LST at MWE-2026 AdMIRe 2: Advancing Multimodal Idiomaticity Representation

Le QIU and **Yu-Yin HSU** and **Emmanuele CHERSONI**
The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Hong Kong SAR

Abstract

This paper presents our methods for the AdMIRe 2.0 shared task, which addresses multilingual and multimodal idiom understanding. Our submission focuses on the text-only track. Specifically, we employ an ensemble of three large language models (LLMs) to directly perform the presented image ranking task. Each model independently produces a ranking of the candidate images, and we aggregate their outputs using a hard voting strategy to determine the final prediction. This ensemble learning framework, by leveraging the complementary strengths of different LLMs, provides a training-free and robust solution to the AdMIRe 2.0 task and places our method in the second position on the leaderboard.

1 Introduction

The AdMIRe 2.0 Shared Task (Arslan et al., 2026; Torunoğlu-Selamet et al., 2026) is an expanded continuation of its precedent, Subtask A of SemEval-2025 Task 1: AdMIRe – Advancing Multimodal Idiomaticity Representation (referred to as AdMIRe 1.0 for distinction). In AdMIRe 1.0, Subtask A is formulated as a static image ranking task (Pickard et al., 2025): Given a Potentially Idiomatic Expression (PIE), specifically a nominal compound (NC), its surrounding context sentence, and a set of five images each accompanied by a descriptive caption, the system is required to rank the images according to how accurately they depict the meaning of the NC in the provided context. A mono-modal track of the task allows participants to perform the ranking using only the textual captions. The images are not randomly generated; rather, each is deliberately associated with the PIE either figuratively or literally, with the fifth a distractor. A demonstration is provided in Figure 1. AdMIRe 1.0 covers two languages: English and Portuguese. Building upon this, AdMIRe 2.0 broadens the scope of the task by extending the dataset to a substantially larger set of 15 languages. Importantly, during the training phase, only the AdMIRe 1.0 data (in English and Portuguese) are available, and participants are not informed of which additional languages will appear in the evaluation. The specific test languages become known only when the test phase begins and the test data are released, with no labeled training data provided for those languages.

Based on the observations from the AdMIRe 1.0 system reports, where the top-performing submissions consistently relied on large language models (LLMs) — for instance, You et al. (2025) in the bimdoal track and Fan et al. (2025) in the text-only track —- we adopted a similar strategy for AdMIRe 2.0. For better performance, robustness and to reduce model-specific biases, we employed multiple LLMs and aggregated their predictions through a hard voting scheme, which formed our final submission for the text-only track.

Although our official submission focuses solely on the text-only setting — and can be viewed as a relatively direct, shortcut-style application of LLM capabilities — it nevertheless achieved a strong result, ranking second on the official leaderboard.

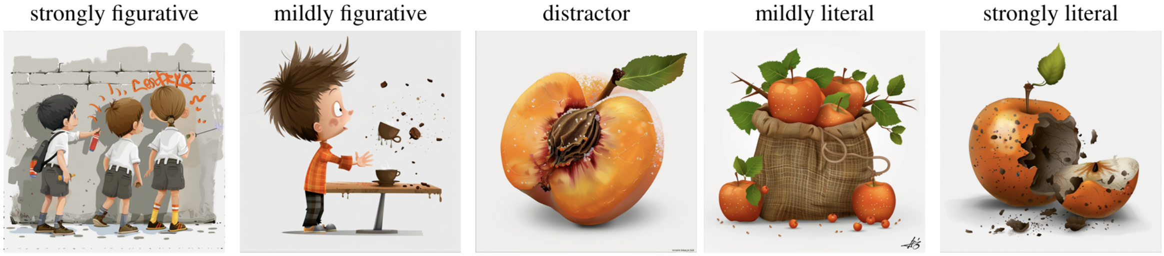
2 Methods and Results

Our method, which is entirely based on prompting, is motivated by both empirical observations from AdMIRe 1.0 and practical considerations specific to AdMIRe 2.0. Given the limited amount of available data[2], the restricted development time, and the fact that AdMIRe 2.0 spans a wide range of languages—including several low-resource ones such as Georgian, Igbo, Kazakh, and Uzbek — we aimed to approach the task in a training-free, zero-shot manner. This was one of the primary motivations for relying on LLMs.

[1] Example source: https://semeval2025-task1.github.io/

[2] Subtask A in AdMIRe 1.0 provides only 70 training instances in English and 32 in Portuguese.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 203–207
March 28, 2026 ©2026 Association for Computational Linguistics



(a) The image depicts three children standing in front of a gray, textured wall...

(b) The image depicts a cartoon-style illustration of a young boy standing at a table...

(c) The image depicts a halved peach with a detailed and realistic appearance...

(d) The image depicts a rustic, burlap sack filled with several bright orange apples...

(e) The image depicts an orange-colored apple that appears to be decomposing or decaying...

Figure 1: Illustration of the Static Image Ranking Task. Consider the NC *bad apple* as in *We have to recognize that this is not the occasional bad apple but a structural, sector-wide problem*, the expected ordering should be (a), (b), (e), (d), and (c).[1]

Our final submission is therefore based entirely on prompting and ensemble learning: we directly provided the textual captions to each model, asked them to perform the ranking, and then combined the predictions from the three LLMs using a hard voting strategy. The captions are cleaned, re-indexed, and concatenated with the prompt prefix below before being fed into the chatbots.

> *Consider the expression [NC] in the [LANG] language, and its meaning in context [SENT]. Rank the following captions based on how well they reflect that meaning. Return only a list of reordered indices, e.g., [1, 3, 2, 4, 5], from most to least similar. Do not include any explanation.*

We found that explicitly indicating the language category in the prompt (i.e., [*LANG*]) has the potential to improve accuracy. This is probably because LLMs internally maintain separate multilingual semantic subspaces. Providing the language label helps the model activate the correct linguistic and cultural knowledge and constrains the reasoning space, thereby improving zero-shot predictions.

Our ensemble consists of three LLMs — GPT-4o (Achiam et al., 2023), Qwen-Plus (Team, 2025), and DeepSeek (Liu et al., 2024). The first two models were selected in that we conducted preliminary experiments on AdMIRe 1.0 Subtask A using a similar prompting strategy and found that their performance was close to that of the top systems reported (see Table 1). A third model was included to facilitates a stable majority-vote ensemble. Also, these models have been referred to in AdMIRe 1.0 system reports, are currently among the most capable and popular multilingual LLMs, and, due to their scale and training diversity, are more likely to provide broad language coverage — probably including for the low-resource languages present in AdMIRe 2.0. The results across all languages are presented in Table 2, which place us in the second position on the leaderboard.

Also, we explored whether incorporating type information (see Table 1 for reference) could provide additional benefits to the final results. To this, we adopted a simple bubble-sorting procedure. Since a PIE used idiomatically in the given context has the opposite usage type of literal, and vice versawe first perform pairwise comparisons to identify the caption most similar to the target usage, followed by the second most likely one. We then conduct another round of pairwise comparisons to determine the caption most similar to the opposite usage and the second most likely one. The remaining caption is treated as the distractor, composing the final ranking.

Due to resource constraints, we only experimented on the Chinese subset using a prompt prefix below:

> *The expression [NC] can be used idiomatically or literally in the LANG language. You are provided with 2 captions, each describing an image that may or may not be related to this expression. Decide which caption most likely relate to*

		Test Set			Extended Evaluation Set		
		Top 1 Acc	DCG	Type Acc	Top 1 Acc	DCG	Type Acc
English	CTYUN-AI	0.64	3.10		0.87	3.51	
	DeepSeek	0.58	3.07	0.89	0.73	3.24	0.80
	GPT-4o	0.59	3.03	0.76	0.60	3.07	0.80
Portuguese	CTYUN-AI	0.92	3.43		0.56	2.97	
	DeepSeek	0.77	3.31	0.77	0.64	3.07	0.76
	GPT-4o	0.77	3.35	0.54	0.42	2.77	0.76

Table 1: Results on the test set and the extended evaluation set of AdMIRe 1.0 Subtask A (text-only track), compared with the best-performing team — CTYUN-AI. Although not required for submission, the AdMIRe tasks expect systems to predict the usage type of a given PIE in its context sentence (*idiomatic* vs. *literal*). Taking the instance in Figure 1 as an example, the system is expected to predict that the PIE *bad apple* in the given sentence is used idiomatically. We therefore also report the type prediction accuracy as *Type Acc* in addition to *Top 1 Acc* (Top Image Accuracy) and *DCG* (NDCG@5). All scores are reported on a scale of 1.

	Top 1 Acc	DCG
Chinese	0.36	0.74
Georgian	0.4	0.74
Greek	0.43	0.76
Igbo	0.33	0.71
Kazakh	0.42	0.76
Norwegian	0.43	0.77
Portuguese-Brazil	0.53	0.81
Portuguese-Portugal	0.45	0.77
Russian	0.51	0.79
Serbian	0.40	0.74
Slovak	0.44	0.78
Slovenian	0.45	0.78
Spanish-Ecuador	0.35	0.73
Turkish	0.40	0.74
Uzbek	0.32	0.73

Table 2: Official results on the test set of AdMIRe 2.0.

> *the [TYPE] meaning or use of this expression. Only output the chosen caption name, such as (A) or (B), do not include any analysis.*

The type information (i.e., *[TYPE]*) is also obtained using a hard-voting strategy. The results are shown in Table 3. Overall, incorporating type information as guidance leads to a slight decrease in performance compared with directly ranking captions using the LLMs (see Table 2). The Top Image Accuracy remains unchanged, while the DCG score drops by 0.01. A closer look at the type-specific scores reveals that performance on literal metrics decreases while on idiomatic metrics improves. This outcome is likely due to the fact that idiomatic usage tends to be more abstract linguistically. Providing explicit type information may therefore help the model identify relevant cues in the captions and align them with the intended usage, therefore improving the accuracy. In contrast, literal interpretations could depend more on visual grounding (again, the *bad apple* example), which is unavailable in the text-only track.

3 Related Work

Studies have shown that language models — ranging from basic BERT to larger generative models such as ChatGPT — continue to exhibit limitations in interpreting idiomatic expressions (IEs) (Shwartz and Dagan, 2019; Wu et al., 2024; Raunak et al., 2023).

Typical solutions to IE representation learns phrase embedding directly from contextual co-occurrence (Mikolov, 2013; Yin and Schütze, 2014, 2016). This is effective for frequent expressions but struggles with sparse IEs. Alternatively, compositional approaches derive phrase embeddings by combining the embeddings of individual components (Mitchell and Lapata, 2010; Yu and Dredze, 2015), but they often fail to capture the semantic opacity characteristic of IEs. More recent work leverages pre-trained language models (PLMs) for IE representation through adaptive modules and contrastive learning, such as Zeng and Bhat (2022); He et al. (2024); Wu et al. (2024). It has also been found that external knowledge, such as synonyms and definitions can enhance model performance (Long et al., 2020; Wang et al., 2020;

	Top 1 Acc	DCG	Literal Acc	Idiomatic Acc	Literal DCG	Idiomatic DCG
1	0.36	0.74	0.46	0.29	0.78	0.70
2	0.36 (0)	0.73 (-0.01)	0.41 (-0.05)	0.33 (+0.04)	0.75 (-0.03)	0.87 (+0.17)

Table 3: Results on the Chinese subset. The first row reports the ranking results obtained directly from LLMs, and the second row presents the results produced by incorporating type information through pairwise comparison. In addition to the overall Top 1 ACC and DCG scores, we also report type-specific performance (*literal* vs. *idiomatic*). The values in parentheses in the second row indicate the increments relative to the first row.

Sha et al., 2023).

Such modelling stratgies have also been noticed in the AdMIRe shared task (1.0), although participation in the text-only track has been relatively limited, likely because images provide stronger cues for the ranking task. Nonetheless, Petersen et al. (2025) fine-tuned SBERT for the task and augmented captions using GPT-4-generated descriptions. The top-performing team (Fan et al., 2025) also applied extensive data augmentation — including synonym substitution and back-translation — for the multilingual setting, and fine-tuned Qwen (Team, 2025) models for the task.

4 Conclusion

This report presents our work on the ADMIRE 2.0 task under the text-only setting. We adopt an ensemble learning framework for the caption ranking task, which places our methods in the second position on the leaderboard. Nevertheless, the results clearly show that ADMIRE 2.0 is substantially more challenging than ADMIRE 1.0, with noticeably lower scores. Even strong LLMs struggle to reliably distinguish literal from idiomatic representations.

Limitations

First, we experimented with only a small subset of available LLMs, which restricts the breadth of our evaluation. Second, our approach relies solely on zero-shot inference combined with a hard-voting ensemble, without exploring more sophisticated or innovative modeling strategies. Due to time and resource constraints, we were unable to investigate alternative architectures, training paradigms, or more creative methods that might further improve performance. We leave these directions for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Yuming Fan, Dongming Yang, Zefeng Cai, and Binghuai Lin. 2025. CTYUN-AI at SemEval-2025 task 1: Learning to rank for idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 16–19, Vienna, Austria. Association for Computational Linguistics.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. *arXiv preprint arXiv:2406.15175*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym knowledge enhanced reader for chinese idiom reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3684–3695.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Wiebke Petersen, Lara Eulenpesch, Ann Piho, Julio Julio, and Victoria Lohner. 2025. Transformer25 at SemEval-2025 task 1: A similarity-based approach. In *Proceedings of the 19th International Workshop on*

Semantic Evaluation (SemEval-2025), pages 2311–2317, Vienna, Austria. Association for Computational Linguistics.

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)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ying Sha, Mingmin Wu, Zhi Zeng, Xing Ge, Zhongqiang Huang, and Huan Wang. 2023. A prompt-based representation individual enhancement method for chinese idiom reading comprehension. In *International Conference on Database Systems for Advanced Applications*, pages 682–698. Springer.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Xinyu Wang, Hongsheng Zhao, Tan Yang, and Hongbo Wang. 2020. Correcting the misuse: A method for the chinese idiom cloze test. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10.

Mingmin Wu, Yuxue Hu, Yongcheng Zhang, Zeng Zhi, Guixin Su, and Ying Sha. 2024. Mitigating idiom inconsistency: A multi-semantic contrastive learning method for chinese idiom reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19243–19251.

Wenpeng Yin and Hinrich Schütze. 2014. An exploration of embeddings for generalized phrases. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 41–47.

Wenpeng Yin and Hinrich Schütze. 2016. Discriminative phrase embedding for paraphrase identification. *arXiv preprint arXiv:1604.00503*.

Runyang You, Xinyue Mei, and Mengyuan Zhou. 2025. PALI-NLP at SemEval 2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1211–1216, Vienna, Austria. Association for Computational Linguistics.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.

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.