# The Devil's in the Translation: Active Learning for Improving Idiom Translation Quality Estimation

**Ishika Agarwal**
UIUC
ishikaa2@illinois.edu

**Dhruva Patil** and **Najmeh Sadoughi** and **Zhu Liu**
Amazon Prime Video
{dhrpatil, nnourab, zhuliu}@amazon.com

## Abstract

Language models face a significant challenge of accurately assessing translating quality of idiomatic expressions. Expressions like "under the weather","let the cat out of the bag", and "let sleeping dogs lie" pose a problem for Neural Machine Translation systems as their meanings cannot be derived from the individual, constituent words. Their non-compositional nature allows them to have both figurative and literal meanings, depending on the context. Hence, due to their tendency to paraphrase translated text, language models tend to default to literal translations. Furthermore, research on idiomatic expression resolution is limited due to automatic methods to quantify these translation errors. Unfortunately, due to the lack of naturally occurring idioms in training corpora, there are very few parallelly translated idioms. In this work, our contributions are three-fold: we (1) release a large dataset of parallelly translated idiomatic expressions, (2) show the limitations of state-of-the-art quality estimation models with idiomatic expressions, (3) show the efficacy of active learning in developing idiom-aware machine translation quality estimation models. Our results show that fewer, informative data are more effective than using the entire training set in two aspects: up to 21.58% more semantic translations are scored higher than literal translations, and the underlying model has not suffered from catastrophic forgetting as the base model capabilities are affected minimally. Our work contributes to the development of data-efficient language models that are more inclusive to various languages and their linguistic nuances.

## 1 Introduction

Idioms, and other non-compositional language, pose a problem for language models, especially in multilingual settings. The difficulty of this task is three-fold. First, their meanings cannot be derived from the individual, constituent words: the idiom "the devil is beating his wife" has no word-to-word semantic equivalence to its corresponding, disambiguated meaning "sunny showers" (Zhou et al., 2023). Second, many idioms do not have semantically equivalent translations: language models tend to paraphrase text during translation, which results in a literal and unnatural translation of idioms across languages

These above issues result in a subpar translation quality by current language models. Due to their tendency to paraphrase text during generation, language models tend to default to literally translating idioms (Rezaeimanesh et al., 2025). This is a crucial shortcoming in how languages are modeled. The effects can be seen in models beyond natural language generation. In particular, we show similar shortcomings in machine translation quality estimation (MTQE) models.

Unfortunately, due to the lack of naturally occurring idioms in training corpora, there are very few parallelly translated idioms. Hence, our proposed solution is to train idiom resolution models with *active learning*. Active learning assumes a pool of unlabeled data (in our case, English idioms without parallel resolved idioms in other languages) and access to an oracle that can provide labels (i.e., parallelly resolved idioms) (Settles, 2009). In order to avoid overusing the oracle, each query to the oracle incurs a cost. The goal is to learn a model that is as performant as possible while minimizing query costs.

**Definitions.** The following are definitions of the common terminology we use in this work.

- **Idiom**: a phrase or sentence that is a metaphor and cannot be understood literally (e.g., "under the weather")
- **Disambiguated Meaning**: the underlying meaning of the idiom, in the same language as the idiom (e.g., "feeling sick or ill")
- **Literal Translation**: an incorrect translation

of the idiom, word-for-word (e.g., "under the weather" and "मौसम के नीचे" / "debajo del clima")

- **Semantic Translation**: a correct translation of the idiom, either (1) a translation of the disambiguated meaning (e.g., "caer enferma") or (2) a culturally appropriate, equivalent idiom (e.g. "ठंड लगना")

In this work, our contributions are two-fold:

1. **MOTHERLODE**: we recognize the lack of large idiomatic datasets with parallelly translated idioms across languages. Hence, we augment two existing datasets (ID10M (Tedeschi et al., 2022) and MAGPIE (Haagsma et al., 2020)) with Claude 3.7-Sonnet. Our dataset (which we call MOTHERLODE) is the largest dataset with parallelly translation idioms, including 430k+ samples across 20k+ unique idiomatic expressions, with each unique idiom parallelly translated across 11 languages.

2. **AL4IDIOMS**: we propose to improve the reliability of machine translation quality estimation models by fine-tuning them to score semantic translations higher than literal translations. To improve the efficiency and reliability of our method (which we call AL4IDIOMS), we employ an active learning-based strategy to select the most informative data samples to fine-tuned MTQE models. This helps alleviate large computational costs as well as mitigates catastrophic forgetting. We show that with just 50% of the training budget, our approach achieves an average 8.83 point improvement in accuracy[1] over the baselines.

## 2 Related Works

Idiom and proverb detection and generation have been studied extensively in literature. Most works focus on English non-compositional phrase detection and generation (Cheng and Bhat, 2024; Li et al., 2024)

Idiom detection and generation have been broken into two broad solutions: rule-based and neural-based (Lai and Nissim, 2024). Rule-based methods rely on identifying linguistic patterns for idioms or rely on knowledge banks to identify and map idioms. Zeng et al. (2023) construct knowledge graphs to augment language models and show significant improvements. While they are simple and can provide interpretable outputs, they do not scale well and are limited in flexibility. On the other hand, neural-based models, while requiring large computational costs, are able to achieve state-of-the-art results and are able to generalize very well in zero/few shot settings.

However, controlling neural models is difficult, and other problems show up. For example, Stowe et al. (2022) release the IMPLI benchmark and additionally show that while models find it difficult to detect idioms, adding in only idiom-specific training data can also deteriorate performance. This highlights a need for research that aims to understand the underlying mechanisms to balance out idiom and general NLP feature representation. Zhou et al. (2023) design curriculum learning metrics to understand the difficulty of a data sample by looking at the representational distance of the idiom and the individual words in the idiom. **This work inspires our uncertainty-based data selection metric in AL4IDIOMS.**

Work in multilingual idiomatic representation is also two fold: contrastive learning algorithms and datasets (Wu et al., 2024). On the contrastive learning algorithm side, He et al. (2024) explore a contrastive triplet loss to pull together correct translations and pull apart negative (literal) translations. On the datasets side, Rezaeimanesh et al. (2025) release an English to Persian parallel idiom translation dataset. For our dataset contributions, we augment two existing datasets: ID10M (Tedeschi et al., 2022) and MAGPIE (Haagsma et al., 2020). ID10M is a dataset for idiom detection with idioms in 10 languages (Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, Spanish). The dataset contains non-idiomatic data, and does not guarantee parallel translations for a majority of the idiomatic data. MAGPIE is a monolingual dataset for English idioms, but does not contain non-idiomatic data. More details of these datasets are in Table 1. Because they are the largest datasets, we augment these datasets to contain multilingual, parallelly translated idiomatic data. To reiterate, we create the **largest dataset of only idiomatic data with parallel translations across 11 languages** (all the

[1]This metric refers to "ACC" in Table 6. ACC is the percentage of data points where a semantic translation is ranked at least 5% higher than a literal translation. When we say our fine-tuned models outperform baselines by an average of 8.83 point improvement, we mean our fine-tuned models rank 8.83% more semantic translations higher than literal translations (by at least 5%) compared to baselines.

| Dataset | Year | # samples | # of langs | PT? |
|---|---|---|---|---|
| MAGPIE (Haagsma et al., 2020) | 2020 | 44,500 | 1 | ✗ |
| ID10M (Tedeschi et al., 2022) | 2022 | 233,491 | 10 | ✗ |
| IMPLI (Stowe et al., 2022) | 2022 | 25,800 | 1 | ✗ |
| PETCI (Tang, 2022) | 2022 | 34,246 | 2 | ✓ |
| (Fadaee et al., 2018) | 2018 | 3,846 | 2 | ✓ |
| (Rezaeimanesh et al., 2025) | 2025 | 2,200 | 2 | ✓ |
| MOTHERLODE (ours) | 2025 | 430,255 | 11 | ✓ |

Table 1: A comparison of existing idiom resolution datasets. "PT?" indicates whether the dataset contains parallel translations (✓ indicates the dataset contains a subset of parallel translations, but not the whole dataset). As shown, our datasets are the largest datasets, with a wide variety of languages and parallel translations.

ones from ID10M and Hindi).

## 3 MOTHERLODE: the Largest Parallelly Translated Idiomatic Dataset

We augment the ID10M and MAGPIE datasets with Claude to include rich, parallel idiom translations, in a zero-shot setting. To enrich the datasets, we not only ask Claude for the correct, semantic translations, but also the disambiguated meaning and incorrect, literal translations (examples and definitions of these metadata can be found in the introduction). Our prompt can be found in Appendix A. MOTHERLODE is created with nearly $1,290 in API calls to Claude. In table 1, we compare our dataset (called MOTHERLODE) with existing datasets. Below we list the novelties:

1. Supports multiple languages: mid-high resource languages and with varying scripts. MOTHERLODE consists of data in Chinese, English, Dutch, French, German, Hindi, Italian, Japanese, Polish, Portuguese, Spanish

2. Contains enriched metadata (corresponding disambiguated meaning, literal translation, and semantic translation)

3. Largest parallel translation dataset consisting of *only* idiomatic data

However, a crucial question to ask is whether Claude is trustworthy to provide parallel translations for idioms. We justify using Claude with the following reasons:

1. **Human Evaluation**: We use human annotators to assess the quality of Claude's translation. We choose a random subset of 100 data points per language. Our prompt to the human annotators is presented in Appendix B. The translations are more often correct than not — according to Table 2, Claude is accurate (on average) 89.43%.

| Language | ✓ Idioms (%) | ✓ Sentences (%) |
|---|---|---|
| Dutch | 75 | 94 |
| French | 85 | 77 |
| German | 65 | 85 |
| Hindi | 92 | 92 |
| Italian | 88 | 93 |
| Portuguese | 80 | 88 |
| Spanish | 96 | 97 |

Table 2: Results from the human verification of the MAGPIE-AUG and ID10M-AUG datasets. "✓ Idioms" indicates the percentage of idioms that were correctly translated from English to another language. "✓ Sentences" indicates the percentage of sentences that were correctly translated from English to another language. The reason for the distinction is to allow for more fine-grained annotation. In particular cases, Claude slightly mistranslates the idioms alone, but is able to translate the idiom in a sentence correctly. For example, Claude translates the idiom "tongue in cheek" as "con ironía" in Spanish – this is not incorrect, but slightly unnatural. When translating the sentence, Claude translate the idiom as "irónicamente", which is appropriate within the sentence and a more natural translation.

2. **Idiom Detection Performance**: We ran an experiment where we test Claude's ability to detect whether an idiom exists in a sentence. We use the samples used for human evaluation. From this, we have a label whether each sample contains an idiom or not. We calculate AUC scores with those labels in Table 3.

3. **Idiom Span Detection Performance**: We ran

| ROC AUC | Precision | Recall |
|---|---|---|
| 89.45 | 69.55 | 86.21 |

Table 3: Claude idiom detection scores (does the following sentence contain an idiom?) on ground truth labels.

another experiment where we test Claude's ability to detect the span of the idiom from a sentence. To provide experimental rigor, we do not use the ID10M, MAGPIE, or MOTHERLODE datasets, but the FLUTE (Chakrabarty et al., 2022) dataset for this experiment. The results are in Table 4. ROUGE and BLEU are n-gram based metrics that calculate how much of the idiom's span Claude is able to detect. Accuracy is whether the detected span and ground truth span are an exact match. To clarify, this experiment was just to benchmark Claude's ability to detect idioms within a sentence – we do not require Claude to extract the idiom spans to create the MOTHERLODE data as they are present in the ID10M and MAGPIE datasets.

| ROUGE | BLEU | Accuracy |
|---|---|---|
| 93.55 | 86.46 | 81.41 |

Table 4: Claude idiom span detection scores (what is the span of the idiom in the following sentence?) on the FLUTE (Chakrabarty et al., 2022) test set.

4. **Past Precedence on English-Persian Idiom Translation**: Previous works rely on Claude for translation (Rezaeimanesh et al., 2025). Out of Claude-3.5-Sonnet, NLLB-200-3.3b, GPT-3.5-Turbo, Qwen-2.5-72b, Command R+-104b, GPT-4o-mini and Google Translate, they show **Claude is the best at translating idioms** to/from Persian, achieving the highest similarity to human-translated idioms.

5. **Public Reports on Translation**: Anthropic reports good translation performance, even on low-resource languages.

## 4 AL4IDIOMS: Active Learning For Idioms

Machine translation quality estimation (MTQE) models estimate the quality of a translation between a source and translated text[2] We see a need to fine-tune idiom-aware MTQE models to make MTQE comprehensive and reliable. We describe the evaluation setup, and our methodology (which we call AL4Idioms) to create idiom-aware MTQE models.

### 4.1 Evaluation Set Up

We fine-tune the WMT22 CometKiwi Direct Assessment model (called the `wmt22-cometkiwi-da`), using the original hyperparameters mentioned in the corresponding HuggingFace repository, with distributed training on 6 NVIDIA Tesla V100 GPUs. The model's backbone is an XLM-RoBERTa model (with 279M parameters), with a learning rate of 1.5e-05, batch size 4, and 5 epochs.

We have two criteria for evaluation: (1) Literal translations MTQE scores should be lower than those for semantic translations, (2) the base model MTQE performance should be maintained.

To ensure the first criteria, we fine-tune the `wmt22-cometkiwi-da` model with a ranking loss that encourages the semantic translations to be scored higher than the literal translations. We use the following metrics to measure the efficacy of the MTQE model:

1. **SEM**: the average score (across languages) for semantic translations across the MOTHERLODE testing set

2. **LIT**: the average score for literal translations across the MOTHERLODE testing set

3. **ACC**: the percentage of points in the MOTHERLODE testing set where SEM - LIT $\geq 0.05$

To ensure the second criteria, we do two things. Firstly, we select a *small* portion of the MOTHERLODE data to train on. We select a subset of highly informative data, detailed in Section 4.2. Secondly, we add in the pre-training data for the WMT22 model (in particular, we use the WMT-2019 Direct Assessment dataset). This data contains a source text, translated text, and a human-annotated MTQE score for the pair — we can use this to calculate the Pearson's correlation ($R^2$, for short) of scores between the MTQE model and the ground truth score. This is how we measure whether the base model capabilities have shifted.

[2]This makes our set up *reference-less*, as opposed to reference-based MTQE models that also require a reference translation to compare the translation against.

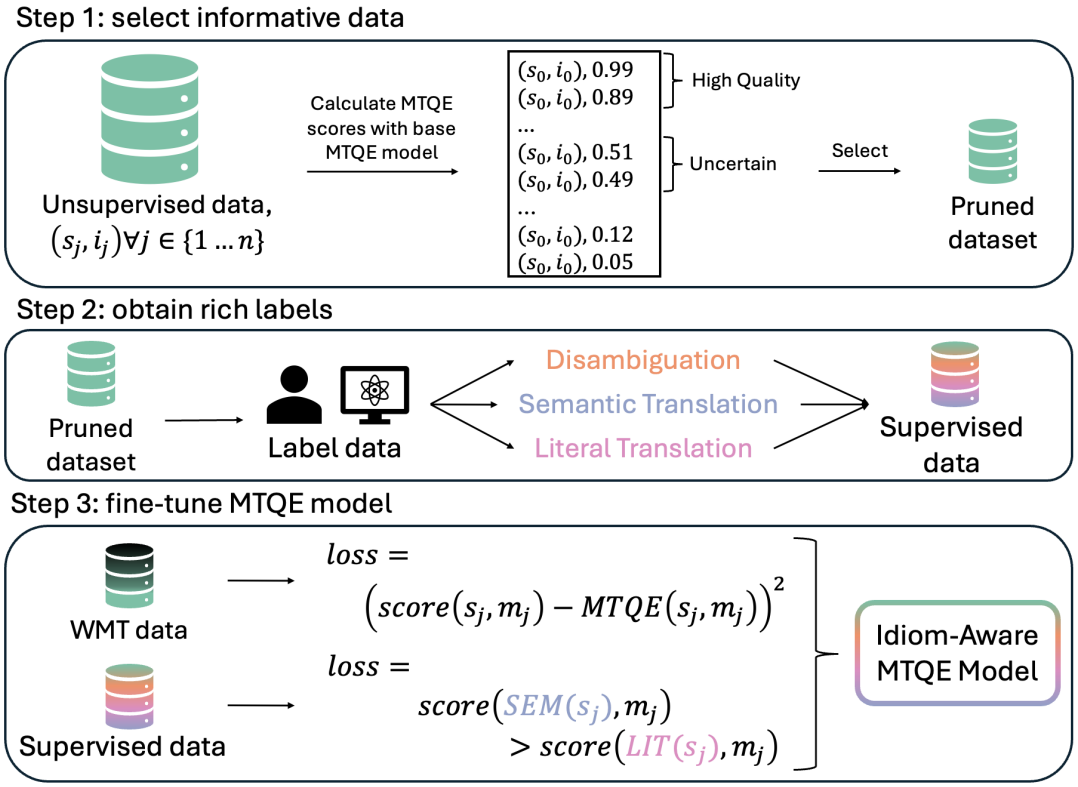


Figure 1: Our methodology to actively fine-tune idiom-aware MTQE models, called AL4IDIOMS: **A**ctive **L**earning For **I**dioms. Note: the ranking loss above is just for illustrative purposes – we use PyTorch's MarginRankingLoss.

For a more fine-grained evaluation, we divide the WMT'19 data into idiomatic and non-idiomatic sets. Encouraged by Claude's idiom and idiom span detection performance, we use Claude to classify idiomatic (denoted as PIE) and non-idiomatic (denoted as NON-PIE) data. The Pearson's correlation between the ground truth scores from WMT'19 data and the predicted scores from the base `wmt22-cometkiwi-da` model is 0.572 on PIE and 0.575 NON-PIE (both of which are statistically significant with a threshold of 0.01). Our fine-tuned model will be successful if the correlation of scores matches these numbers. Hence, we add the following metrics to our evaluation:

4. **PIE-$R^2$**: the correlation of MTQE scores between the MTQE model and ground truth score of the idiomatic WMT'19 data
5. **PIE-p-val**: the associated p-value for PIE-$R^2$
6. **NON-PIE-$R^2$**: the correlation of MTQE scores between the MTQE model and ground truth score of the non-idiomatic WMT'19 data
7. **NON-PIE-p-val**: the associated p-value for NON-PIE-$R^2$

Given our comprehensive evaluation plan, we outline our methodology that will optimize learning to discriminate between semantic and literal translations, while minimizing the effect on base model capabilities. Note: the WMT'19 dataset shares only a small overlap of common languages with MOTHERLODE (Chinese and German). Hence, when we report the correlation of the fine-tuned models to the base model, we only select data from MOTHERLODEfrom the overlap of common languages. Concretely, Table 6 shows results of models trained on only Chinese and German data, while Table 5 shows results of models trained on all languages in MOTHERLODE.

### 4.2 Methodology

From the results of the human evaluation on the MOTHERLODE dataset, we recognize the difficulty of obtaining high quality, human-verified idiom translations. Hence, we design an active learning-based method for filtering out low-quality data. Our methodology's is illustrated in Figure 1. As shown, there are three steps.

| Setting | Lang. | SEM | LIT | ACC(%) |
|---|---|---|---|---|
| BASE MODEL | ZH | 0.774 | 0.738 | 58.18 |
| | NL | 0.797 | 0.768 | 57.58 |
| | FR | 0.810 | 0.776 | 43.18 |
| | DE | 0.783 | 0.765 | 45.38 |
| | HI | 0.785 | 0.776 | 46.15 |
| | IT | 0.789 | 0.776 | 42.11 |
| | JA | 0.801 | 0.788 | 52.34 |
| | PL | 0.757 | 0.746 | 44.79 |
| | PT | 0.780 | 0.735 | 57.45 |
| | ES | 0.790 | 0.763 | 52.44 |
| | **AVG** | **0.787** | **0.763** | **49.96** |
| AL4IDIOMS-Uncertain with 50% | ZH | 0.573 | 0.537 | 63.30 |
| | NL | 0.645 | 0.623 | 43.20 |
| | FR | 0.640 | 0.621 | 51.00 |
| | DE | 0.649 | 0.631 | 47.80 |
| | HI | 0.571 | 0.550 | 60.00 |
| | IT | 0.631 | 0.604 | 55.60 |
| | JA | 0.582 | 0.554 | 61.60 |
| | PL | 0.618 | 0.600 | 42.40 |
| | PT | 0.636 | 0.614 | 44.50 |
| | ES | 0.628 | 0.607 | 44.30 |
| | **AVG** | **0.617** | **0.594** | **51.37** |
| AL4IDIOMS-High with 50% | ZH | 0.508 | 0.446 | 77.10 |
| | NL | 0.568 | 0.529 | 60.00 |
| | FR | 0.570 | 0.522 | 66.00 |
| | DE | 0.567 | 0.525 | 60.20 |
| | HI | 0.504 | 0.465 | 73.90 |
| | IT | 0.552 | 0.521 | 50.50 |
| | JA | 0.530 | 0.476 | 70.30 |
| | PL | 0.554 | 0.504 | 63.60 |
| | PT | 0.548 | 0.503 | 63.40 |
| | ES | 0.572 | 0.525 | 61.10 |
| | **AVG** | **0.547** | **0.502** | **64.61** |
| Full Data | ZH | 0.535 | 0.472 | 77.10 |
| | NL | 0.608 | 0.556 | 62.50 |
| | FR | 0.603 | 0.557 | 57.90 |
| | DE | 0.595 | 0.539 | 73.70 |
| | HI | 0.514 | 0.478 | 67.00 |
| | IT | 0.586 | 0.530 | 69.50 |
| | JA | 0.559 | 0.499 | 75.50 |
| | PL | 0.572 | 0.529 | 55.40 |
| | PT | 0.573 | 0.519 | 71.40 |
| | ES | 0.584 | 0.537 | 63.80 |
| | **AVG** | **0.573** | **0.522** | **67.38** |

Table 5: Breaking down the performance of the best performing models from Table 6 by language. Generally there is an increasing trend in performance as more data is added, with 'Full Data" and AL4IDIOMS-High showing similar performance.

**In Step 1**, we use the base model to quantify the informativeness of data and select a subset of data to label. To clarify, unlabeled data is data from the ID10M and MAGPIE datasets (the entire sentence and the span of the contained idiom), and the labeled dataset is the metadata (disambiguated meaning, semantic translation, literal translation). We use the MTQE base model to score the semantic similarity between the sentence and the idiom contained.

The reason for this is we use the MTQE model as a proxy to determine the quality of samples. Since the underlying mechanism of an MTQE model is to measure the (language-agnostic) semantic similarity between the source text and translated text, we see that sentence-idiom pairs with high semantic similarity indicate high-quality samples, with medium semantic similarity indicate samples where the model is most uncertain about, and with low semantic similarity indicate low-quality samples. In active learning literature, high-quality sampling (Yuan et al., 2025; Shi et al., 2025) and uncertainty-based sampling (Xia et al., 2025; Niekerk et al., 2025) are used, hence we report results on both settings. We select a portion of these samples to label[3].

**In Step 2**, we use either human labelers or very large, closed-source language models (i.e, Claude Sonnet 3.7) to generate metadata as labels. The supervised subset of data contains a sentence, the corresponding idiom, the disambiguated meaning of the idiom, the translated sentence with the semantic meaning of the idiom (either a culturally equivalent idiom, or a translation of the disambiguation), and the translated sentence with the literal meaning of the idiom.

We use the metadata to augment the labeled set of data. In our training, we used the supervised metadata to not only train on sentence-translation sentence pairs, but also disambiguated sentence-translation sentence pairs. Disambiguated sentences are the original sentences with the disambiguated meaning instead of the idiom. This technique increases the similarity between the idiom and the underlying disambiguated meaning.

**In Step 3**, we fine-tune the MTQE model. As outlined in the previous section (where we describe our experimental setup), we use two datasets for fine-tuning: WMT 2019's Direct Assessment (Barrault et al., 2019) (we use the entire training set) and our select MOTHERLODE data subset. Each of these datasets use different loss functions. Because we want to maintain the base model performance, for the WMT dataset, we apply an MSE loss between the predicted score (indicated by

[3]Here, we provide clarification on the sampling. In AL4Idioms-High, we select the k samples with top MTQE scores out of all sentence-idiom pairs. In AL4Idioms-Uncertain, we select the k samples with "medium" MTQE scores out of all sentence-idiom pairs. Medium scores implies the middle k of the dataset. If the dataset is 100 points, and we want to select the middle 3 points, middle-k selects the 49th, 50th, and 51st ranked point.

$score((s_j, m_j))$ in the figure) and the base model score (indicated by $MTQE((s_j, m_j))$ in the figure). For the AL4IDIOMS data subset, we use a ranking loss to ensure the score of the semantic translations are higher than the literal translations. Finally, we obtain our Idiom-Aware MTQE Model.

We do not use the MSE loss for the MOTHERLODE dataset simply because we do not have ground truth MTQE scores for it. The WMT dataset, on the other hand, does. They use humans to annotate the quality of the translations. Hence, we can use MSE for those scores. For MOTHERLODE, we only know that semantic translations need to be scored higher than literal, so we use a ranking loss for that.

### 4.3 Results

Table 6 contains the results. For reference, the "Full Data" setting takes 25 hours to fine-tune. The "Only WMT" setting does not perform well as the model in this setting cannot distinguish semantic and literal translations – this shows the quality of MOTHERLODE. Models in the "Only MOTHERLODE" setting do not perform because their correlation to the ground truth scores on the WMT'19 dataset are far from the base model, and hence highlights the effects of catastrophic forgetting. These are both things that AL4IDIOMS is able to overcome. As shown, **with just 50% of the budget, our method outperforms** and improves ACC on the (1) base model (`wmt22-cometkiwi-da`) by 21.58 points, (2) "Full Data" setting by 0.41 points and (3) best model setting using "Random" selection by 4.5 points. Our method also yields models that are **more correlated to the base model** than the baselines.

Table 5 shows the performance of the best models from Table 6 and breaks it down by the language

| | | MTQE | | | PIE | | NON-PIE | |
|---|---|---|---|---|---|---|---|---|
| Selection | % | SEM | LIT | ACC (%) | $R^2$ | p-val | $R^2$ | p-val |
| BASE MODEL | 0 | 0.747 | 0.774 | 29.47 | 0.578 | 0.0 | 0.575 | 0.0 |
| Only WMT | 0 | 0.709 | 0.716 | 0.0 | 0.572 | 1.51e-90 | 0.577 | 7.81e-42 |
| Only MOTHERLODE | 10 | 0.583 | 0.579 | 5.17 | 0.377 | 3.13e-237 | 0.325 | 1.81e-140 |
| | 25 | 0.545 | 0.509 | 33.23 | 0.424 | 2.23e-307 | 0.394 | 2.85e-210 |
| | 50 | 0.545 | 0.509 | 33.38 | 0.425 | 9.08e-308 | 0.319 | 1.94e-210 |
| | 75 | 0.544 | 0.509 | 33.23 | 0.424 | 2.23e-307 | 0.394 | 2.85e-210 |
| Random | 10 | 0.704 | 0.700 | 3.99 | 0.388 | 5.97e-254 | 0.323 | 2.35e-139 |
| | 25 | 0.649 | 0.616 | 32.07 | 0.457 | 0.0 | 0.408 | 7.11e-228 |
| | 50 | 0.606 | 0.577 | 29.72 | 0.434 | 2.00e-323 | 0.488 | 9.44e-206 |
| | 75 | 0.582 | 0.527 | 46.55 | 0.463 | 0.0 | 0.516 | 8.26e-238 |
| AL4IDIOMS-Uncertain | 10 | 0.678 | 0.679 | 7.99 | 0.453 | 1.41e-206 | 0.490 | 2.51e-111 |
| | 25 | 0.647 | 0.609 | 32.82 | 0.546 | 0.0 | 0.541 | 1.65e-240 |
| | 50 | 0.556 | 0.492 | 51.05 | 0.559 | 0.0 | 0.561 | 1.02e-238 |
| | 75 | 0.611 | 0.556 | 45.29 | 0.571 | 0.0 | 0.542 | 8.59e-249 |
| AL4IDIOMS-High | 10 | 0.673 | 0.686 | 2.54 | 0.409 | 3.14e-284 | 0.349 | 2.36e-163 |
| | 25 | 0.647 | 0.648 | 8.64 | 0.483 | 0.0 | 0.474 | 3.33e-189 |
| | 50 | 0.544 | 0.485 | 49.20 | 0.542 | 0.0 | 0.539 | 1.90e-216 |
| | 75 | 0.558 | 0.514 | 39.06 | 0.504 | 0.0 | 0.536 | 1.27e-214 |
| Full Data | 100 | 0.567 | 0.503 | 50.64 | 0.546 | 0.0 | 0.542 | 1.17e-237 |

Table 6: Results on fine-tuning WMT22-CometKiwi-DA models with varying budgets of data. As shown, AL4IDIOMS performs with 50% of data, slightly outperforming "Full Data". The "Only WMT" setting is without MOTHERLODE data, and only uses the MSE loss function in Step 3 of Figure 1. The "Only MOTHERLODE" setting is without WMT'19 data, and only uses the ranking loss function in Step 3. In darker green, we highlight the settings that yield the best performing model with respect to ACC. In light green, we highlight settings which yield models that outperforms the base model on ACC.

| Sentence Language | He kicked the bucket at 80. Spanish | | |
|---|---|---|---|
| Translation (English Translation) | Se murió a los 80. (He died at 80.) | Estiró la pata a los 80. (He stretched his leg at 80.) | Pateó el balde a los 80 años. (He kicked the bucket at 80). |
| Notes | Disambiguated translation | Linguistically/culturally equivalent idiomatic translation | Literal translation |
| Base model MTQE score | 0.451 | 0.555 | 0.499 |
| Fine-tuned MTQE score | 0.551 | 0.482 | 0.436 |

Table 7: Qualitative example of the fine-tuned MTQE model's performance on a Spanish example.

to see where the model is gaining/lacking. We only compute the SEM, LIT and ACC scores here, as we train with all language pairs from MOTHERLODE. Table 5 shows the trend of languages that improve/deteriorate from base, to fine-tuned with AL4IDIOMS, to Full Data. Some languages have peak performance with at least one variant of AL4IDIOMS (French, Hindi, Polish), but all languages improve from the base model. These results show the wide applicability of AL4IDIOMS as it improves performance across languages of different resource types, scripts, and ethnic culture. We see that certain languages dip in performance with one variant of AL4IDIOMS (for example, Portugese and Spanish with AL4IDIOMS-High – they both are better in AL4IDIOMS-High, however). An interesting future work could be to determine the allocation of data across languages.

### 4.4 Qualitative Results

Below, we show a few qualitative examples of how AL4IDIOMS helps.

**Disambiguated and culturally specific idioms are better represented.** Tables 7 and 8 demonstrates the efficacy of our method by translating the same sentence in three different ways (see "Notes"). It lost representation on the culturally equivalent translation because this particular model was not trained on Spanish or Hindi data, but it has improved MTQE for disambiguated meaning (because the idiom and disambiguated meaning shows up in other languages during fine-tuning). Overall, the correct translations are still ranked higher than the literal translation, which was not the case in the base model.

**Performance on non-idiomatic data is more aligned.** Table 9 contains a translation (left) and reference translation (right) from the WMT'19 test dataset, without any idioms. We highlight in red and green the in/accuracies of each translation. For reference, from the WMT'19 dataset, the translation on the left is given a score of 0.43 by a human annotator. Our fine-tuned model maintains the same ranking as the base model of the two translations. Our fine-tuned models outputs a score for the left translation that is closer to 0.43, compared to the base model.

**Properties of the selected subsets.** In Figure 1 and Table 6, we present two variants of our method ("-High" and "-Uncertain"). These methods vary by the range of MTQE scores between each sentence-idiom pair assigned by the base model, which can be thought of as measuring the ambiguity of the idiom in the particular sentence. Extreme MTQE scores (top-k and bottom-k) between sentence-idiom pairs indicate low ambiguity (particularly, high MTQE scores indicate low ambiguity and high quality, defining our "-High" variant); sentence-idiom pairs with medium MTQE scores indicate high ambiguity, which defines our "-Uncertain" variant. In Table 10, we analyze certain properties of these chosen subsets to understand how the data in these variants differ.

Sentence lengths for high quality samples are longer than uncertain samples, which could indicate that **more context helps ground the idiom's disambiguated meaning**. Idiom lengths do not differ much between the variants. When measuring the n-gram similarities between semantic and literal translations, we see that these translations

tend to be more similar in high quality samples than in uncertain samples. This indicates that **high quality samples could contain easier idioms** – the edits required to make incorrect, literal translations semantically correct are minimal. The more interesting result is when we see the distribution of the top-3 languages for the source sentence. This result indicates that **sentence-idiom pairs tend to have less ambiguity in languages *other than English***.

To further investigate, we read through a portion of the samples in each language to identify patterns. We notice that **the samples selected by the "-High" variant contain cultural and linguistic nuances**. For example, "Niemcy" is used in a few of the samples. In Polish, it can mean "Germany" or "the Germans", showcasing a characteristic of the Polish language, where an article[4] is not required in front of the word to distinguish the country from the ethnic group, unlike other languages. In a few of the French sentence-idiom pairs, the phrase "À la suite" is counted as the idiom, which can mean "following" (for a sequential outcome) or "as a result of" (for a causal outcome). Finally, we see a few samples were deemed as high quality if the semantic and literal translation match – this means that the same idiom occurs in both languages.

On the other hand, our hypothesis is confirmed that **samples selected in the "-Uncertain" variant contain idioms in ambiguous contexts**. To list a few examples: (1) "He tossed the bundle of joy to his friend." where "bundle of joy" refers to an infant, (2) "No whole thing's perfectly above board" where "above board" refers to being honest and non-deceptive, and (3) "They weren't out of the woods yet — far from it." which could mean being out of danger, or literally out of a forest[5]. As shown in Table 10, a large portion of these are in English.

## 5 Conclusion

In this work, we notice that machine translation quality estimation models are unreliable for idioms — they assign equivalent scores for incorrect, literal translations and correct, semantic translations of the idiomatic sentences. Here, we present two main contributions to improve MTQE models to be idiom-aware: (1) MOTHERLODE, the largest parallelly translated dataset —with only idiomatic data— across 11 languages, generated by Claude, and verified by humans to be roughly 89.43% accurate across languages; and (2) AL4IDIOMS, an active learning approach to carefully select training data can improve MTQE models' ability to distinguish between semantic and literal translations while maintaining base model capabilities. We show that with just 50% of the training budget, our approach achieves an average 8.83 point im-

[4]Like, "the" in English, "los" in Spanish, "les" in French, and "wenn" in German.

[5]In this case, it actually isn't clear which is the true meaning, as there is no clarification in the original sample from the MAGPIE dataset! Claude chooses to interpret this as the former.

| Sentence Language | My son can play video games until the cows come home if we don't stop him - he doesn't know when to quit! Hindi | | |
|---|---|---|---|
| Translation (English Translation) | Agar hum roke nahi toh mera beta bahut lambe samay tak video game khelta rahega – usse rukna hi nahi aata! (If we don't stop, then my son will play video games for a very long time – he doesn't know how to stop!) | Agar hum roke nahi toh mera beta jab tak chaand-taare rahenge video game khelta rahega – usse rukna hi nahi aata! (If we dont stop, then my son will play video games until the moon and stars last – he doesn't know how to stop! | Agar hum roke nahi toh mera beta jab tak gaye ghar nahi aa jaati video game khelta rahega – usse rukna hi nahi aata! (If we don't stop, then my son will play video games until the cows come home – he doesn't know how to stop! |
| Notes | Disambiguated translation | Linguistically/culturally equivalent idiomatic translation | Literal translation |
| Base model MTQE score | 0.698 | 0.740 | 0.855 |
| Fine-tuned MTQE score | 0.571 | 0.565 | 0.552 |

Table 8: Qualitative example of the fine-tuned MTQE model's performance on a Hindi example. Green and red show semantic and literal translations, respectively. We show the romanized sentences here for easy PDF rendering, but the scores were obtained with the corresponding text in Devanagari script.

| Sentence Language | At first he thought it was just excitement of catching a lobster, but then he realized that he was yelling, 'I got bit!' German | |
|---|---|---|
| Translation (English Translation) | Zunächst dachte er, es sei nur Aufregung, einen Hummer zu fangen, aber dann hat er gemerkt, dass er ruft: 'Ich wurde etwas!' (At first he thought it was just excitement of catching a lobster, but then he noticed that he was yelling 'I got something!') | Zunächst dachte er, es sei die Begeisterung über den Fang eines Hummers, aber dann realisierte er, dass er rief: 'Ich wurde gebissen!" (At first he thought it was just excitement of catching a lobster, but then he realized that he was screaming: 'I was bitten!') |
| Base model MTQE score | 0.601 | 0.814 |
| Fine-tuned MTQE score | 0.510 | 0.671 |

Table 9: Qualitative example of the fine-tuned MTQE model's performance on non-idiomatic data.

| Metric Description | High | Uncertain |
|---|---|---|
| Character of the sentence | 158.948 | 140.538 |
| Number of words in idioms | 2.993 | 2.490 |
| ROUGE between sem. and lit. | 0.855 | 0.539 |
| BLEU between sem. and lit. | 0.680 | 0.479 |
| Top-3 languages of the source sentence | Polish (20.67%), French (17.17%), Spanish (12.67%) | English (98.33%), French (1.11%), Spanish (0.56%) |

Table 10: Properties of the selected subset of the data for the AL4IDIOMS-High and AL4IDIOMS-Uncertain variants. Sem are semantic translations, and lit are literal translations.

provement in accuracy over the baselines. These contributions represent meaningful progress toward more reliable and efficient idiom translation assessment.

## 6 Limitations and Ethics Statement

In this work, we rely on Anthropic's Claude model to provide ground truth translations. Claude could be wrong by either providing an incorrect translation or a culturally insensitive translation – caution must be used. Furthermore, using a strong closed source model reduces the accessibility of our method. Future work can explore using small language models as oracles who can also provide a confidence score for their translations, enabling a confidence-aware active learning approach with weak oracles. Our dataset, MOTHERLODE, should be used with *a pinch of salt* as the underlying datasets (MAGPIE and ID10M) could contain unintended biases which were propagated to our dataset. This work focuses on high to mid-resource European and Asian languages, and we leave to future work to show the generalization on low resource languages.

## References


Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations.

Kellen Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pretrained language models. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys*, 56(10):1–34.

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

Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2025. A confidence-based acquisition model for self-supervised active learning and label correction. *Transactions of the Association for Computational Linguistics*, 13:167–187.

Sara Rezaeimanesh, 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.

Burr Settles. 2009. Active learning literature survey.

Haochen Shi, Xinyao Liu, Fengmao Lv, Hongtao Xue, Jie Hu, Shengdong Du, and Tianrui Li. 2025. A pre-trained data deduplication model based on active learning.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. Impli: Investigating nli models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388.

Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *arXiv preprint arXiv:2202.09509*.

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.

Mingmin Wu, Guixin Su, Yongcheng Zhang, Zhongqiang Huang, and Ying Sha. 2024. 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.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Kenneth Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, Zhehao Zhang, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi, and Julian McAuley. 2025. From selection to generation: A survey of llm-based active learning.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2025. Hide and seek in noise labels: Noise-robust collaborative active learning with llm-powered assistance.

Ziheng Zeng, Kellen Tan Cheng, Srihari Venkat Nanniyur, Jianing Zhou, and Suma Bhat. 2023. Iekg: A commonsense knowledge graph for idiomatic expressions.

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.

# A Prompt to Claude 3.7-Sonnet for creating MOTHERLODE

Given the following idiom and the context, can you provide the following five things:

1. "disambiguation": a short phrase or sentence describing the semantic meaning of the idiom (in the same language as the idiom, please)
2. "semantic_translation_of_idiom": if the idiom is idiomatic, provide either a translation of the semantic meaning of the idiom in {target_language} or an equivalent idiom in {target_language}. If the idiom is simply a phrase (use the context to determine this), simply tranlsate it in {target_language}.
3. "semantic_translation_of_sentence": a SEMANTIC translation of the entire sentence {target_language}.
4. "literal_translation_of_idiom": a literal translation of the idiom in {target_language}.
5. "literal_translation_of_sentence": a translation of the sentence in {target_language} containing a literal translation of the idiom.

Below are some examples to help.

Example 1:
Context: “He is under the weather.”
Idiom: "under the weather"
disambiguation: "becoming ill, falling asleep”
semantic_translation_of_idiom: "ठंड लगना"
semantic_translation_of_sentence: "उसको ठंड लग गई"
literal_translation_of_idiom: "मौसम के नीचे"
literal_translation_of_sentence: "वो मौसम के नीचे है"

Example 2:
Context: "the iron ball was hanging by a thread."
Idiom: "hanging by a thread"
disambiguation: "to be physically suspended by a thin string or fiber"
semantic_translation_of_idiom: "colgar de un hilo”
semantic_translation_of_sentence: "La bola de hierro colgaba de un hilo"
literal_translation_of_idiom: "colgar de un hilo”
literal_translation_of_sentence: "La bola de hierro colgaba de un hilo"

Example 3:
Context: "He kicked the bucket at 80."
Idiom: "kicked the bucket"
disambiguation: "died; passed away"
semantic_translation_of_idiom: "gestorben”
semantic_translation_of_sentence: "Er ist mit 80 gestorbe"
literal_translation_of_idiom: "Den Eimer gekickt"
literal_translation_of_sentence: "Er hat den Eimer mit 80 gekickt"

Example 4:
Context: "He kicked the bucket to the curb."
Idiom: "kicked the bucket"
disambiguation: "To physically strike a bucket with one's foot"
semantic_translation_of_idiom: "踢一个桶”
semantic_translation_of_sentence: "他在路边踢了一个桶"
literal_translation_of_idiom: "踢一个桶”
literal_translation_of_sentence: "他在路边踢了一个桶"

To help generate the three sentences above, you are also provided some context in which the idiom is used:
Context: "{context}"
Idiom: "{idiom}"

Output your final answer in the following JSON:
{{
"disambiguation": <output a short phrase or sentence describing the semantic meaning of the idiom (in the same language as the idiom, please)>,
"semantic_translation_of_idiom": <depending on the context, translate it semantically or literally in {target_language}>,
"semantic_translation_of_sentence": <translate the entire sentence semantically in {target_language}>,
"literal_translation_of_idiom": <output a literal translation of the idiom in {target_language}>
"literal_translation_of_sentence": <output a literal translation of the sentence in {target_language}>
}}

Figure 2: Prompt to Claude to generate the disambiguated meaning, semantic translation, and literal translation for an idiom. “context” and “idiom” are inputs from the dataset.

## B Human Annotator Prompt

In Figure 3, we provide our prompt to the human annotators to verify Claude's output.

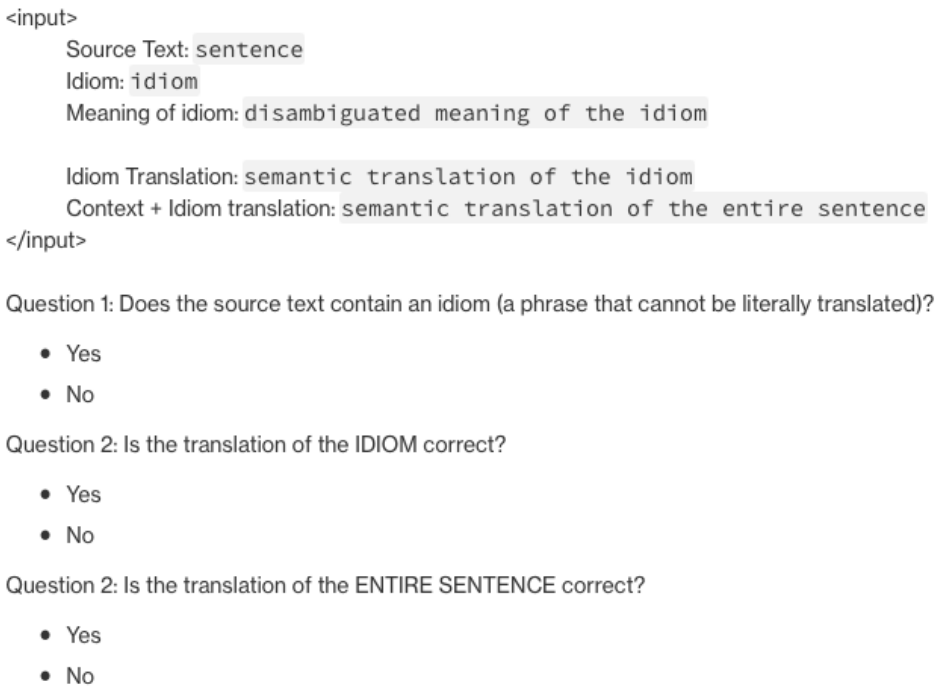


Figure 3: Screenshot of the prompt provided to the human annotators to verify Claude's output.