How Large Language Models Translate Chinese Metaphors?

Anonymous ACL submission

Abstract

Large language models (LLMs) achieve strong performance in machine translation, yet their underlying mechanisms remain unclear for metaphorical language processing. We investigate whether LLMs rely on semantic understanding or learned alignment patterns using a Chinese–English dataset that includes conventional metaphors, novel metaphors, and non-metaphorical expressions, annotated with metaphor topic (tenor) and imagery (vehicle) and accompanied by human reference translations.

Across models of different scales, we observe a consistent dissociation between metaphor identification and translation quality: recognition of metaphorical meaning declines in more challenging cases, while translation performance often remains stable. This gap suggests that accurate translation does not necessarily reflect human-like semantic processing, but may instead be supported by learned cross-lingual correspondences. Moreover, prompting models to identify metaphorical meaning before translation substantially improves smaller models but yields limited gains for larger ones, indicating scale-dependent sensitivity to explicit semantic guidance.

1 Introduction

In recent years, large language models (LLMs) have made strong progress in machine translation. Prior work has focused mainly on improving translation quality (Vilar et al., 2023; Jiao et al., 2023; Cui et al., 2025), while paying less attention to how LLMs translate—especially when handling abstract or metaphorical language. It remains unclear whether LLMs follow a human-like process, first understanding the source meaning and then producing a translation, or whether they mainly rely on statistical patterns learned during training.

Metaphor translation offers a useful test case. Unlike literal expressions, metaphors often require the translator to infer the intended meaning behind the surface form before generating an appropriate expression in the target language. For example, the Chinese metaphor "热锅上的蚂蚁" literally means "ants on a hot pan," but it describes a state of intense anxiety. A natural English translation would therefore be "like a cat on a hot tin roof" or a suitable paraphrase. Successful metaphor translation depends on capturing underlying meaning rather than matching words directly.

To examine how LLMs handle this challenge, we construct a new Chinese–English metaphor translation dataset with three categories: (1) conventional metaphors likely seen during training, (2) novel metaphors that emerged after 2025 and are less likely to have been extensively represented in publicly documented pretraining corpora, and (3) non-metaphorical expressions that require minimal inference. This design helps separate semantic understanding from pattern matching. If models struggle to identify a metaphor's meaning but still produce fluent translations, this would suggest reliance on learned correspondences. Better performance on non-metaphorical expressions than metaphors, and on conventional than novel metaphors, would further support this view. In contrast, strong and consistent performance across all categories would provide stronger evidence for a more meaning-driven translation process.

We also test whether prompting models to explicitly identify metaphorical meaning improves translation quality. We treat identifying metaphorical meaning as a proxy for source-language understanding and examine

whether a two-step prompting strategy—identification followed by translation—leads to better results.

Our study addresses three research questions: 1) How accurately can LLMs identify and translate metaphorical expressions? 2) Do models of different sizes rely more on semantic understanding or on learned alignment patterns when translating metaphors? 3) Does prompting models to first identify metaphorical meaning improve translation quality?

Our contributions are threefold:

- We introduce a Chinese–English metaphor translation corpus annotated with metaphor topic and imagery, human reference translations, and translation strategy labels. The dataset will be publicly released upon publication.

- We identify a systematic gap between metaphor identification and translation quality: as identification accuracy declines in more complex cases, translation performance often remains stable, suggesting reliance on learned alignment patterns rather than full semantic understanding.

- We show that identification–translation prompting has scale-dependent effects: smaller models benefit substantially from explicit structural guidance, whereas larger models are less sensitive and exhibit more stable lexical choices across prompting conditions.

2 Related works

Research on automatic metaphor identification has evolved from early rule-based systems (Fass, 1991) to data-driven approaches. Recent work typically formulates metaphor identification as binary classification (Birke and Sarkar, 2006; Rai et al., 2017) or sequence labeling (Gao et al., 2018; Wu et al., 2018), while more recent studies explore LLM-based approaches (Fuoli et al., 2026; Sanchez-Bayona and Agerri, 2025). These methods are commonly evaluated on benchmark datasets such as the VUA Corpus (Steen et al., 2010). Building on Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 2008; Lakoff and Turner, 2024), subsequent studies extend beyond metaphor identification to metaphor domain recognition (Rosen, 2018; Ge et al., 2022; Wachowiak and Gromann, 2023) and broader metaphor understanding (Ye et al., 2025; Sengupta et al., 2025). Some work further examines models' ability to extract metaphor components, namely the *tenor* and *vehicle* (Liu et al., 2026; Richards, 1965).

In translation studies, scholars have proposed criteria for metaphor translation quality. Newmark (1988) argues that translation should prioritize conveying sense and imagery rather than literal form. Similarly, Schäffner (2004) emphasizes preserving underlying conceptual structure over surface metaphorical expressions. For example, translating the German *Freundschaftsbrücke* ("bridge of friendship") simply as "friendship" may be acceptable if the core meaning is retained.

Recent work investigates how machine translation systems and LLMs handle metaphor (Wang et al., 2024; Tang et al., 2024; Liu et al., 2023). Reported issues include overly literal renderings and omission of metaphorical meaning (Dorst, 2023; Karakanta et al., 2025). Benchmarks such as MMTE (Wang et al., 2024) have been introduced to evaluate metaphor translation performance.

However, prior studies largely treat metaphor identification and translation as separate tasks, without examining their relationship. It remains unclear whether successful metaphor translation reflects structured semantic understanding or reliance on contextual similarity and statistical alignment. While it is commonly assumed that metaphor understanding supports translation, empirical evidence linking the two is limited. Identification analyzes metaphor structure within a language, whereas translation requires re-expressing meaning across languages.

Our work bridges this gap by jointly evaluating LLMs' metaphor identification and translation performance to investigate the mechanisms underlying metaphor translation.

3 Dataset

Our dataset consists of two self-constructed corpora, totaling 580 sentences: 380 metaphorical instances and 200 non-metaphorical base-

lines. Each metaphorical sentence is annotated with its two core components: **tenor**, the topic being described, and **vehicle**, the figurative image used to describe it. For example, in "time is money," "time" is the tenor and "money" is the vehicle. These annotations enable separate evaluation of metaphor identification and component recognition, providing a basis for assessing LLMs' metaphor understanding.

The dataset consists of three categories: conventional metaphors, novel metaphors, and a non-metaphorical control set (see Table 1). Conventional metaphors ($N = 350$) are drawn from CMC (Li et al., 2022) ($N = 100$), CMDAG (Shao et al., 2024) ($N = 100$), the UM-Corpus (Tian et al., 2014) (CC BY-NC-ND 4.0) ($N = 50$), and 100 additional instances from poetry, fiction, and song lyrics. Novel metaphors ($N = 30$) are collected from Chinese social media platforms (Weibo, Xiaohongshu, and Douyin), consisting of metaphorical Internet slang expressions attested in timestamped posts from 2025 onwards and manually verified as involving metaphorical mappings during data collection (March 2026). The non-metaphorical control set ($N = 200$) is drawn from the UM-Corpus and manually verified to exclude metaphorical expressions. All sentences are limited to 50 Chinese characters. CMC and CMDAG are publicly available for research use. For social media data, only textual content is retained; no usernames, account IDs, or other personally identifiable information are collected. All examples are manually screened to remove offensive or sensitive content.

The conventional and non-metaphorical sentences largely consist of high-frequency expressions likely seen during model training, whereas the novel metaphors are unlikely to have appeared in existing corpora. This design allows us to test whether LLMs rely primarily on memorized distributional patterns or on meaning-based processing during metaphor translation.

Metaphor identification followed the Metaphor Identification Procedure (MIP) (Group, 2007). Two annotators independently performed identification, achieving high inter-annotator agreement (Cohen's $\kappa = 0.98$). For confirmed metaphors, tenor–vehicle boundaries were determined through joint discussion. Ambiguous cases were excluded, resulting in the final set of 580 sentences. Among them, 250 UM-Corpus sentences include existing English translations. The remaining 330 sentences were translated by two certified Chinese–English professional translators.

Type	Source	N	T-V	Trans.
Conv.	CMDAG	100	src	ours
	CMC	100	src	ours
	PNL	100	ours	ours
	Idioms (UM-corpus)	50	ours	src
Novel	Internet slang	30	ours	ours
Non-meta.	UM-corpus	200	ours	src

Table 1: Corpus composition. Conv. = Conventional; Non-meta. = Non-metaphorical; T-V = tenor–vehicle annotation; Trans. = human reference translation; PNL = self-collected Poems/novels/lyrics src = existing annotation/translation; ours = annotated/translated in-house.

4 Experimental Setting

4.1 Models

We evaluate metaphor processing across four major LLMs: Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Team, 2024), GPT-5, and DeepSeek-Chat (Bi et al., 2024). Qwen models are deployed locally using HuggingFace Transformers, while GPT-5 and DeepSeek-Chat are accessed through their respective APIs. All experiments use deterministic decoding with temperature set to 0.

4.2 Prompts

We design two prompting frameworks: *direct translation* and *identification–translation*.

For **direct translation**, we use three prompting conditions:

Prompt 1 (Basic). The model is instructed to directly translate a Chinese sentence into English without additional guidance: "Translate the following Chinese sentence into English."

Prompt 2 (Natural). The model is further encouraged to produce natural and idiomatic English expressions: "Translate the following Chinese sentence into natural and id-

iomatic English. Where appropriate, use common or corresponding expressions in English."

Prompt 3 (Few-shot). The model is provided with six Chinese–English translation examples prior to translation in order to supply stylistic and translational reference patterns.

For **identification–translation**, we use two prompting conditions:

Prompt 4 (Chain-of-Thought, CoT). The model is first instructed to determine whether the sentence contains metaphorical language following the MIP procedure. If a metaphor is identified, the model is further asked to identify the tenor and vehicle before generating the translation.

Prompt 5 (Gold Structure). The model is provided with manually annotated tenor and vehicle information as explicit structural guidance prior to translation.

Prompts 1–4 are applied to the full corpus of 580 sentences; Prompt 5 is applied to the 380 metaphorical sentences only, as it requires gold tenor–vehicle annotation as input. The full prompts are provided in Appendix A.

4.3 Evaluation – Metaphor Identification

We evaluate metaphor identification using accuracy, precision, recall, and F1. Component recognition is assessed with both exact string matching and semantic similarity, as model outputs often vary lexically while remaining semantically appropriate.

For exact matching, a prediction is considered *correct* if both the identified topic and imagery match the gold annotations, *partially correct* if only one matches, and *incorrect* otherwise.

To account for paraphrastic variation, we additionally compute cosine similarity between model outputs and gold annotations. Following prior work (Mohamed and Oussalah, 2020), we adopt a threshold of 0.7 to indicate semantic equivalence. A prediction is deemed semantically correct if both components reach cosine similarity scores of at least 0.7 with their respective gold references.

4.4 Evaluation – Metaphor Translation

Translation quality is evaluated using automatic metrics and manual assessment. We report BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and COMET (Rei et al., 2022).

Preliminary analysis showed relatively high automatic scores on the Conv-CMC and Conv-CMDAG subsets, likely due to their structurally explicit metaphors (e.g., marked by "像" or "是"), which are easier to translate. In contrast, the other subsets contain more implicit metaphors with less explicit structure. We therefore conduct manual evaluation on the remaining 180 metaphorical instances across four models and five prompting conditions, yielding 3,600 translations.

Two annotators with Master's degrees in translation studies independently evaluated the outputs. Following functional equivalence theory (Waard and Nida, 1986) and cognitive approaches to metaphor translation (Schäffner, 2004), we assess whether translations convey the intended meaning and pragmatic effect rather than literal correspondence.

Each translation is scored as 1 if it preserves the intended meaning and is linguistically appropriate, and 0 otherwise. Final labels are assigned as follows: 1 if both annotators assign 1, 0 if they disagree, and −1 if both assign 0.

4.5 Translation Strategies

We adopt three metaphor translation strategies from Van den Broeck (1981) and add a mistranslation category.

(1) Paraphrase. The source metaphor is rendered as a non-metaphorical expression in the target language (e.g., translating "热锅上的蚂蚁" as "a state of intense anxiety").

(2) Strict translation. The source topic (tenor) and imagery (vehicle) are directly transferred into the target language. This yields natural translations when equivalent metaphors exist cross-linguistically (e.g., "她是黑马" → "She's a dark horse"), but may sound unnatural when no corresponding vehicle exists.

(3) Substitution. The source imagery (vehicle) is replaced with a culturally appropriate target-language metaphor conveying a similar meaning (e.g., "落汤鸡" → "wet rat").

(4) Mistranslation. The output fails to preserve the intended meaning, including semantically unrelated, severely distorted, untrans-

lated, or otherwise invalid translations.

Although a few cases involve addition or reduction, these are rare. Our analysis focuses on the major strategies above to compare prompting effects, model behavior, and human translation preferences.

5 Results

5.1 Metaphor Identification

We first evaluate models' ability to detect metaphors and recognize their components (tenor and vehicle).

Model	Accuracy	Precision	Recall	F1
Qwen-3B	41.2	100.0	10.3	18.6
Qwen-7B	92.2	94.9	93.2	94.0
GPT-5	93.4	93.4	96.8	95.1
DeepSeek	**96.0**	**97.3**	**96.6**	**97.0**

Table 2: Metaphor identification performance of four models.

Table 2 reports overall metaphor identification performance. With the exception of Qwen-3B, all models achieve high accuracy, precision, recall, and F1. DeepSeek performs best overall (F1 = 97.0), followed closely by GPT-5 (95.1) and Qwen-7B (94.0). In contrast, Qwen-3B exhibits extremely low recall (10.3), indicating that it fails to detect a substantial proportion of metaphorical expressions. These results suggest that smaller models face clear limitations in metaphor identification.

We next examine tenor–vehicle recognition (Table 3). DeepSeek achieves the highest or near-highest scores across genres under both string-based (Str-C) and semantic-based (Sem-C) evaluation, indicating a comparatively strong ability to capture structural relations in Chinese metaphors.

Performance varies systematically by metaphor type. On the Conv-CMC and Conv-CMDAG subsets—where both tenor and vehicle are explicitly marked—models (except Qwen-3B) perform very well, with semantic matching scores ranging from 0.77 to 0.88. However, performance drops substantially for more implicit metaphor types, including Conv-PNL, Conv-Idioms, and Novel-Slang, in which only one of the tenor or vehicle is explicitly expressed, or where no explicit comparator is present to connect them. For example, even strong models achieve Sem-C scores below 0.32 on idioms and below 0.43 on novel slang. This pattern indicates that explicit structural cues greatly facilitate component recognition, whereas implicit or culturally embedded metaphors remain challenging.

Genre	Model	Acc.	Str-C	Sem-C
Overall	Qwen-3B	0.10	0.07	0.08
	Qwen-7B	0.93	0.46	0.48
	GPT-5	0.97	0.50	0.54
	DeepSeek	0.97	0.56	0.60
Conv-CMC+DAG	Qwen-3B	0.17	0.13	0.14
	Qwen-7B	0.99	0.79	0.82
	GPT-5	0.98	0.77	0.78
	DeepSeek	0.99	0.85	0.88
Conv-PNL	Qwen-3B	0.04	0.02	0.02
	Qwen-7B	0.92	0.12	0.14
	GPT-5	0.97	0.19	0.25
	DeepSeek	0.93	0.20	0.24
Conv-Idioms	Qwen-3B	0.00	0.00	0.00
	Qwen-7B	0.80	0.06	0.06
	GPT-5	0.98	0.24	0.32
	DeepSeek	1.00	0.24	0.30
Novel-Slang	Qwen-3B	0.00	0.00	0.00
	Qwen-7B	0.80	0.07	0.07
	GPT-5	0.87	0.23	0.27
	DeepSeek	0.87	0.37	0.43

Table 3: Metaphor component recognition results across models and corpora (N = total number of instances per genre: Overall = 380, Conv-CMC+DAG = 200, Conv-PNL = 100, Conv-Idioms = 50, Novel-Slang = 30). Acc. = metaphor identification accuracy; Str-C/Sem-C = proportion of Correct predictions under string-based and semantic-based evaluation respectively. Light blue indicates the highest Str-C or Sem-C score per genre.

5.2 Metaphor Translation

We next examine translation performance. To validate automatic metrics, we compute correlations with human evaluation across 3,600 model-generated translations. At the system level, COMET shows the strongest correlation with human judgments (Spearman's $r = 0.85$, $p < .001$), followed by BERTScore ($r = 0.76$, $p < .001$), while BLEU exhibits substantially weaker correlation. Accordingly, we rely primarily on COMET in subsequent analyses.

Table 4 compares metaphorical and non-metaphorical sentences. Across all models, non-metaphorical baseline sentences re-

	Metaphor			Non-meta.		
Model	BS	**CM**	BL	BS	**CM**	BL
Qwen-3B	0.57	0.76	19.55	0.57	0.82	18.91
Qwen-7B	0.59	0.78	22.66	0.58	0.83	21.23
GPT-5	0.65	0.81	28.91	0.61	0.84	23.43
DeepSeek	0.66	0.81	30.49	0.63	0.84	27.50

Table 4: Translation scores averaged across prompts. BS = BERTScore; CM = COMET; BL = BLEU. BERTScore and COMET: mean; BLEU: corpus-level.

	Conventional			Novel		
Model	BS	**CM**	BL	BS	**CM**	BL
Qwen-3B	0.58	0.77	20.56	0.42	0.67	7.69
Qwen-7B	0.61	0.79	23.89	0.43	0.67	8.39
GPT-5	0.66	0.81	29.94	0.54	0.73	16.95
DeepSeek	0.68	0.82	31.70	0.53	0.72	16.33

Table 5: Translation scores averaged across prompts (non-metaphorical sentences excluded). BS = BERTScore; CM = COMET; BL = BLEU. BERTScore and COMET: mean; BLEU: corpus-level

ceive consistently higher scores than metaphorical expressions. Moreover, GPT-5 and DeepSeek substantially outperform Qwen-3B and Qwen-7B. Table 5 further distinguishes conventional and novel metaphors (excluding non-metaphorical sentences). Conventional metaphors are translated much more accurately than novel metaphors.

This performance gap is especially pronounced for smaller models, but remains visible even for GPT-5 and DeepSeek. Together, these results suggest that translation quality is strongly influenced by familiarity and distributional frequency.

For manual evaluation, we focus on 180 metaphorical instances (150 conventional and 30 novel), which show lower component recognition performance. Across four models and five prompting conditions, this yields 3,600 translations. Two annotators independently evaluated the outputs, achieving 88.2% agreement and Cohen's $\kappa = 0.572$, indicating moderate agreement.

5.3 Translation Mechanism

We now relate metaphor understanding to translation performance. Figure 1 compares component recognition with human translation scores. A systematic dissociation emerges:

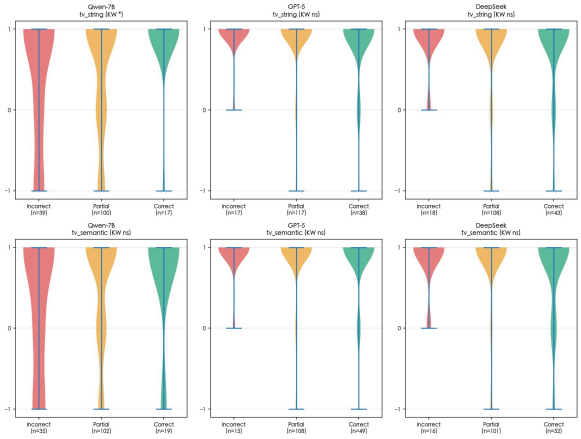

Figure 1: Correlation between metaphor component recognition and translation scores across models and corpora. Qwen2.5-3B is excluded from the figure due to near-zero metaphor identification rates across all conditions.

translation scores remain consistently high even when component recognition is incorrect or only partially correct. Under a human-like understanding-to-generation account, translation quality would be expected to decline substantially when metaphor component recognition fails. However, this pattern is not consistently observed in the results.

Combined with earlier findings—namely, that non-metaphorical expressions outperform metaphorical ones, and conventional metaphors outperform novel metaphors—this dissociation suggests that metaphor translation may rely more on semantic-space alignment and distributional correspondence than on explicit metaphor understanding.

5.4 Translation Strategies

To analyze strategy use, we employ *Claude Sonnet 4.6* for initial annotation, followed by human verification and correction. Prompts are detailed in Appendix A. Prompting conditions exert limited influence on strategy distribution; instead, metaphor type is the primary determinant. We therefore analyze Conv-PNL, Conv-idioms, and Novel-slang separately (Figure 2).

Overall, LLMs show broadly similar strategy patterns to human translators. For Conv-PNL metaphors, models predominantly use *strict translation*. For Conv-Idioms, they employ more *substitution* and *paraphrase*. Translation errors are relatively infrequent overall.

This pattern helps explain the earlier dissociation between weak component recognition and relatively strong translation performance. For Conv-PNL, many expressions involve conventionalized or creatively constructed metaphors (e.g., poems and fiction), allowing direct lexical transfer to incidentally preserve much of the intended meaning even without understanding metaphors. For Conv-Idioms, frequent exposure during training likely enables models to retrieve appropriate target-language equivalents through learned alignment. Together, these findings suggest that acceptable performance on conventional metaphors may rely more on distributional semantic alignment and familiarity rather than human-like metaphor understanding.

However, this mechanism becomes less reliable for Novel-Slang metaphors. Models rely on *strict translation* more frequently than human translators, and most incorrect outputs occur under this strategy, suggesting difficulty adapting when cross-linguistic correspondences are weak or culturally specific.

Human translators, in contrast, more flexibly apply *substitution* and *paraphrase* to Novel-Slang, inferring pragmatic implications and adapting to target-language conventions. They also occasionally employ addition and reduction, strategies rarely used by models.

5.5 Prompting Effect

Finally, we examine prompting effects (Table 6). Identification–translation prompting (Prompts 4–5) benefits smaller models more substantially. For Qwen-3B and Qwen-7B, the highest human evaluation scores for both conventional and novel metaphors consistently occur under these prompts, suggesting that explicit structural scaffolding supports metaphor translation in lower-capacity models.

In contrast, GPT-5 and DeepSeek show limited sensitivity to prompting strategy for conventional metaphors, with performance remaining largely stable across all five prompts. For novel metaphors, GPT-5 achieves its highest score under Prompt 5, whereas DeepSeek exhibits relatively low sensitivity overall, indicating that explicit tenor–vehicle guidance provides limited additional benefit.

Overall, identification–translation prompting improves metaphor translation, particularly for smaller models.

Model	Prompt	Conv.	Novel
	1	0.55	−0.33
	2	0.59	−0.13
	3	0.51	−0.27
	4	0.59	0.03
Qwen-3b	5	0.62	0.07
	1	0.59	0.00
	2	0.55	−0.03
	3	0.61	−0.13
	4	0.61	0.00
Qwen-7b	5	0.65	0.30
	1	0.94	0.47
	2	0.94	0.53
	3	0.95	0.40
	4	0.93	0.50
GPT-5	5	0.93	0.70
	1	0.91	0.63
	2	0.91	0.83
	3	0.87	0.63
	4	0.86	0.47
DeepSeek	5	0.91	0.63

Table 6: Human Evaluation results (range from -1 to 1). Light blue indicates the highest scores achieved under *Identification-Translation Prompting (4-5)*, while light green indicates the highest scores under *Direct Translation Prompting (1-3)*.

6 Discussion and Conclusion

This study introduces a Chinese–English metaphor translation dataset and systematically investigates the relationship between metaphor identification and metaphor translation in large language models. Our central question is whether successful metaphor translation depends on explicit metaphor understanding.

Across analyses, a consistent pattern emerges: metaphor understanding and metaphor translation are only partially aligned. While larger models demonstrate strong performance in both identification and translation, fine-grained results reveal that accurate tenor–vehicle recognition is not a necessary condition for producing acceptable translations. Models can generate fluent and semantically adequate outputs even when their structural analysis of the metaphor is incomplete or inaccurate.

This dissociation suggests that LLM-based metaphor translation does not strictly follow a human-like pipeline of "understand metaphor

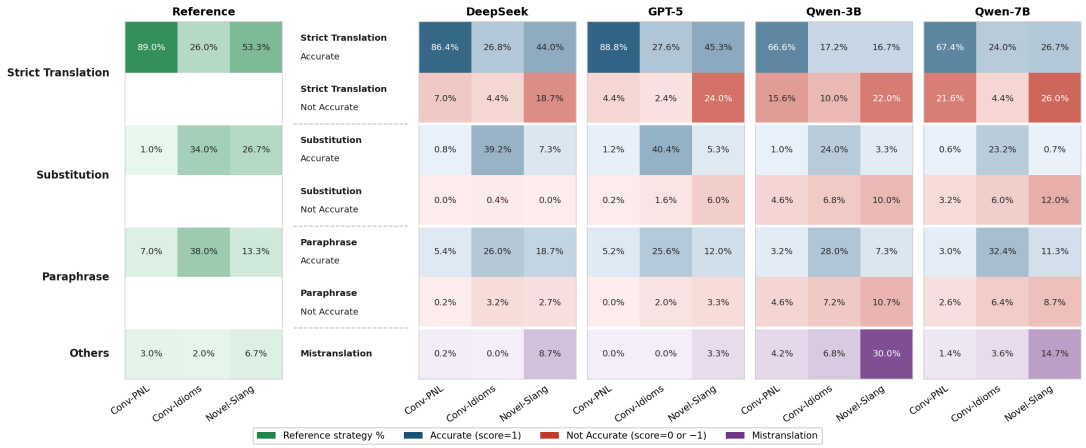

Figure 2: Translation strategy distribution and accuracy by metaphor type and model. Values are percentages averaged across five prompting conditions. Minor strategies (addition and reduction, each <3%) are omitted for brevity.

→ generate translation." Instead, translation appears to rely heavily on semantic-space alignment and distributional correspondence across languages. When metaphors are conventional and distributionally well represented, cross-lingual mappings are readily available, enabling high-quality translation without explicit metaphor reasoning. In contrast, culturally specific and novel metaphors—particularly Internet slang—lack stable cross-linguistic correspondences, making direct transfer less reliable and increasing the need for adaptive strategies.

Our strategy analysis reinforces this interpretation. Models predominantly employ strict transfer for structurally transparent metaphors, a strategy that often succeeds when cross-lingual mappings are strong. However, for novel metaphors, strict transfer frequently results in errors, whereas human translators more flexibly adopt substitution and paraphrase to accommodate pragmatic and cultural differences. This contrast highlights a key limitation of current LLMs: relative rigidity in situations requiring culturally grounded reinterpretation.

Prompting results further clarify the mechanism. Explicit identification–translation prompting benefits smaller models, suggesting that structural scaffolding can partially compensate for limited internalized representations. In stronger models, however, prompting exerts limited influence, implying that their performance primarily reflects entrenched distributional knowledge rather than dynamically constructed structural analysis.

Taken together, our findings point to a partial independence between metaphor understanding and translation in LLMs. The results are consistent with the hypothesis that cross-lingual distributional alignment and familiarity play a substantial role in metaphor translation performance. These results contribute empirical evidence to ongoing debates about the nature of semantic processing in LLMs and provide new insight into the mechanisms underlying cross-linguistic metaphor translation.

7 Limitations

First, the size of the novel metaphor subset is relatively limited, as it is difficult to collect and reliably verify expressions that emerged after 2025. Although we apply strict filtering criteria based on source tracing and usage timestamps, the resulting dataset remains small compared to the other subsets, which may constrain the generalizability of findings related to novel metaphor processing.

Second, our experiments are restricted to Chinese–English translation. While this setting allows for controlled investigation of metaphor transfer mechanisms, it remains unclear whether the observed patterns generalize to other language pairs, particularly those with different typological or cultural metaphor systems. In the future, we may expand this experiment to other languages.

References

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443.

Aletta G Dorst. 2023. Metaphor in literary machine translation: style, creativity and literariness. In *Computer-Assisted Literary Translation*, pages 173–186. Routledge.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2026. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Applied Corpus Linguistics*, page 100204.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10681–10689.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Alina Karakanta, Mayra Nas, and Aletta G. Dorst. 2025. Metaphors in literary machine translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 276–286, Geneva, Switzerland. European Association for Machine Translation.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago Press.

George Lakoff and Mark Turner. 2024. *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In *Proceedings of the 29th international conference on computational linguistics*, pages 6468–6479.

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.

Hongde Liu, Chenyuan He, Senbin Zhu, Changyong Niu, and Yuxiang Jia. 2026. Metaphor components identification with feedback-enhanced feature-driven in-context learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(4):1–30.

Muhidin Mohamed and Mourad Oussalah. 2020. A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics. *Language Resources and Evaluation*, 54(2):457–485.

Peter Newmark. 1988. *A textbook of translation*, volume 66. Prentice hall New York.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sunny Rai, Shampa Chakraverty, and Devendra K Tayal. 2017. Identifying metaphors using fuzzy conceptual features. In *Proceedings of the International Conference on Information, Communication and Computing Technology*, pages 379–386. Springer.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabelist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

I. A. Richards. 1965. *The philosophy of rhetoric*. Oxford University Press.

Zachary Rosen. 2018. Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based

construction grammatical cues. In *Proceedings of the Workshop on Figurative Language Processing*, pages 102–109, New Orleans, Louisiana. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477.

Christina Schäffner. 2004. Metaphor and translation: some implications of a cognitive approach. *Journal of Pragmatics*, 36(7):1253–1269.

Meghdut Sengupta, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, Debanjan Ghosh, and Henning Wachsmuth. 2025. Investigating the impact of conceptual metaphors on llm-based nli through shapley interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17393–17403.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Wenhao Huang, Ge Zhang, and Jie Fu. 2024. Cmdag: A chinese metaphor dataset with annotated grounds as cot for boosting metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366.

Gerard Steen, Lettie Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. A method for linguistic metaphor identification from mip to mipvu. *Converging Evidence in Language and Communication Research*, 14.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and context-aware translation of east asian idioms with gpt-4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9285–9305.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*, pages 1837–1842.

Raymond Van den Broeck. 1981. The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4):73–87.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Jan de Waard and Eugene Albert Nida. 1986. *From one language to another: Functional equivalence in Bible translating*. Nelson.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024. Mmte: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 11343–11358.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Fengying Ye, Shanshan Wang, Lidia S Chao, and Derek F Wong. 2025. Unveiling llms' metaphorical understanding: Exploring conceptual irrelevance, context leveraging and syntactic influence. *arXiv preprint arXiv:2510.04120*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A: Prompts

A.1 Direct Translation Prompts

Prompt 1 (Basic)

Translate the following Chinese sentence into English. Only output the English translation. Do not explain.

Prompt 2 (Natural)

Translate the following Chinese sentence into natural and idiomatic English. Where appropriate, use common or corresponding expressions in English. Only output the English translation. Do not explain.

Prompt 3 (Few-shot)

Below are some examples of Chinese-to-English translation. You may refer to the style of these examples.

Example 1:
中文：有没有帽子叔叔来回答个问题？执勤的时候是希望被叫”帅哥”还是”同志”。
英文：Any guys in blue here to answer a question? When you're on duty, do you prefer being called 'handsome' or 'comrade'?

Example 2:
中文：中华民族像一个巨人，屹立世界在东方。
英文：The Chinese nation stands like a giant in the East.

Example 3:
中文：我今天飞到东明天飞到西，就像一只小蜜蜂采花蜜。
英文：Flying east today, west tomorrow, I'm just like a little bee gathering nectar.

Example 4:
中文：我关掉月亮，你却更明亮。
英文：I turn off the moon, yet you shine even brighter.

Example 5:
中文：这种产品一直是我们公司的摇钱树。
英文：This product has always been a cash cow for our company.

Example 6:
中文：判断一个人应根据他的行动，而不是他的言词。
英文：A man should be judged by his deeds, not his words.

Now translate the following Chinese sentence into English. Only output the English translation. Do not explain.

A.2 Identification–Translation Prompts

Prompt 4 (Chain-of-Thought)

You are a professional Chinese-to-English translator and an expert metaphor annotator. For each Chinese sentence, do two tasks:

1. Identify whether the sentence contains metaphorical use. If metaphorical, identify the tenor and vehicle.
2. Translate the sentence into English.

Use the Metaphor Identification Procedure (MIP) in a practical way:

- First understand the overall meaning of the sentence.
- For relevant words or expressions, compare their contextual meaning with a more basic, concrete, physical, or historically earlier meaning.
- If the contextual meaning contrasts with the basic meaning but can be understood through comparison with it, treat it as metaphorical use.

Decision rule: If any word or expression is used metaphorically, set `"is_metaphor"` to `1`. Otherwise set it to `0`.

If metaphorical: `"tenor"` = the target topic being described (in Chinese); `"vehicle"` = the source image or concept used to describe it (in Chinese).
If not metaphorical: `"tenor"` = `""`, `"vehicle"` = `""`.

Translation rule: The `"translation"` field must contain English only. Do not leave any Chinese characters in the translation.

Output ONLY valid JSON in exactly this format:

```
{
  "is_metaphor": 0,
  "tenor": "",
  "vehicle": "",
  "translation": ""
}
```

Do not output explanations, Markdown, bullet points, or any text outside the JSON.

Prompt 5 (Gold Structure)

You are a professional Chinese-to-English translator. Translate the Chinese sentence into natural English.

Chinese sentence:
`[sentence]`

If `is_metaphor = 1`:
Metaphor annotation:
- Tenor: `[tenor]`
- Vehicle: `[vehicle]`

The annotation is provided only to help understand the sentence meaning.

If `is_metaphor = 0:`
No metaphor annotation is provided.

Return only the English translation. Use English only. Do not output Chinese characters. Do not explain.

A.3 Strategy Annotation Prompt

Strategy Annotation Prompt

You are a translation strategy annotator. Given a Chinese source sentence containing a metaphor and its English translation, identify the translation strategy used for the **metaphorical expression only** (ignore non-metaphorical parts).

Annotate using **one** of the following labels:

- **strict translation**: both tenor and vehicle are preserved in the target language.
- **substitution**: the source vehicle is replaced by a different target vehicle conveying the same tenor.
- **paraphrase**: the metaphor is rendered as a non-metaphorical expression.
- **mistranslation**: the translation fails to preserve the intended meaning (semantically unrelated, severely distorted, untranslated, or otherwise invalid).

Example:

> 他通过走后门把他儿子弄进了一所飞行学校。
> [metaphor: **走后门**]

- "He got his son into a flight school through *backdoor connections*." → **strict translation**
- "He *pulled some strings* to get his son into a flight school." → **substitution**
- "He *used his personal connections* to get his son into a flight school." → **paraphrase**

Return **only** the label, nothing else.

Appendix B: Detailed Results

Genre	Model	N	Met. Acc.	Str-T	Str-P	Str-F	Sem-T	Sem-P	Sem-F
Overall	Qwen-3B	380	0.10	0.07	0.02	0.90	0.08	0.02	0.90
	Qwen-7B	380	0.93	0.46	0.33	0.21	0.48	0.32	0.20
	GPT-5	380	0.97	0.50	0.38	0.12	0.54	0.36	0.11
	DeepSeek	380	0.97	0.56	0.33	0.11	0.60	0.30	0.11
Conv-CMC+CMDAG	Qwen-3B	200	0.17	0.13	0.04	0.83	0.14	0.03	0.83
	Qwen-7B	200	0.99	0.79	0.12	0.08	0.82	0.10	0.07
	GPT-5	200	0.98	0.77	0.13	0.11	0.78	0.14	0.09
	DeepSeek	200	0.99	0.85	0.08	0.07	0.88	0.06	0.07
Conv-PNL	Qwen-3B	100	0.04	0.02	0.01	0.97	0.02	0.01	0.97
	Qwen-7B	100	0.92	0.12	0.54	0.34	0.14	0.54	0.32
	GPT-5	100	0.97	0.19	0.67	0.14	0.25	0.62	0.13
	DeepSeek	100	0.93	0.20	0.61	0.19	0.24	0.58	0.18
Conv-Idioms	Qwen-3B	50	0.00	0.00	0.00	1.00	0.00	0.00	1.00
	Qwen-7B	50	0.80	0.06	0.60	0.34	0.06	0.64	0.30
	GPT-5	50	0.98	0.24	0.66	0.10	0.32	0.60	0.08
	DeepSeek	50	1.00	0.24	0.68	0.08	0.30	0.64	0.06
Novel-Slang	Qwen-3B	30	0.00	0.00	0.00	1.00	0.00	0.00	1.00
	Qwen-7B	30	0.80	0.07	0.53	0.40	0.07	0.53	0.40
	GPT-5	30	0.87	0.23	0.57	0.20	0.27	0.53	0.20
	DeepSeek	30	0.87	0.37	0.43	0.20	0.43	0.37	0.20

Table 7: Results of metaphor processing across all models. Conv. = Conventional; Met. Acc. = metaphor identification accuracy; Str./Sem.-T/P/F = the proportions of True, Partial, and False predictions under string-based and semantic-based evaluation in structural component recognition.

		Metaphor			Non-meta.		
Model	**P**	BS	CM	BL	BS	CM	BL
Qwen-3B	P1	0.56	0.76	19.36	0.57	0.82	20.05
	P2	0.57	0.76	19.76	0.57	0.82	18.66
	P3	0.56	0.76	18.74	0.56	0.82	18.44
	P4	0.58	0.77	20.12	0.57	0.82	18.48
	P5	0.58	0.77	19.77	–	–	–
Qwen-7B	P1	0.59	0.78	23.17	0.59	0.83	21.56
	P2	0.59	0.78	22.27	0.59	0.83	21.81
	P3	0.59	0.78	22.60	0.57	0.83	21.09
	P4	0.59	0.78	22.83	0.58	0.83	20.47
	P5	0.59	0.78	22.43	–	–	–
GPT-5	P1	0.66	0.81	30.37	0.62	0.84	24.37
	P2	0.64	0.80	27.53	0.60	0.84	22.57
	P3	0.65	0.80	29.13	0.61	0.84	22.69
	P4	0.66	0.81	29.36	0.61	0.84	24.08
	P5	0.65	0.81	28.19	–	–	–
DeepSeek	P1	0.66	0.81	30.21	0.64	0.85	28.57
	P2	0.67	0.81	30.50	0.63	0.85	27.88
	P3	0.66	0.81	30.76	0.63	0.84	26.85
	P4	0.66	0.81	29.90	0.63	0.84	26.69
	P5	0.67	0.81	31.06	–	–	–

Table 8: Translation scores across prompts. BS = BERTScore; CM = COMET; BL = BLEU.

		Conventional			Novel		
Model	**P**	BS	CM	BL	BS	CM	BL
Qwen-3B	P1	0.57	0.76	20.43	0.42	0.66	6.83
	P2	0.58	0.77	20.70	0.42	0.66	8.72
	P3	0.57	0.77	19.71	0.42	0.66	7.41
	P4	0.59	0.77	21.19	0.44	0.68	7.65
	P5	0.59	0.77	20.79	0.42	0.68	7.83
Qwen-7B	P1	0.61	0.78	24.34	0.43	0.67	9.58
	P2	0.60	0.79	23.49	0.41	0.67	8.10
	P3	0.60	0.79	23.80	0.44	0.67	8.66
	P4	0.61	0.79	24.33	0.42	0.68	5.35
	P5	0.61	0.79	23.47	0.42	0.68	10.28
GPT-5	P1	0.67	0.82	31.64	0.54	0.73	15.50
	P2	0.65	0.81	28.42	0.52	0.72	17.15
	P3	0.66	0.81	29.97	0.56	0.74	19.27
	P4	0.67	0.82	30.45	0.56	0.74	16.58
	P5	0.66	0.81	29.21	0.53	0.74	16.26
DeepSeek	P1	0.67	0.82	31.33	0.53	0.72	17.20
	P2	0.68	0.82	31.80	0.55	0.74	15.41
	P3	0.68	0.82	31.87	0.53	0.73	17.81
	P4	0.67	0.82	31.31	0.52	0.70	13.44
	P5	0.68	0.82	32.20	0.55	0.73	17.76

Table 9: Translation scores across prompts (metaphor only). BS = BERTScore; CM = COMET; BL = BLEU.

		BERTScore					COMET					BLEU				
Type	**Model**	P1	ΔP2	ΔP3	ΔP4	ΔP5	P1	ΔP2	ΔP3	ΔP4	ΔP5	P1	ΔP2	ΔP3	ΔP4	ΔP5
Conv-CMC	Qwen-3B	**.638**	0.020	0.009	0.018	0.023	**.810**	0.015	0.013	0.010	0.014	**24.58**	1.73	−0.79	1.32	0.20
	Qwen-7B	**.679**	−0.010	−0.015	0.006	−0.005	**.834**	−0.001	−0.005	0.002	0.003	**30.48**	−1.49	−0.74	1.58	−0.51
	DeepSeek	**.747**	0.002	−0.002	0.008	0.003	**.864**	−0.002	−0.004	0.002	0.001	**39.53**	−1.28	−1.04	0.40	−0.54
	GPT-5	**.739**	−0.036	−0.041	−0.019	−0.031	**.858**	−0.011	−0.016	−0.007	−0.012	**37.47**	−5.78	−6.06	−4.03	−5.88
Conv-CMDAG	Qwen-3B	**.514**	0.016	0.014	0.027	0.038	**.744**	0.008	0.010	0.009	0.013	**14.88**	1.11	1.87	2.78	2.83
	Qwen-7B	**.561**	0.021	0.008	0.018	0.027	**.765**	0.017	0.009	0.014	0.021	**18.35**	0.47	0.60	1.04	−0.49
	DeepSeek	**.624**	0.015	0.012	0.002	0.015	**.796**	0.009	0.003	0.002	0.011	**23.87**	2.96	2.23	0.50	1.25
	GPT-5	**.627**	0.022	0.016	0.022	0.025	**.795**	0.010	0.007	0.014	0.011	**24.74**	3.38	4.34	1.17	3.18
Conv-PNL	Qwen-3B	**.608**	−0.008	−0.025	0.000	−0.012	**.766**	0.000	−0.006	0.006	−0.005	**24.94**	−1.79	−2.48	−1.27	−1.78
	Qwen-7B	**.611**	−0.018	−0.005	−0.002	−0.019	**.778**	−0.009	−0.001	−0.005	−0.008	**25.48**	−1.64	0.30	−0.67	−1.07
	DeepSeek	**.682**	0.002	0.014	0.012	0.015	**.809**	0.003	0.005	0.003	0.006	**34.83**	−0.10	1.21	−0.09	1.90
	GPT-5	**.701**	−0.056	−0.024	−0.001	−0.029	**.817**	−0.025	−0.012	0.000	−0.010	**37.90**	−8.15	−4.02	−0.37	−5.62
Conv-Idioms	Qwen-3B	**.488**	0.001	0.017	0.009	0.019	**.706**	0.017	0.006	0.015	0.030	**14.22**	−0.22	−2.27	−0.36	0.05
	Qwen-7B	**.534**	0.012	0.004	−0.014	−0.002	**.738**	0.018	0.010	0.006	0.009	**21.77**	−0.65	−4.11	−3.94	−1.95
	DeepSeek	**.587**	0.000	−0.010	−0.020	−0.004	**.771**	0.005	−0.004	−0.017	−0.004	**22.84**	0.13	−1.01	−1.70	0.86
	GPT-5	**.571**	−0.026	−0.010	−0.025	−0.016	**.773**	−0.018	−0.012	−0.009	−0.004	**21.26**	−1.47	−0.22	−1.85	−0.35
Novel-Slang	Qwen-3B	**.417**	0.000	0.004	0.027	0.003	**.660**	−0.002	0.004	0.016	0.015	**6.83**	1.89	0.58	0.81	1.00
	Qwen-7B	**.428**	−0.015	0.010	−0.005	−0.003	**.670**	0.002	0.003	0.006	0.013	**9.58**	−1.48	−0.92	−4.23	0.71
	DeepSeek	**.526**	0.024	0.003	−0.009	0.021	**.724**	0.017	0.003	−0.019	0.006	**17.21**	−1.79	0.61	−3.77	0.56
	GPT-5	**.537**	−0.016	0.018	0.022	−0.007	**.727**	−0.012	0.009	0.017	0.014	**15.50**	1.66	3.77	1.08	0.76
Non-Metaphorical	Qwen-3B	**.574**	−0.005	−0.013	−0.007	–	**.822**	−0.002	−0.005	−0.003	–	**20.05**	−1.39	−1.61	−1.57	–
	Qwen-7B	**.593**	−0.005	−0.020	−0.012	–	**.833**	−0.004	−0.008	−0.006	–	**21.56**	0.25	−0.48	−1.09	–
	DeepSeek	**.638**	−0.003	−0.005	−0.009	–	**.847**	−0.001	−0.002	−0.005	–	**28.57**	−0.69	−1.72	−1.88	–
	GPT-5	**.616**	−0.012	−0.001	−0.001	–	**.842**	−0.005	−0.006	−0.004	–	**24.37**	−1.80	−1.68	−0.29	–

Table 10: Translation quality scores by model and metaphor type (Prompt 1 baseline; Prompts 2–5 shown as Δ from Prompt 1; BERTScore and COMET: mean; BLEU: corpus-level).