decoding. We consider three subsets u: *All*, *Correct-only*, and *Wrong-only*, under two conditions c: *With-gloss* and *No-gloss*.

Per-sample head scores. Let $A^{(\ell,h)}_{i,t\to k}$ denote the attention weight from generation position t to key position k at layer ℓ and head h for sample i. Let $\mathcal{T}_i$ be the set of template-defined generation positions and K_i the tracked token span (e.g., idiom span or gloss span). The head score is

$$S_i^{(\ell,h)} = \frac{1}{|\mathcal{T}_i|} \sum_{t\in\mathcal{T}_i} \sum_{k\in K_i} A^{(\ell,h)}_{i,t\to k},$$

or the corresponding ratio-based variant for contrastive analyses.

Salient heads. For each sample i, we retain the top-k heads ranked by $S_i^{(\ell,h)}$ (k=10). We then count how often each head appears across samples and define the salient head set $\mathcal{H}^{(c,u)}_{\text{sal}}$ as the m=50 most frequent heads.

Salient layers. We aggregate layers from the per-sample top-k heads. For each layer ℓ, we sum the number of per-sample top-k entries belonging to ℓ across all samples, and retain the top-n layers by this count as the salient layer set $\mathcal{L}^{(c,u)}_{\text{sal}}$ (n=10).

Cross-condition overlap. For each subset u, we report the Jaccard similarity between *With-gloss* and *No-gloss* salient sets, for both heads and layers. Higher values indicate greater structural overlap across conditions. These are the values reported in Table 6.

M Token-level Attention Aggregation Details

This section details the implementation of the token-level attention diagnostics used in Section 5.3. All attention weights are extracted from Qwen3-8B hooks and correspond to post-softmax decoder self-attention under greedy decoding.

Attention extraction. For each generation, we collect the self-attention tensor at each layer and head, yielding attention weights $A^{(\ell,h)}_{t\to k}$ from generation query position t to key token position k. Attention is defined over the full prompt token sequence, including template tokens.

Generation positions. Let $\mathcal{T}$ denote the set of template-defined generation positions used for analysis. In practice, $\mathcal{T}$ corresponds to the decoding positions associated with the content-bearing segment of the output. No additional filtering based on token type (e.g., punctuation) is applied beyond this template-based selection.

Token span identification. Token spans corresponding to the idiom and (when present) the gloss are identified by mapping character offsets in the prompt text to token index ranges using the model tokenizer. These spans are defined with respect to the full prompt tokenization and are not re-indexed to exclude template tokens.

Token-level attention mass. For each key token position k, we compute the aggregated attention mass

$$\bar{A}(k) = \frac{1}{|\mathcal{T}|LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{t\in\mathcal{T}} A^{(\ell,h)}_{t\to k}.$$

This quantity represents the average post-softmax attention mass assigned to token k across layers, heads, and selected generation positions.

Using the aggregated attention mass $\bar{A}(k)$, we define the token-level diagnostics reported in Section 5.3.