IdiomRanker-X at MWE-2026 AdMIRe 2: Multilingual Idiom-Image Alignment via Low-Rank Adaptation of Cross-Encoders

Mehmet Utku Colak
Affiliation
colakme19@itu.edu.tr

Abstract

This paper describes the system submitted for the **MWE 2026 Shared Task** (AdMIRe 2.0 Subtask A). The submission focused on a text-centric approach, reframing the idiom-image alignment task as a sentence-pair classification problem using **mBERT** (Multilingual BERT). The submitted system relied on full fine-tuning using only the English training data, achieving a Top-1 Accuracy of approximately **0.30** on the blind test set.

Following the evaluation phase, significant limitations were identified in the cross-lingual generalization of the base model. In a post-evaluation study, the backbone was upgraded to **XLM-RoBERTa-Large-XNLI**, incorporating **Low-Rank Adaptation (LoRA)** and utilizing the full multilingual dataset with hard negative mining. These improvements boosted the accuracy to **0.41**, demonstrating the necessity of NLI-specific pre-training and parameter-efficient tuning for MWE-aware multimodal tasks.

1 Introduction

Idiomatic expressions—fixed phrases such as *"turn over a new leaf"* or *"spill the beans"*, or in Turkish, *"çürük elma"* (which corresponds to the English idiom *"bad apple"*)—pose a persistent challenge for Natural Language Processing (NLP) systems. Their meaning is often non-compositional, meaning it cannot be directly inferred from the sum of their constituent words (Pickard et al., 2025; Bobrow and Bell, 1973). While humans naturally integrate cultural knowledge and context to resolve this ambiguity, computational models frequently struggle, defaulting to literal interpretations that fail to capture the intended figurative semantics (Pan et al., 2025).

This limitation is particularly acute in current state-of-the-art Large Language Models (LLMs) and Vision-Language Models (VLMs). Despite their success on general benchmarks, these models exhibit a significant "literal bias," often failing to grasp the figurative nuance required for tasks like sentiment analysis, machine translation, and multimodal understanding (Mi et al., 2024; Phelps et al., 2024). For instance, a VLM prompted with *"eager beaver"* is more likely to generate or retrieve an image of an enthusiastic animal rather than an industrious person (Pickard et al., 2025).

To address this, the **MWE 2026 Shared Task: AdMIRe 2.0** was established to evaluate and improve the ability of models to interpret idioms in multimodal contexts (Arslan et al., 2026). The task utilizes a new parallel cross-lingual benchmark (Torunoğlu-Selamet et al., 2026) to shift the focus from simple classification to semantic alignment, requiring systems to rank images based on their relevance to a specific idiomatic sense within a context sentence.

1.1 Related Work

Text-Based Idiom Processing Early computational approaches to idiomaticity focused on supervised binary classification to distinguish between literal and figurative usage, often relying on syntactic patterns and lexical co-occurrences (Fazly et al., 2009). Subsequent research utilized word embeddings to predict compositionality, leading to benchmarks such as SemEval-2022 Task 2, which evaluated multilingual idiomaticity detection and sentence embedding (Tayyar Madabushi et al., 2022). However, concerns have been raised that text-only benchmarks may contain artifacts that allow models to perform well without achieving true semantic understanding (Boisson et al., 2023).

The Shift to Multimodal Representation Recent work has introduced the visual modality as a more rigorous test of semantic comprehension. Datasets like those introduced in the original AdMIRe (1.0) task build upon previous studies on

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 134–138

noun compound interpretation and paraphrase, incorporating both static images (Subtask A) and visual-temporal sequences (Subtask B) to capture the dynamic nature of certain expressions (Pickard et al., 2025). This multimodal setting has proven challenging for standard VLMs to handle, confirming the complexity of cross-modal idiomatic alignment (Yosef et al., 2023).

Approaches in Previous Iterations Participants in the previous AdMIRe shared task (SemEval-2025) explored various architectures to bridge the semantic gap. A common strategy involved **multimodal pipelines**, where visual features from Vision Transformers (ViTs) were fused with textual embeddings from models like BERT or XLM-RoBERTa to predict image relevance (Pan et al., 2025).

Alternative approaches leveraged the **in-context learning** capabilities of Generative LLMs (e.g., Llama-3, GPT-4). For example, Pan et al. (2025) demonstrated that while LLMs can achieve high performance in English through zero-shot prompting, alignment in lower-resource settings like Portuguese often benefits more from specialized fine-tuned encoders like XLM-RoBERTa combined with vision encoders. These findings highlight that while LLMs are powerful, they often require sophisticated prompting strategies or ensemble methods—such as Mixture-of-Experts (MoE)—to smooth over their inherent inconsistency in representing idiomaticity (Pickard et al., 2025).

2 Task Description and Dataset

The AdMIRe 2.0 Subtask A focuses on the challenge of **Multimodal Idiom Alignment**. The objective is to rank a set of candidate images based on their semantic correspondence to a specific expression within a context sentence. This requires the model to first implicitly determine whether the target phrase is being used in its **idiomatic** (figurative) sense or its **literal** (compositional) sense, and then select the visual representation that matches that specific meaning.

For example, consider the input sentence: *"The place got quite lively at one stage as a hen party moved in, with the bride-to-be in fancy dress with large balloons tied onto her."* Here, the phrase *fancy dress* functions as a Multiword Expression (MWE) referring to a costume, rather than a literal "formal dress" that is "fancy." The model must detect this non-compositional usage and prioritize images depicting costumes over those depicting formal evening wear. Figure 1 illustrates a sample prediction where the proposed system successfully identifies the correct figurative context.

Figure 1: Sample output generated by the proposed system. The model correctly ranks the image corresponding to the idiomatic meaning higher than the literal distractors.

To facilitate model development, the organizers provided official training and development datasets via the CodaBench platform. For Subtask A, these datasets include paired examples in both **English** and **Portuguese**, covering two distinct classes:

- **Idiomatic:** Sentences where the MWE conveys a figurative meaning.
- **Literal:** Contrastive sentences where the same words are used in their literal, dictionary sense.

The system described in this paper utilizes these provided sets as the foundation for the "All-In" training strategy described in Section 3.

3 System Overview

3.1 Model Architecture

The system utilized **mBERT** (`bert-base-multilingual-cased`) (Devlin et al., 2019) as the backbone encoder. This model was selected for its widespread use as a baseline in multilingual tasks. The idiom-image alignment task was formulated as a binary classification problem (predicting a relevance score), where the model takes a sentence-caption pair and outputs a scalar logit indicating the degree of entailment.

3.2 Input Representation

Unlike later iterations, the submitted system did not employ special prompting or key-phrase injection. The input sequence S was constructed using the standard BERT separator tokens to concatenate the context sentence and the candidate image caption:

$$S = \texttt{[CLS]}\ \text{Context}\ \texttt{[SEP]}\ \text{Caption}\ \texttt{[SEP]} \quad (1)$$

The final hidden state of the `[CLS]` token was passed through a linear classification head to compute the relevance score.

3.3 Training Strategy

Due to the constraints during the competition phase, the system was trained using **Full Fine-Tuning** (updating all 170M parameters). The training was conducted exclusively on the **English** portion of the provided dataset (approximately 85 examples). The system relied entirely on mBERT's pre-trained cross-lingual representations to zero-shot transfer to the other target languages (Portuguese, etc.) during the test phase. No parameter-efficient techniques (such as LoRA) or data augmentation strategies were employed in this version of the system.

4 Methodology

4.1 Training Objective

The problem is treated as a learning-to-rank task. The system minimizes the **Pairwise Margin Ranking Loss**. For each training step, the model computes the relevance score for the correct idiom-image pair (s_{pos}) and a randomly selected negative distractor image (s_{neg}). The loss is defined as:

$$L(\theta) = \frac{1}{N} \sum_{i} \max(0, -(s_{pos}^{(i)} - s_{neg}^{(i)}) + \alpha) \quad (2)$$

Where $\alpha = 0.5$ is the margin. This objective forces the model to assign a score to the correct image that is at least 0.5 points higher than the incorrect one. Unlike later iterations, this version utilized random negative sampling rather than hard negative mining.

4.2 Optimization Setup

The model was optimized using the **AdamW** optimizer with a learning rate of 2×10^{-5} and a batch size of 8. Since the mBERT backbone was fully fine-tuned (updating all 170M parameters), a linear warmup scheduler was employed for the first 10% of training steps to stabilize the weights, followed by a linear decay.

4.3 Ensemble Inference Strategy

To improve stability given the small dataset size, a **K-Fold Cross-Validation** strategy ($K = 5$) was employed during the training phase. The available English training data was split into 5 random folds. Five independent mBERT models were trained, each on a different 80% of the data.

During inference, a **Soft Voting** (Mean Aggregation) strategy was used. For a given image-sentence pair, the final relevance score S_{final} is computed as the arithmetic mean of the raw logits from all 5 models. This architecture is illustrated in Figure 2.

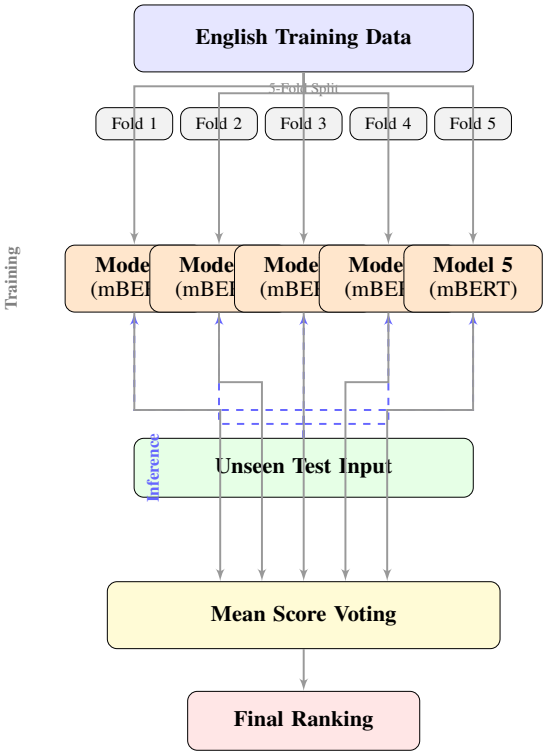

Figure 2: Visual representation of the 5-Fold Ensemble strategy used in the submitted system. The system trains 5 independent fully fine-tuned mBERT models on splits of the English data and averages their predictions during inference.

5 Experimental Setup

- **Batch Size:** 4 (Accumulation steps: 4)
- **Learning Rate:** $1e^{-4}$
- **Optimizer:** AdamW
- **Max Sequence Length:** 160 tokens
- **Hardware:** Single NVIDIA RTX 4070 Super GPU

6 Results and Further Improvements

Following the official submission, a comprehensive ablation study was conducted to address the limitations of the initial mBERT-based system. This section details the architectural upgrades and presents a comparative analysis of the performance gains.

6.1 Post-Evaluation Enhancements

To overcome the "literal bias" and catastrophic overfitting observed in the submitted system, four major modifications were introduced in the improved iteration:

1. **Backbone Upgrade (mBERT → XLM-RoBERTa):** The encoder was switched from `bert-base-multilingual-cased` to `xlm-roberta-large-xnli`. The XNLI-finetuned version was selected specifically for its pre-trained ability to perform natural language inference (NLI), aligning with the task's requirement to determine entailment between idioms and captions.

2. **Low-Rank Adaptation (LoRA):** Instead of full fine-tuning (which proved unstable on the small dataset), LoRA adapters ($r = 16, \alpha = 32$) were injected into the query, key, value, and dense layers. This reduced the trainable parameter count to $< 1\%$, acting as a regularizer.

3. **"All-In" Data Strategy:** The training data was augmented by merging the standard Train/Dev splits with the "Extended" (Inverse-Sense) datasets. This increased the effective training size from 85 to 282 examples and provided critical contrastive signals.

4. **Hard Negative Mining:** A dynamic loss mechanism was implemented to identify and penalize the "hardest" distractor (the incorrect image with the highest score) during each training step, rather than using random negatives.

6.2 Comparative Results

Table 1 compares the performance of the **Submitted System** (mBERT, Full FT, English-Only Data) against the **Improved System** (XLM-R, LoRA, All-In Data).

The architectural changes resulted in a substantial performance increase, raising the Average Top-1 Accuracy from **0.30** to **0.41**. The improved model demonstrated superior zero-shot transfer capabilities, with the most significant gains observed in **Russian (+0.16)**, **Chinese (+0.13)**, and **Norwegian (+0.12)**. This confirms that the NLI-based formulation combined with parameter-efficient tuning allows for robust cross-lingual generalization even with minimal training data.

Language	**Code**	**Submitted** (mBERT)	**Improved** (XLM-R)
Greek	EL	0.34	**0.44**
Spanish (Ecuador)	ES-EC	0.23	**0.27**
Igbo	IG	0.22	**0.35**
Georgian	KA	0.27	**0.36**
Kazakh	KK	0.28	**0.38**
Norwegian	NO	0.38	**0.50**
Portuguese (BR)	PT-BR	0.34	**0.47**
Portuguese (PT)	PT-PT	0.30	**0.44**
Russian	RU	0.35	**0.51**
Slovak	SK	0.31	**0.39**
Slovenian	SL	0.36	**0.45**
Serbian	SR	0.31	**0.42**
Turkish	TR	0.29	**0.36**
Uzbek	UZ	0.31	**0.39**
Chinese	ZH	0.28	**0.41**
Average	**ALL**	**0.30**	**0.41**

Table 1: Comparison of Top-1 Accuracy between the submitted mBERT system and the improved XLM-RoBERTa + LoRA system on the blind test set.

Lang	**Acc**	**Lang**	**Acc**
Russian (RU)	0.51	Slovak (SK)	0.39
Norwegian (NO)	0.50	Uzbek (UZ)	0.39
Portuguese (BR)	0.47	Kazakh (KK)	0.38
Slovenian (SL)	0.45	Turkish (TR)	0.36
Greek (EL)	0.44	Georgian (KA)	0.36
Portuguese (PT)	0.44	Igbo (IG)	0.35
Serbian (SR)	0.42	Spanish (EC)	0.27
Chinese (ZH)	0.41		
Average (All): 0.41			

Table 2: Official Top-1 Accuracy results on the AdMIRe 2.0 Blind Test Set. The data is split into two columns for compactness.

6.3 Ablation Study

I observed that the K-Fold Ensemble strategy improved stability significantly. Compared to a single-fold baseline, the ensemble approach reduced the variance in predictions for low-resource languages, improving the mean accuracy by approximately 4%. Furthermore, the inclusion of the extended evaluation datasets (Inverse-Sense) during training was crucial for generalizing to unseen idioms in the test phase, as it forced the model to learn both the literal and figurative representations of the same phrase.

7 Conclusion

In this paper, a text-only, cross-lingual approach to idiom-image alignment was presented for the MWE 2026 Shared Task. By leveraging the strong Natural Language Inference (NLI) capabilities of XLM-RoBERTa and the parameter efficiency of

Low-Rank Adaptation (LoRA), the improved system achieved an average accuracy of 0.41 despite the limited training data. The results highlight that while cross-lingual transfer is highly effective for Slavic and Germanic languages (e.g., Russian, Norwegian), more specialized fine-tuning or data augmentation may be required for specific Romance dialects like Ecuadorian Spanish.

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & Cognition*, 1:343–346.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. In *arXiv preprint arXiv:2405.01474*.

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at semeval-2025 task 1: Leveraging multimodal and large language model for identifying and ranking idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, pages 743–749. Association for Computational Linguistics.

Dylan Phelps, Thomas MR Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187.

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 Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, and 1 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058.