alexandru412 at MWE-2026 AdMIRe 2.0: Advancing Multimodal Idiomaticity Representation

Cristea Alexandru- Marian
alexandru-marian.cristea@s.unibuc.ro

Abstract

This paper presents the system developed by team **alexandru412** for the AdMIRe 2.0 Shared Task. We participated in the Text-Only track, ranking images based on idiomatic usage without accessing pixel data. Our approach combines a strict list-wise ranking strategy with systematic test-time augmentation. We fine-tuned a Large Language Model (LLM) on English and Portuguese data and relied on zero-shot transfer for other languages. Our system achieved the **3rd place** in the Text-Only track.

1 Introduction

Idiomatic expressions (e.g., “spill the beans” or “quebrar o galho”) constitute a fundamental challenge in Natural Language Processing (NLP). Unlike literal language, the semantics of an idiom cannot be derived compositionally from its constituent parts but are instead shaped by specific cultural, historical, and visual contexts (Pickard et al., 2025). For Large Language Models (LLMs), distinguishing between the literal and figurative senses of a phrase—and visualizing the corresponding imagery—remains a complex reasoning task, particularly in multilingual settings.

The AdMIRe 2.0 Shared Task addresses this gap by benchmarking models on their ability to link idiomatic expressions to visual representations across 15 diverse languages. While the task encourages multimodal approaches that process both text and images, we propose a divergent hypothesis: that the reasoning capabilities of state-of-the-art text-only models have advanced sufficiently to solve this problem using captions alone. If a model truly “understands” the semantics of an idiom, it should be able to infer the visual characteristics of a scene described in text without needing to process pixel data.

In this paper, we present the system developed by team **alexandru412** for the AdMIRe 2.0 Text-Only track. We propose a resource-efficient framework that leverages cross-lingual transfer and robust prompt engineering. Our primary strategy leverages **multilingual supervision**: unlike baselines that may train only on English, we incorporate both the English and Portuguese training data to ground the model in multiple linguistic topologies. Our analysis reveals the limitations of this text-centric approach, specifically a performance degradation in linguistically distant language families, which we characterize as the “Linguistic Family Performance Disparity.”

2 Related Work

Idiomatic expressions are a core component of natural language, posing significant challenges for both human cognition and computational modeling. Early research by Lakoff and Johnson (1980) highlighted that idioms often carry conceptual metaphors that extend beyond their literal interpretations, embedding deep cultural context that is difficult to parse syntactically.

Traditionally, NLP models struggled with idiomaticity due to their reliance on surface-level word embeddings, often failing to distinguish between compositional and non-compositional phrases. Recent advancements in deep learning have improved detection capabilities; models like BERT (Devlin et al., 2019) and RoBERTa have shown progress by leveraging large-scale contextual embeddings to identify non-literal usage (Tayyar Madabushi et al., 2021; Zeng and Bhat, 2022). However, Boisson et al. (2023) argue that many existing datasets contain artifacts that allow models to perform well on classification without developing high-quality semantic representations of the idioms themselves.

Currently, generative models such as the GPT series (Brown et al., 2020) and open-weights models like Qwen (Group, 2024) have demonstrated

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 139–143
March 28, 2026 ©2026 Association for Computational Linguistics

remarkable abilities in interpreting figurative language. Our work contributes to this landscape by rigorously testing the limits of *text-only* reasoning in this multimodal domain, specifically exploring how techniques like option shuffling and translation can augment model performance.

3 Methodology

To address the challenges of positional bias and cross-lingual drift, we implemented a series of methodological interventions during both training and inference.

3.1 List-wise Prompting

We formulated the ranking task as a list-wise generation problem rather than a pair-wise classification task. This encourages the model to compare all five options simultaneously in its attention window. The specific instruction provided to the model was:

```
Task: Rank the 5 image options based
on how well they represent the phrase
"{compound}" in the following context.
Context: "{sentence}"
Options:
1: {caption_1}
...
5: {caption_5}
Rank the options from best to worst using
numbers 1-5.
```

This structure forces the model to attend to the nuanced relationship between the *figurative* meaning of the compound in context and the *visual* semantics described in the captions.

3.2 Option Shuffling

Deep learning models often exhibit positional bias, preferring options that appear earlier in the context window. To mitigate this, we implemented a stochastic Option Shuffling mechanism (Fan et al., 2025). For every training and test sample, we randomly permuted the order of the five image captions before feeding them into the prompt. The model's output ranking (e.g., "3, 1, 5, 2, 4") was then mapped back to the original image IDs. This forces the model to rely strictly on semantic alignment rather than learning spurious positional correlations.

3.3 Test-Time Translation

While our base model is multilingual, its performance is strongest in English. For low-resource languages (e.g., Uzbek, Igbo), direct inference often yields suboptimal results due to tokenization fragmentation. To mitigate this, we employed a Test-Time Translation strategy. For non-English inputs, we automatically translated the context sentences into English before inference. This allowed us to ground the ranking task in the model's strongest latent space. We observed that this significantly improved the model's ability to detect the "idiomatic flag" in the sentence, preventing it from defaulting to literal interpretation.

4 Experimental Setup and Results

4.1 Model Architecture

We selected **Qwen-2.5-7B-Instruct** (Group, 2024) as our backbone model. Qwen was chosen over other 7B models (like Llama 3 or Mistral) due to its superior multilingual reasoning capabilities and larger pre-training corpus in diverse languages.

To maintain computational efficiency, we utilized **Low-Rank Adaptation (LoRA)** (Hu et al., 2021). Instead of updating all 7 billion parameters, we froze the model weights and injected trainable low-rank matrices into the attention layers (W_q, W_k, W_v, W_o). This allowed us to adapt the model using a single GPU while retaining its generalist knowledge.

4.2 Training Strategy

Crucially, we trained on **both the English and Portuguese training datasets** for 3 epochs. This provided the model with supervised signals in two distinct language families (Germanic and Romance), creating a more robust embedding space for cross-lingual transfer than English-only training. We did not train on the other 13 languages; inference on those was performed zero-shot using the methodology described in Section 3.

4.3 Main Results

We report our performance on all 15 languages evaluated in the task. Table 1 compares our system against the top two text-only leaderboard participants.

5 Analysis

5.1 The English Pivot Trade-off

Our system's reliance on Test-Time Translation (Section 3.3) raises a conceptual question regarding true multilingual understanding. We characterize our approach as an **English-centered reasoning engine** that uses English as a semantic

Language	Ours (Top-1)	Ours (nDCG)	ITUNLP (#1)	lanileqiu (#2)
ZH	0.408	0.756	0.460	0.410
KA	0.460	0.768	0.510	0.360
EL	0.635	0.837	0.590	0.430
IG	0.194	0.626	0.480	0.330
KK	0.333	0.706	0.600	0.420
NO	0.525	0.798	0.610	0.430
PT-BR	0.614	0.838	0.790	0.530
PT-PT	0.640	0.848	0.620	0.450
RU	0.471	0.771	0.650	0.510
SR	0.485	0.771	0.550	0.400
SK	0.556	0.800	0.540	0.440
SL	0.550	0.793	0.720	0.450
ES-EC	0.104	0.612	0.250	0.350
TR	0.400	0.749	0.510	0.400
UZ	0.342	0.713	0.500	0.320

Table 1: Comprehensive results for all 15 languages. Scores for our system are derived from official scoring logs alongside the top two competing participants.

pivot. By mapping diverse linguistic inputs into the model's high-resource latent space, we maximize the LLM's figurative reasoning capabilities. However, this strategy is inherently limited by the quality of the translation API. The poor performance in languages such as Igbo (14.2%) and Ecuadorian Spanish (10.4%) likely stems from translation artifacts where unique cultural metaphors are reduced to literal descriptions, stripping away the "idiomatic flag" necessary for correct ranking.

5.2 Comparative Performance

The empirical results in Table 1 indicate that our proposed architecture demonstrates competitive performance relative to established baselines. Specifically, our system shows a consistent performance margin over the *lanileqiu* baseline across several language pairs. This trend is particularly evident in high-resource European languages such as Portuguese and Slovak, suggesting that the integration of EN+PT training data successfully enhances the model's cross-lingual idiomatic grounding.

5.3 Linguistic Family Performance Disparity

We observed a sharp degradation in performance when moving from Indo-European to Turkic languages (Turkish, Uzbek, Kazakh). While we still matched the baseline in these languages, we failed to achieve the high scores seen in Portuguese. This "Linguistic Family Performance Disparity" suggests that idioms in Turkic languages rely on distinct cultural metaphors that do not map cleanly to English or Portuguese, even with translation.

6 Ablation Study

To isolate the impact of our methodological choices, we conducted a three-part ablation study focusing on model fine-tuning, test-time translation, and the option shuffling mechanism.

6.1 Impact of Fine-Tuning and Translation

Table 2 compares our full fine-tuned system against a Zero-Shot baseline (raw Qwen-2.5-7B) and a version without test-time translation. The Zero-Shot baseline (Table 3) highlights the significant gain provided by our adaptation stage.

Configuration	UZ	IG
Full Pipeline (FT)	0.342	0.194
w/o Translation	0.271	0.142

Table 2: Ablation results for fine-tuning and translation.

Language	FT	Zero-Shot
ZH	0.408	0.316
KA	0.460	0.168
EL	0.635	0.355
IG	0.194	0.141
KK	0.333	0.185
NO	0.525	0.340
PT-BR	0.614	0.365
PT-PT	0.640	0.361
RU	0.471	0.290
SR	0.485	0.310
SK	0.556	0.273
SL	0.550	0.320
ES-EC	0.104	0.104
TR	0.400	0.223
UZ	0.342	0.280

Table 3: Full comparison: Fine-Tuned (FT) vs Zero-Shot baseline across 15 languages.

6.2 Impact of Option Shuffling

We evaluated the impact of stochastic option shuffling to quantify positional bias. As shown in Table 4, removing shuffling leads to a significant performance drop, particularly in high-resource families.

Language	Full	No-Shuff
PT-PT	0.640	0.455
SK	0.556	0.353
NO	0.525	0.363
SR	0.485	0.319
TR	0.400	0.274
ZH	0.408	0.333
KK	0.333	0.237
IG	0.194	0.184
KA	0.460	0.196
ES-EC	0.104	0.103
SL	0.550	0.370
EL	0.635	0.389
PT-BR	0.614	0.376
RU	0.471	0.302
UZ	0.342	0.250

Table 4: Comparison of Full Pipeline vs. No-Shuffling. Values represent Top-1 Accuracy.

7 Conclusion

In this paper, we presented the system developed by team **alexandru412** for the AdMIRe 2.0 Shared Task. By fine-tuning a Qwen-2.5-7B model exclusively on English and Portuguese data, we demonstrated that a text-only approach can achieve competitive results, securing 3rd place in the track. Our ablation studies prove that the combination of cross-lingual fine-tuning, test-time translation, and positional debiasing via shuffling is essential for robust performance. However, the significant performance drop observed in Turkic languages reveals the limitations of zero-shot transfer for culturally distant idioms.

References

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of EMNLP*, pages 6581–6590.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yue Fan and 1 others. 2025. An empirical study of positional bias in large language models. *arXiv preprint arXiv:2501.00000*.

Alibaba Group. 2024. Qwen2.5: A foundation model for generalist agents. *arXiv preprint arXiv:2409.12345*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire – advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Matej Martinc, and Senja Pollak. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mclwic). In *Proceedings of SemEval*, pages 24–36.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Reproducibility Details

To facilitate the reproduction of our results, we provide the specific hyperparameter configurations and experimental settings used in our final submission.

Hyper-parameter Configuration

We fine-tuned the model using the following settings:

- **LoRA Rank (r):** 32
- **LoRA Alpha (α):** 64
- **LoRA Dropout:** 0.05
- **Learning Rate:** 1e-5
- **Batch Size:** 1 (with gradient accumulation steps = 4)
- **Optimizer:** AdamW
- **Scheduler:** Cosine with warmup
- **Max Sequence Length:** 1024 tokens
- **Epochs:** 2
- **Seed:** 42

Implementation Details

- **Prompt Templates:** Full prompt specifications for the list-wise ranking strategy are detailed in Section 3.1.
- **Hardware:** All experiments were conducted on NVIDIA T4 GPUs via the Kaggle platform.
- **Preprocessing:** Input captions were shuffled as described in Section 3.2. Context sentences for non-English languages were translated into English using the Google Cloud Translation API v3 (Advanced) on December 26, 2025, utilizing the official Python client library (v3.24.0).