

A Rising Tide Lifts All Boats: MTQE Rewards for Idioms Improve General Translation Quality

Ishika Agarwal^{*1}, Zhenlin He^{*1}, Dhruva Patil², Dilek Hakkani-Tür¹

¹University of Illinois, Urbana Champaign, ²Independent

{ishikaa2, zhenlin5, dilek}@illinois.edu, dhruvakpatil@gmail.com

Abstract

Non-compositional expressions (e.g., idioms, proverbs, and metaphors) pose significant challenges for neural machine translation systems because their meanings cannot be derived from individual words alone. These expressions encode rich, cultural meaning, and have both figurative and literal meanings, making accurate translation difficult. Because models are fairly good at translating compositional text, we investigate GRPO-style fine-tuning using Machine Translation Quality Estimation (MTQE) models as reward functions to train models to better translate idioms. Using Chinese and Hindi idiom datasets, we find that idiom translation abilities improve by ~ 16 points¹, general, non-idiomatic translation implicitly improves by ~ 11 points, and cross-lingual translation abilities (trained on one language, evaluated on another) improves by ~ 9 points. Overall, our work quantifies the non-compositional translation gap and offers insights for developing LLMs with stronger cross-cultural and figurative language understanding.

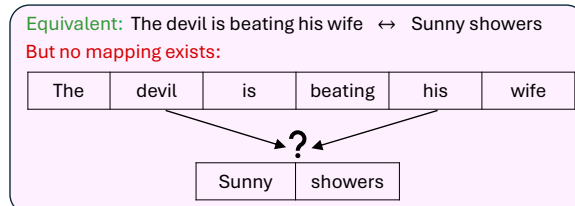
1 Introduction

The focus on multilingual language modeling has grown significantly (Adelnia and vahid dastjerdi, 2011; Cheng and Bhat, 2024; Ramos et al., 2026; Wu et al., 2018) because of how accessible LLMs become when they can converse in different languages. However, multilingual language modeling is difficult, due to alignment based problems. Broadly, these problems can be categorized into *language specific knowledge* where models contain different facts about the same topic in different languages (Agarwal et al., 2025; Jin et al., 2025), *non-isomorphic* phrases or words that do

^{*}These authors contributed equally to this work. Correspondence to ishikaa2@illinois.edu.

¹This is an average improvement across all metrics measuring n-gram similarity and semantic similarity. More details are available in Section 4.

1. Meaning cannot be derived from words



2. Idioms cannot be literally translated

Idiom	पानी-पानी होना
Literal translation	To become water-water
Semantic translation	To feel embarrassed
Equivalent idiom	Egg on your face

Not the same: To become water-water ↯ egg on your face

3. Idiom meanings change with context

Figurative Meaning	Literal Meaning
"Their lives were <i>hanging by a thread</i> "	"The iron ball was <i>hanging by a thread</i> ."
↓	↓
"Their lives were <i>in danger</i> ."	"The iron ball was <i>hanging by a thread</i> ."

Figure 1: There are three challenges when modeling and translating non-compositional phrases.

not have direct translations in other languages (like the word "jugaad" in Hindi) (Wu et al., 2024a; Meng et al., 2025), and *non-compositional* phrases whose meaning cannot be derived from individual words) (Zhou and Bhat, 2024). The focus of this work is on improving the translation abilities of non-compositional sentences.

Specifically, translating non-compositional phrases poses three challenges, as illustrated in Figure 1. **Firstly**, their meanings cannot be derived from the individual, constituent words (Zhou et al., 2023). Moreover, they compress historical stories and cultural assumptions, which make their literal translations and intended meanings differ

greatly.

Second, many idioms do not have semantically equivalent translations: language models tend to paraphrase text during translation, and thus show a strong literal translation bias towards non-compositional phrases (Adelnia and vahid dast-jerdi, 2011). These literal decodings erase the idiom’s figurative intent and leads to incorrect, misleading translations.

Third, and finally, non-compositional phrases are highly context dependent and can have different meanings (De Luca Fornaciari et al., 2024a). These issues result in subpar translation quality, and highlight a broader cultural translation gap that current LLMs are not equipped to bridge without specialized training.

In this work, we propose to improve the translation quality of language models in the form of a training-free solution, and training-based solution. The training-free solution is a simple, three-step prompting pipeline that encourages the model to think about the cultural context and semantic meaning of an idiom before translating it.

The training-based solution uses GRPO-style fine-tuning to encourage models to translate idioms semantically. In particular, we show the efficacy of using MTQE (Machine Translation Quality Estimation) models, such as COMET (Rei et al., 2020, 2022, 2023), as a reward model for fine-tuning LLMs with GRPO (Group Relative Policy Optimization) (Shao et al., 2024). MTQE models are trained on human preference data, which contains implicit signals of non-compositional phrases. We use MTQE rewards as a form of distillation to train LLMs for better translation quality. All methods are outlined in Section 3.

We use Chinese and Hindi idioms from existing datasets (outlined in Section 3.1) to evaluate our methods. We evaluate on small language models (Qwen’s 3B model and Llama’s 8B model) to show that we can improve small language models for cheap and accessible translation. We measure translation quality across various dimensions of semantic and n-gram based similarity, and find that by using MTQE rewards during GRPO fine-tuning on idiomatic data:

1. Idiomatic translation quality improves by ~ 16 points (Figures 3 and 4),
2. Non-idiomatic, general, translation quality improves by ~ 11 points (Figures 5 and 6), and

3. Cross-lingual semantic representation (idiom training in one language transfers to another) improves by ~ 9 points (Figures 7 and 8).

2 Background

2.1 MTQE Models

MTQE models estimate the quality of machine translated text (Rei et al., 2020). There are two kinds: reference-free and reference-based (Rei et al., 2022, 2023).

Reference-free models are given a source text and a translated text as input. Their output is a scalar score between 0 to 1 that indicates the semantic equivalence between the source and translation (the closer it is to 1, the stronger the semantic equivalence) (Zhao et al., 2024). Thus, during training, sentences that are semantically closer will output higher scores and those that are semantically less equivalent will receive lower scores.

Reference-based models, on the other hand, are given a source text, translated text, and reference text. Their output is also a scalar between 0 to 1, and also indicate the semantic equivalence between source, translation, and reference. They rely on gold standard, human-annotated references off of which to base the numerical MTQE scores. During training, the model learns to output direct assessment scores to the given source, translated, and reference text (Rei et al., 2022).

We posit that due to MTQE models being trained on parallelly translated and/or human annotated data, they autonomously learn how to model non-compositional language. We can use such models as a form of weak distillation to teach models to translate non-compositional language effectively.

2.2 Idiom Translation

Idiom and proverb detection and generation have been studied extensively in literature (Cheng and Bhat, 2024; Lai and Nissim, 2024). Some works focus on detection and generation (Wu et al., 2024b; Zhou et al., 2023; He et al., 2024; De Luca Fornaciari et al., 2024b), while others work on creating high-quality curated datasets (Zeng et al., 2023; Rezaeimanesh et al., 2025; Tedeschi et al., 2022; Haagsma et al., 2020).

While detection and generation abilities are still improving, translation remains a problem (Cheng

and Bhat, 2024). Sources show that while proprietary, closed-source language models are able to translate idioms well, open-source language models have not yet reached SOTA results (Obeidat et al., 2024). Kim et al. (2025) suggests language models do indeed know the semantic meaning behind idioms, they just need to be extracted properly. Previous work has mostly used prompting methods (Gao et al., 2025; Rafatbakhsh et al., 2021). Rather than default to translating literal meanings (Rezaeimanesh et al., 2025), language models must be fine-tuned for translating semantic meanings behind idioms. Recent literature suggests that reinforcement learning can sharpen output distributions towards a specific task (Yue et al., 2025). *We borrow this finding and use GRPO to improve a language model’s ability to translate non-compositional language.*

3 Improving Non-Compositional Translation

3.1 Dataset Creation

Our evaluation spans two languages: Chinese and Hindi. We release our datasets, along with all of our code [here](#). The same training and testing splits are used for all baselines, which allows for direct comparison between all methods.

Chinese. We use the PETCI dataset (Tang, 2022) for Chinese–English idiom translation. We first run a preprocessing script on the original release that trims stray whitespace and removes rows where either the Chinese idiom or the English translation is empty, contains only whitespace, or is marked as missing. We extract 1,623 Chinese–English Idiom pairs use 1,000 for training and 623 for testing. We emphasize that this preprocessing defines the subset used in our experiments, not intended as a quality estimator of the remaining PETCI entries.

Hindi. We compile a dataset from multiple sources, combining mined Hindi–English idioms from the OPUS OpenSubtitles corpus (Zhang et al., 2020) with additional high-quality synthetic pairs generated using GPT-5. We filter this collection by removing duplicates, discarding entries with incomplete or overly literal translations, and manually validating idioms that appear ambiguous or context-dependent. After cleaning, we select 1,000 valid Hindi–English idiom pairs, which we split deterministically into 800 training examples and 200

test examples.

3.2 Training-Based

We explore using GRPO (Shao et al., 2024) for fine-tuning models to improve an LLM’s translation abilities. In particular, the reward model is an MTQE model, and the LLM is rewarded based on how effective translations are based on the ground truth. As mentioned before, there are two kinds of MTQE models: reference-free and reference-based. Within these variations, we contrive different settings for rewards (see Figure 2):

1. **QE-Positive.** QE models take in as input the source text (`src`) and the machine translated text (`mt`). In this setting, `src` is the input idiom in a different language, and `mt` is the LLM generated translation during training. This reward encourages LLMs to generate translations that are semantically equivalent to the original input idiom. We denote this as $QE_{pos}(\text{idiom}, \text{mt})$.
2. **QE-Negative.** Here, `src` is the literal meaning of the idiom (in English) and `mt` is, again, the generated translation during training. Note: both `src` and `mt` are in English in this setting. This is a creative misuse of MTQE. Although it is expected to have crosslingual outputs, it is ultimately a semantic similarity metric. In this case, we apply a negative reward for the MTQE between `src` and `mt`, encouraging the LLM to generate translations that are not semantically equivalent to the literal translations. We denote this as $QE_{neg}(\text{literal}, \text{mt})$.
3. **QE-Constrained.** The issue in the QE-Positive setting is that bad translations do not get *discouraged*. The issue in the QE-Negative setting is that some idioms can be correctly literally translated in another language (for example: "plenty of fish in the sea" in English and "Hay más peces en el mar" in Spanish are correct, literal translations of each other). To bridge these gaps, we test a third setting. We simply assign the joint reward: $QE_{pos}(\text{idiom}, \text{mt}) - QE_{neg}(\text{literal}, \text{mt})$. This helps encourage the LLM to semantically translate the idiom, and discourage the LLM to literally translate.
4. **QE-DA.** This setting follows the **QE-Positive** setup, where `src` is the source idiom, and `mt`

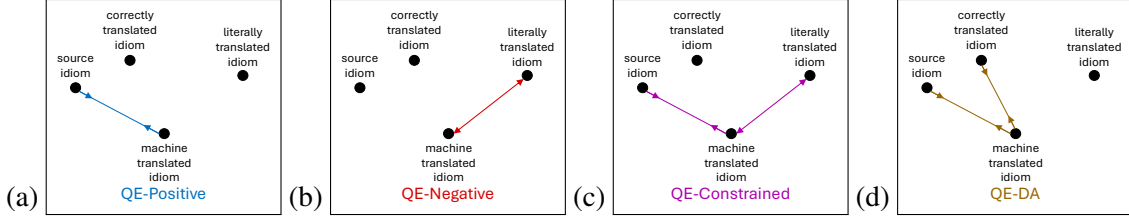


Figure 2: Illustration of the distinction between all three GRPO-QE-* methods. (a) **QE-Positive** pulls semantically equivalent texts closer. (b) **QE-Negative** pulls semantically inequivalent texts apart. (c) **QE-Constrained** balances both. (d) **QE-DA** uses a ground-truth reference translation to inform MTQE.

is the machine translated idiom, but it also receives a ref, which is the ground truth translated idiom. This reward will encourage models to translate idioms towards a specific target translation.

3.3 Training-Free Structured Prompting

We also develop a training-free idiom translation method designed to reduce literal-translation bias with just prompt engineering. The core idea is to dissect idiom translation into three explicit reasoning stages:

1. **Idiomatic Explanation.** The model explains the idiom’s figurative meaning in English, emphasizing cultural meaning over surface semantics, denoted by E (prompt in Figure 9, Appendix A).
2. **Literal Semantics.** The model provides a word-by-word literal translation of the idiom, yielding L , which helps to disentangle literal and figurative interpretations (prompt in Figure 10, Appendix A).
3. **Natural Idiomatic Translation.** Given both E and L , the model produces a single fluent English expression that captures the idiomatic sense (prompt in Figure 11, Appendix A).

Unlike decoding-based reranking or reinforcement learning, this method introduces no additional optimization, instead relies entirely on prompt engineering and explicit reasoning decomposition. This structured prompting approach aims to mitigate the literalism observed in the raw model.

4 Experiments

4.1 Setup

Models. To show the efficacy of our method, we use two small language

models: (1) Qwen/Qwen2.5-3B (shorthand: Qwen) (Yang et al., 2025), and (2) meta-llama/Llama-3.1-8B (shorthand: Llama) (Grattafiori et al., 2024). We employ small LMs to showcase the efficacy of our method compared to larger LMs. For all model inference, we use a temperature of 0.3. For SFT, we train models for 3 epochs. For GRPO, we use the ver1 library (Sheng et al., 2024), with 4 completions per group, and 5 epochs. All other settings can be found in our [Github repository](#).

Metrics. Our evaluation contains a variety of metrics to test different aspects of the translation quality.

1. We report the scores of the same MTQE models that we use as reward models for GRPO, to show the final reward of the tuned models. **DA** are the Direct Assessment scores (specifically, the Unbabel/wmt22-comet-da model), and **QE** are the Quality Estimation scores (specifically, the Unbabel/wmt22-cometkiwi-da model).
2. We use other MTQE models to evaluate performance: specifically "Unbabel/XCOMET-XL" (DA, denoted by **DA_XL**), "Unbabel/XCOMET-XXL" (DA, denoted by **DA_XXL**), "Unbabel/wmt23-cometkiwi-da-xl" (QE, denoted by **QE_23_XL**), and "Unbabel/wmt23-cometkiwi-da-xxl" (QE, denoted by **QE_23_XXL**).
3. To evaluate the n-gram similarity between the predicted translation and the ground truth translation, we use **ROUGE**.
4. Because ROUGE is too rigid a metric (it measures the n-gram overlap ration and does not measure for semantic similarity), we use

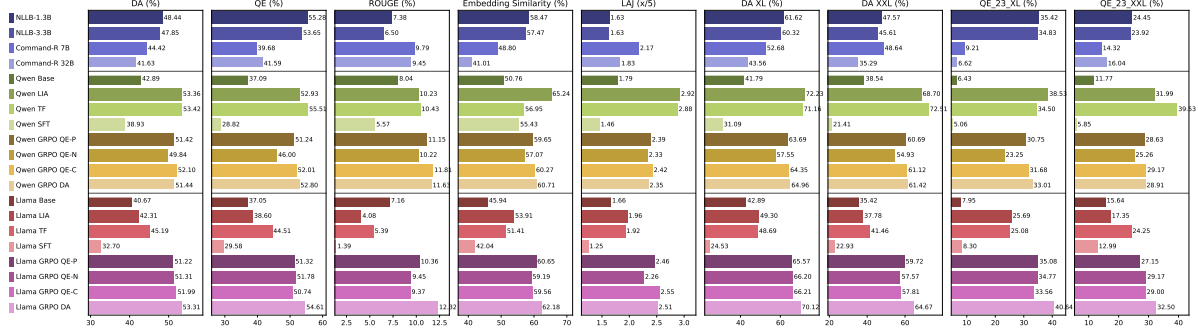


Figure 3: Evaluation of translation abilities of **Chinese idioms**. Here, we see that the LIA and TrainingFree (TF) baselines do well on Qwen-2.5-3B, but not on Llama-3.1-8B (hence is unreliable). The GRPO based methods are not only **performant**, but also **reliable**.

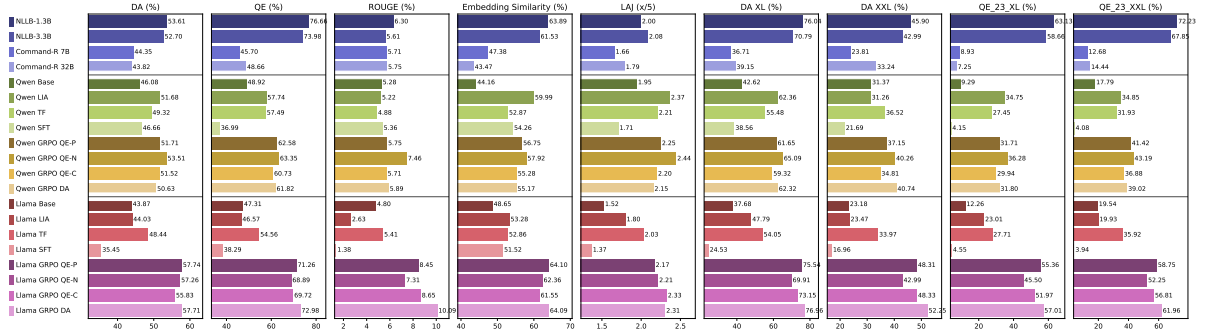


Figure 4: Evaluation of translation abilities of **Hindi idioms**. Here, we see that the core translation models (NLLB and Command-R) translate Hindi idioms well, but not Chinese idioms (Fig. 3). The GRPO based methods are not only **performant**, but also **reliable**.

embedding models to embed the semantic similarity of the predicted translation and ground truth translation; we, then, report the cosine similarity between these embeddings (i.e., **Embedding Similarity**).

- Finally, to make our evaluation comprehensive, we employ an LLM-as-a-Judge (abbreviated to **LAJ**) to score the semantic similarity between the predicted and ground truth translation. The particular LLM-as-a-Judge we use is the Prometheus-7v-V2.0 model (Kim et al., 2024).

Baselines We employ a variety of baseline methods. First, we use core translation models: sequence-to-sequence (NLLB (Team et al., 2022)) and autoregressive (Command-R (Cohere et al., 2025)). To compare our training-free method, we use the LIA (Language Model Based Idiom Alignment) (Donthi et al., 2025)². To compare our

²The same paper also proposes SIA (Semantic Idiom Alignment), but their setting requires an external database of idioms to retrieve from, which is then fed into SIA to select highly matching translations. We do not choose to compare

training-based method, we fine-tune models using supervised fine-tuning (SFT) using a cross-entropy loss.

Non-Idiomatic Performance In order to ensure our idiom-specific methods (especially, the training-based) do not deteriorate in translating non-idiomatic text, we measure the performance of our methods on non-idiomatic text translation. We use the same metrics as mentioned above, and we employ the Opus-100 (Zhang et al., 2020) dataset. This dataset contains pairs of source and machine translated text pairs. We use the en-zh and en-hi pairs from the dataset, and use 400 randomly selected data points from the test splits of the Opus dataset for our evaluation.

4.2 Results and Analysis

Idiom Translation. Figures 3 and 4 contain the evaluation results on Chinese and Hindi idioms, respectively. The NLLB and Command-R models tend to be better at the Qwen and Llama base

against this particular baseline as we do not assume access to an external database.

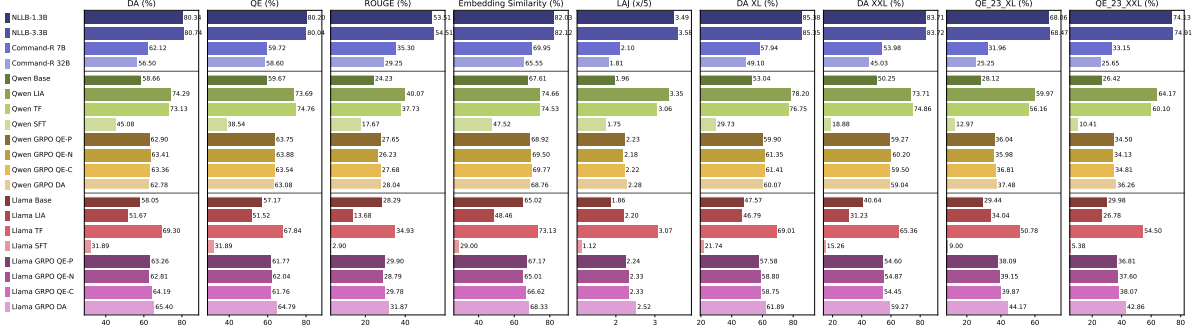


Figure 5: Evaluation of translation abilities of **regular Chinese sentences**. Here, it is shown that performance does not deteriorate when models are trained on idiomatic data.

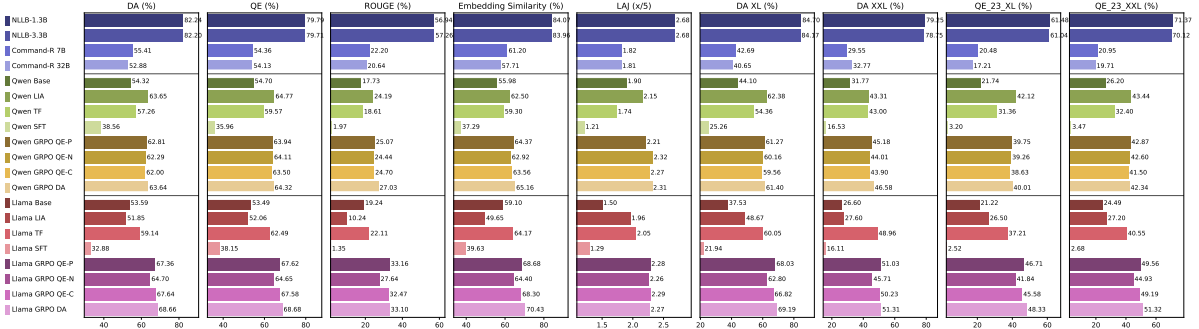


Figure 6: Evaluation of translation abilities of **regular Hindi sentences**. Here, it shows that performance does not deteriorate when models are trained on idiomatic data.

models: they score an average of 6.23 (Qwen) and 7.03 (Llama) points³ higher for Chinese translation, and 10.15 (Qwen) and 12.02 (Llama) points higher for Hindi translation. Still, the results show that *training-free/-based methods can improve flexible language models so they can perform on-par with translation models that have dedicated machine translation or parallelly translated datasets*⁴.

Comparing the training-free methods (LIA and TrainingFree (denoted by TF in the figures)), we see that Qwen performs better with LIA (LIA is 1.4 points better than TrainingFree), while

³To aggregate these metrics into one reportable score, we take the average of all 9 metrics: DA, QE, ROUGE, Embedding Similarity, LAJ, DA_XL, DA_XXL, QE_23_XL, QE_23_XXL. All are on the same scale, while LAJ is on a scale of 1-5. To calibrate them all on the same scale, we multiply the LAJ score by 20. Thus, to calculate the performance of a particular method, we use the following formula: $p = (DA + QE + ROUGE + EmbeddingSimilarity + 20 * LAJ + DA_XL + DA_XXL + QE_23_XL + QE_23_XXL) / 9$. To find the difference in performance between baseline A and method B , we calculate $p_A - p_B$.

⁴According to the model cards (Yang et al., 2025; Grattafiori et al., 2024), the training sets have multilingual data, but the data is not necessarily parallelly translated in other language. These multilingual data are generally for multilingual question-answering, reasoning, and knowledge tasks.

Llama performs better with TrainingFree (it is 4.1 points better than LIA). These inconsistent results show that prompting is not a reliable method, motivating the need for fine-tuning.

Supervised fine-tuning the models consistently cause models to perform worse than even the base models – this is because the models are being trained for outputting particular sentences, rather than understanding semantic meaning of sentences. Thus, *using RL is our final approach, which significantly outperforms the base and SFT models*: The **QE-Positive** models outperform the base and SFT models by 16.7 and 22.8 points, respectively; the **QE-Negative** models outperform them by 15.3 and 21.4 points, respectively; the **QE-Constrained** models outperform them by 16.1 and 22.2 points, respectively; finally, the **QE-DA** models outperform them by 18.1 and 24.2 points, respectively. **Overall, the RL models bring an absolute point improvement of 16.54 points in idiom-translation ability.**

Human Evaluation. To verify that the improvements from MTQE-rewarded GRPO are not solely due to reward hacking, we conduct a small-scale human evaluation on 20 randomly selected trans-

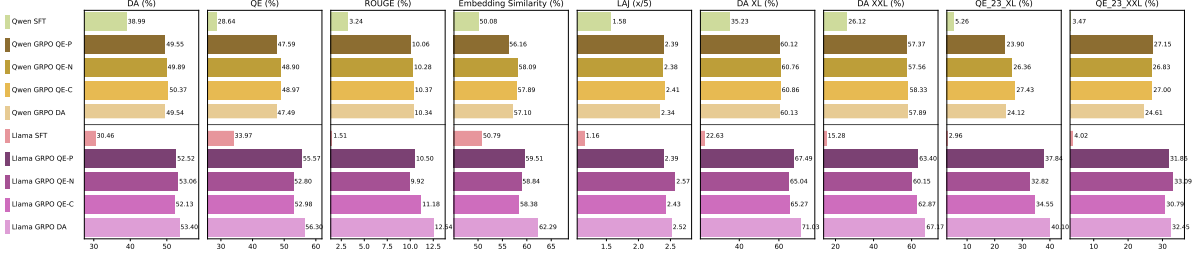


Figure 7: Evaluation of **Hindi-trained models translating Chinese idioms**. Per Fig. 3, Qwen-2.5-3B achieved DA: 42.89, QE: 37.09, ROUGE: 8.04, ES: 50.76, LAJ: 1.79; Llama-3.1-8B achieved DA: 40.67, QE: 37.05, ROUGE: 7.16, ES: 45.94, LAJ: 1.66. GRPO models outperform base models.

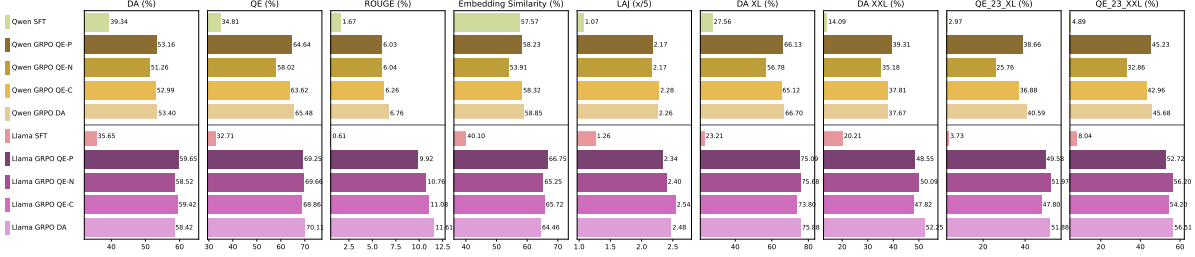


Figure 8: Evaluation of **Chinese-trained models translating Hindi idioms**. Per Fig. 4, Qwen-2.5-3B achieved DA: 46.08, QE: 48.92, ROUGE: 5.28, ES: 44.16, LAJ: 1.95; Llama-3.1-8B achieved DA: 43.87, QE: 47.31, ROUGE: 4.80, ES: 48.65, LAJ: 1.52. GRPO methods outperform base models.

Setting	Chinese	Hindi
NLLB-1.3B	1.4	2.3
Qwen Base	3.2	2.1
Qwen GRPO QE-P	3.5	3.8
Llama Base	2.3	2.8
Llama GRPO QE-P	2.8	3.6

Table 1: Small-scale human evaluation of idiom translation quality. Scores are averaged over 20 randomly selected translations and use a 1–5 Likert scale, where higher scores indicate more semantically accurate and natural translations.

lations for each language and model setting. We compare NLLB-1.3B, the Qwen and Llama base models, and the corresponding GRPO QE-Positive models. Each translation is rated on a 1–5 Likert scale, ranging from a nonsensical or malformed output (1) to a perfect semantic translation (5). As shown in Table 1, GRPO improves the average human rating from 3.2 to 3.5 for Qwen and from 2.3 to 2.8 for Llama on Chinese, and from 2.1 to 3.8 for Qwen and from 2.8 to 3.6 for Llama on Hindi. These results provide complementary human evidence that the gains observed across automatic metrics, including embedding similarity and MTQE-based metrics, reflect improved semantic idiom translation rather than simply reward optimization.

Performance on non-idiomatic sentence translation. Figures 5 and 6 show the results of translation abilities of non-idiomatic text translation in Chinese and Hindi respectively. These tables show that **the performance on general, mostly non-idiomatic text translation not only maintains, but also improves with the GRPO models, compared to base (average of 10.9 points) and SFT (average of 29.7 points)**. Of course, the translation models are better performing (the best translation model is 18.9 points better than the best GRPO model). We hypothesize that the GRPO-tuned language models have improved semantic representation across languages, so they could ultimately outperform the translation models on reasoning, knowledge, and question-answering tasks.

Effects of language-specific training. Figures 7 and 8 contain the evaluation results for the transfer settings. Comparing the results from Figures 4 and 8 versus Figures 3 and 7, there do exist significant transferability capabilities between the trained models. In particular, models trained on Chinese idioms but evaluated on Hindi idioms perform 13.1(Qwen)/24.0(Llama) points *better* than base models and 0.8(Qwen)/0.8(Llama) points *better* than models trained on Hindi idioms. On the contrary, models trained on Hindi idioms but evaluated on Chinese idioms perform

12.3(Qwen)/17.8(Llama) points *better* than base models, Qwen performs -1.8 *worse* than models trained on Chinese, and Llama performs 1.2 points *better*. The overall boost of 8.53 absolute points in translation improvement shows that **training on one language does not hinder the performance on other languages**.

Effect of reward model. Even though we had broken down the rewards into four, **there are no noticeable effects of each reward**. This is encouraging because they all have different kinds of supervision. **QE-Positive** requires just an input idiom, which can be extracted from large corpora with idiom detectors. **QE-Negative** and **QE-Constrained** requires a source *and* a literal translation, which might be a bit expensive to get. Although, costs can be avoided by using a small enough language model that reliably generates literal translations. **QE-DA**, however, is the most expensive as it requires a ground truth translation of idioms – obtaining this is expensive, as it necessitates annotation from either humans or closed-source large language models. **Since the least expensive reward QE-Positive does not lag far behind the most accurate reward QE-DA, we are able to showcase the robustness of using MTQE rewards.**

4.3 Performance v.s. Compute Tradeoff

While performant, the GRPO training inherent in our method is the most time consuming out of the baseline methods. Using 4 H100 NVIDIA GPUs, it takes around 6-12 hours; in total, training takes 24-to-48-GPU hours. In comparison, SFT takes just 45 minutes. Neither the training-free methods nor the translation models (NLLB and Command-R) incur training costs (although, creating them incurs massive training and data curation costs). In the tradeoff of performance versus compute, we argue that our method still wins because it is the most performant, most generalizable, and most reliable method out of the baselines.

5 Conclusion

In this work, we explore structured approaches to improving non-compositional language translation. As our main contribution, we present GRPO-based fine-tuning using MTQE models as reward models. Our experimentation uncovers three results: (1) idiom translation abilities increase by an average of 16.5 absolute points over base models across languages and architectures, (2) non-

idiomatic translation abilities are implicitly improved by 10.9 absolute points, and (3) cross-lingual translation abilities are also improved by 8.5 absolute points. These results show that MTQE rewards effectively distill their multilingual language embedding capabilities to language models: they improve semantic relationships in multilingual language models, are reliable to not compromise current language model abilities, and enables models to generalize to other languages meaningfully. These results encourage future work to understanding the complementary nature of mapping semantics in various languages for improving multilingual language modeling.

6 Limitations

While MTQE models can handle a broad variety of languages and have shown to be aligned with human preferences, our method is upper-bounded by the performance of MTQE models – the GRPO-trained translation models can only be as good as the MTQE models are. Plus, MTQE models also require many parallel data to be trained. Furthermore, RL is an expensive algorithm in general – even after choosing small model sizes and small datasets, training took around 6-12 hours to train on 4 NVIDIA H100s, which is not accessible. Future work involves understanding how to reduce the computational cost to train better multilingual models. Furthermore, we focused on high resource languages because low resource languages would require another axis of improvement to also improve the low-representation of low-resource languages within the dataset corpus. Finally, we do not study in this paper the case of models incorrectly interpreting figurative language in unsafe or harmful ways.

Acknowledgments

We are very grateful for Preeti Gomathinayagam, Zhihao Cheng, and Zhuchen Shao for their efforts in the early stages of this project.

References

- Amineh Adelnia and Hossein vahid dastjerdi. 2011. [Translation of idioms: A hard task for the translator](#). *Theory and Practice in Language Studies*, 1:879–883.
- Ishika Agarwal, Nimet Beyza Bozdog, and Dilek Hakkani-Tür. 2025. [Language specific knowl-](#)

- edge: Do models know better in x than in english? *Preprint*, arXiv:2505.14990.
- Kellen Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pre-trained language models. Association for Computational Linguistics.
- Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammam, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. *Command a: An enterprise-ready large language model*. *Preprint*, arXiv:2504.00698.
- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024a. *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.
- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024b. *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.
- 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.
- Hui Gao, Jing Zhang, Peng Zhang, and Chang Yang. 2025. *Consistency rating of semantic transparency: an evaluation method for metaphor competence in idiom understanding tasks*. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. *MAGPIE: A large corpus of potentially idiomatic expressions*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.
- Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. *Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.
- Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. *Language model alignment in multilingual trolley problems*. In *The Thirteenth International Conference on Learning Representations*.
- Jisu Kim, Youngwoo Shin, Uji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. *Memorization or reasoning? exploring the idiom understanding of LLMs*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710, Suzhou, China. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. *Prometheus 2: An open source language model specialized in evaluating other language models*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Huiyuan Lai and Malvina Nissim. 2024. *A survey on automatic generation of figurative language: From rule-based systems to large language models*. *ACM Comput. Surv.*, 56(10).
- Lei Meng, Yinlin Li, Wei Wei, and Caipei Yang. 2025. Resolving linguistic asymmetry: Forging symmetric multilingual embeddings through asymmetric contrastive and curriculum learning. *Symmetry*, 17(9):1386.
- Mohammed M Obeidat, Ahmad S Haider, Sausan Abu Tair, and Yousef Sahari. 2024. Analyzing the performance of gemini, chatgpt, and google translate in rendering english idioms into arabic. *FWU Journal of Social Sciences*, 18(4).
- Elaheh Rafatbakhsh, Alireza Ahmadi, Amirsaeid Moloodi, and Saeed Mehrpour. 2021. Development and validation of an automatic item generation system for english idioms. *Educational Measurement: Issues and Practice*, 40(2):49–59.
- Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2026. *Fine-grained reward optimization for machine translation using error severity mappings*. *Transactions of the Association for Computational Linguistics*, 14:733–754.

- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Sara Rezaeiemanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. [Large language models for Persian-English idiom translation](#). 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 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *arXiv preprint arXiv:2202.09509*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- 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.
- Di Wu, Yibin Lei, Andrew Yates, and Christof Monz. 2024a. Representational isomorphism and alignment of multilingual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14074–14085.
- Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. [A study of reinforcement learning for neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.
- Mingmin Wu, Guixin Su, Yongcheng Zhang, Zhongqiang Huang, and Ying Sha. 2024b. Refining idioms semantics comprehension via contrastive learning and cross-attention. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13785–13795.
- 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. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Ziheng Zeng, Kellen Tan Cheng, Srihari Venkat Nan-niyur, Jianing Zhou, and Suma Bhat. 2023. [IEKG: A commonsense knowledge graph for idiomatic expressions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. [Improving massively multilingual neural machine translation and zero-shot translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.
- Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. [From handcrafted features to llms: A brief survey for machine translation quality estimation](#). *Preprint*, arXiv:2403.14118.
- Jianing Zhou and Suma Bhat. 2024. Non-compositional expression generation and its continual learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2828–2839.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2023. [Non-compositional expression generation based on curriculum learning and continual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4320–4335, Singapore. Association for Computational Linguistics.

A Prompts

Tables 9, 10, and 11 contain the prompts for the TrainingFree method described in Section 3.3.

B Licenses

All datasets are publicly available and do not contain any personally identifiable information. The PETCI dataset is under the Creative Commons 4.0 license, the Open Subtitles dataset is under the GNU General Public License, the NLLB and Command-R models are under the Creative Commons Attribution Non Commercial 4.0 license, the Qwen model is under the Apache 2.0 license, and the Llama model is under the Llama license.

C Human Annotators

In creating the dataset, we used human annotators to verify the validity of the dataset. The authors of this paper were appropriate enough to create this dataset as they knew either Chinese or Hindi, and they were unbiased in their dataset verification process. We wrote up guidelines to make sure the idiomatic data was not "too literal" or "overly ambiguous" of whether it had a figurative or literal meaning.

D Usage of AI Assistants

There were only two things AI Assistants were used for: (1) writing code for plotting figures (after processing the data ourselves, we described to ChatGPT the format of the figures that we wanted and asked it how to add a color scheme), and (2) revising small chunks of text like the title, the abstract, and caption figures (we had already written the title and abstract, but we instructed Claude to refine the writing by making it clearer and shorter). No AI Assistants were used for any of the other tasks: paper writing, code writing, results analysis, or ideation (or even, this section).

Idiomatic Explanation Prompt

Explain the meaning of the following {lang} idiom in English.

- Audience: educated readers; be concise (<= 2 sentences).
- Do not translate word-by-word; provide the ****idiomatic sense****.

Idiom: {idiom}

Figure 9: Prompt to elicit Idiomatic Explanation, which is Step 1 in our training-free, structured prompting approach outlined in section 3.3.

Literal Semantics Prompt

Provide a ****literal, word-by-word**** English translation for the following {lang} idiom.

- Keep it terse and faithful to each component.
- No commentary, just the literal gloss.

Idiom: {idiom}

Figure 10: Prompt to elicit Literal Semantics, which is Step 2 in our training-free, structured prompting approach outlined in section 3.3.

Natural Idiomatic Translation Prompt

Produce a ****natural English idiomatic translation**** given:

- (1) An idiom explanation (idiomatic meaning) and
- (2) A literal word-by-word gloss.

Rules:

- Output a single short English phrase/sentence that a native speaker would actually say.
- Prefer clarity and naturalness over literalness.
- No extra commentary.

Idiom: {idiom}
Explanation: {explanation}
Literal: {literal}
Result:

Figure 11: Prompt to elicit Natural Idiomatic Translation, which is Step 3 in our training-free, structured prompting approach outlined in section 3.3.