ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking

Atakan Site*, **Oğuz Ali Arslan, Gülşen Eryiğit**
Department of Artificial Intelligence and Data Engineering
Istanbul Technical University
{site21, arslanog20, gulsenc}@itu.edu.tr

Abstract

This paper presents our system for AdMIRe 2 (Advancing Multimodal Idiomaticity Representation), a shared task on multilingual multimodal idiom understanding. The task focuses on ranking images according to how well they depict the literal or idiomatic usage of potentially idiomatic expressions (PIEs) in context, across 15 languages and two tracks: a text-only track, and a multimodal track that uses both images and captions. To tackle both tracks, we propose a hybrid zero-shot pipeline built on large vision–language models (LVLMs). Our system employs a chain-of-thought prompting scheme that first classifies each PIE usage as literal or idiomatic and then ranks candidate images by their alignment with the inferred meaning. A primary–fallback routing mechanism increases robustness to safety-filter refusals, while lightweight post-processing recovers consistent rankings from imperfect model outputs. Without any task-specific fine-tuning, our approach achieves 55.9% Top-1 Accuracy in the text-only track and 60.1% in the multimodal (text+image) track, ranking first overall on the official leaderboard. These results suggest that carefully designed zero-shot LVLM pipelines can provide strong baselines for multilingual multimodal idiomaticity benchmarks.

1 Introduction

Idioms constitute a subclass of multi-word expressions (MWEs) and remain a challenging problem even for state-of-the-art large language models (LLMs). The core difficulty stems from the fact that idiomatic meaning is non-compositional and often unpredictable from the constituent words; that is, it cannot be reliably inferred by composing the semantics of the individual words. For instance, the expression "bad apple" rarely refers to a defective piece of fruit; instead, it typically denotes a person whose negative behavior can corrupt or undermine a group. Idioms can also introduce ambiguity between a literal reading suggested by their surface form and the intended idiomatic interpretation. These properties make idioms a valuable probing ground for examining how NLP models represent and compose meaning.

In recent years, progress has been made toward modeling idiomatic meaning (Umut et al., 2025; Kim et al., 2025; Khoshtab et al., 2025). Nevertheless, language models still struggle with figurative and abstract meaning, often failing to go beyond surface-level lexical cues and relying on shallow lexical associations (Mi et al., 2025; Leon et al., 2025). This weakness has practical implications for downstream tasks that require meaning beyond straightforward compositional semantics. In particular, failures in idiom understanding can lead to incorrect reasoning in natural language inference (Stowe et al., 2022), mismatches in retrieval and semantic similarity (Tayyar Madabushi et al., 2022), and erroneous decisions in question-answering systems (Rakshit and Flanigan, 2022). Improving idiom understanding is therefore an important step toward more robust language understanding, motivating targeted benchmarks and analyses of idiomatic language to better assess model generalization.

Idiom processing has been evaluated with idiom-specific datasets and benchmarks at both token and sentence level, covering idiomaticity detection and representation learning (Cook et al., 2008; Haagsma et al., 2020; Saxena and Paul, 2020; Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). More recently, some datasets have incorporated visual information to create multimodal evaluation settings, revealing that grounded figurative and idiomatic understanding can be substantially more challenging for LVLMs than text-only benchmarks (Saakyan et al., 2025; Yosef et al., 2023). SemEval-2025 Task 1 (AdMIRe: Advancing Multimodal Idiomaticity Representation) frames idiomaticity through image-based meaning representation and

*Corresponding author.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 226–236

prediction, providing a benchmark for grounding idiom interpretation beyond text-only cues (Pickard et al., 2025). Building on this foundation, AdMIRe 2 further expands the evaluation to a broader multilingual and multimodal setting centered on potentially idiomatic expressions (PIEs). The shared task is run in two tracks, a text-only track and an image+text track, which enables direct measurement of how visual grounding affects literal–idiomatic disambiguation (Arslan et al., 2026).

In this paper, we propose a hybrid zero-shot system for both tracks that combines two LVLMs through a primary–fallback mechanism and uses efficient prompting strategies to produce robust image rankings.

We show that our proposed system ranks first on the official leaderboard in terms of average performance across tracks. Our pipeline achieves 55.9% accuracy in the text-only track and 60.1% accuracy in the multimodal (text+image) track, without requiring any task-specific fine-tuning.

All the code is available on our GitHub[1].

2 Background

This section reviews prior work on LLMs and LVLMs, with a particular focus on their applications to idiom processing and figurative language understanding.

Early work on idioms and PIEs predates large-scale LLMs and focuses on building dedicated resources and supervised models. Classical idiom datasets target token-level usage labeling and sentence-level idiomaticity, and have shown that idioms are systematically harder for distributional models than compositional expressions (Cook et al., 2008; Saxena and Paul, 2020). More recent corpora such as MAGPIE scale this line of work up to large, annotated collections of PIEs drawn from the British National Corpus, with tens of thousands of instances labeled as literal or idiomatic across more than 1,700 expressions (Haagsma et al., 2020). Other idiom-focused resources extend this direction to naturally occurring English and Portuguese sentences containing multiword expressions, annotated with fine-grained sense labels to evaluate idiom usage detection and sentence-level representation learning (Madabushi et al., 2021). SemEval-2022 Task 2 further consolidates this direction by casting multilingual idiomaticity detection and idiom-aware sentence embeddings as a shared task in English, Portuguese, and Galician, and by providing a common evaluation protocol for idiom-sensitive representations (Tayyar Madabushi et al., 2022). In parallel, large language models have also been explored as tools for idiom corpus construction, generating synthetic idiom corpora across multiple languages and assessing their value for idiomaticity detection (Arslan et al., 2025). Taken together, these efforts establish idiom and PIE processing primarily as a supervised, text-only classification and representation problem, and highlight both the usefulness of idiom-focused resources and the difficulty of encoding idiomatic meaning even for strong pretrained transformers.

With the emergence of LLMs, idiom processing has increasingly been revisited through prompt-based evaluation, particularly in zero-shot and few-shot settings. Recent work constructs controlled contrastive datasets of minimal sentence pairs where the same idiom is used in either a literal or a figurative context, and shows that LLMs often fail precisely when disambiguation requires careful use of contextual cues rather than surface-level associations (Mi et al., 2025). Another line of work investigates LLMs as classifiers for multiword expressions and PIEs, finding that carefully engineered prompts can match supervised baselines on some idiom and MWE identification benchmarks, but that performance does not generalize reliably across datasets and is highly sensitive to annotation choices (Hashiloni et al., 2025). Complementary work evaluates conversational LLMs on challenging idiom detection test suites and reports systematic errors, including over-predicting idiomatic readings in literal contexts and difficulty dealing with polysemous expressions (De Luca Fornaciari et al., 2024). Overall, these studies show that, despite notable progress, state-of-the-art LLMs still rely on shallow lexical cues and frequency statistics when processing idioms and PIEs, and often fail to recover the intended non-compositional meaning from context.

Figurative language has also been studied more broadly, beyond idioms, as a testbed for LLMs' semantic and reasoning capabilities. A natural language inference benchmark for figurative language frames the task as recognizing entailment between around nine thousand premise–hypothesis pairs covering various figurative phenomena, each annotated with an entailment label and a human-written explanation (Chakrabarty et al., 2022). Experiments on such benchmarks show that even

[1] https://github.com/oguzaliarslan/idiom-nlp

strong sequence-to-sequence models fine-tuned on the data exhibit substantial gaps in both prediction accuracy and explanation quality, indicating that figurative language remains challenging even in text-only settings.

More recently, research has begun to explore figurative language in multimodal settings, where images provide additional grounding. The IRFL dataset pairs idioms, metaphors and similes with both figurative and literal candidate images and defines recognition tasks that require models to identify which image best reflects the figurative meaning (Yosef et al., 2023). State-of-the-art LVLMs achieve only around 22% accuracy on IRFL, compared to 97% for humans, underscoring the difficulty of multimodal figurative understanding. V-FLUTE extends this direction by framing visual figurative language understanding as an explainable visual entailment task, covering metaphors, similes, idioms, sarcasm, and humor. Given an image and a caption, a model must decide whether the image entails the caption and provide a textual explanation (Saakyan et al., 2025). Experiments on V-FLUTE reveal that LVLMs struggle to generalize from literal to figurative meaning, particularly when figurative cues are primarily present in the visual modality, and often produce hallucinated or incomplete explanations.

Within this broader figurative-language landscape, multimodal idiom understanding has only recently become a dedicated research focus. IRFL includes an idiom subset, but treats idioms alongside other figurative phenomena in a unified recognition setting, without explicitly modeling potentially idiomatic expressions or literal–idiomatic ambiguity. In contrast, SemEval-2025 Task 1: AdMIRe (Advancing Multimodal Idiomaticity Representation) offers a more targeted benchmark centered on idioms and PIEs in multilingual, multimodal contexts (Pickard et al., 2025). AdMIRe introduces datasets where nominal compounds with both literal and idiomatic readings are embedded in context sentences and paired with images generated to depict either the literal or idiomatic interpretation, across English and Brazilian Portuguese. The task description reports that the strongest participating systems rely on mixtures of pre-trained LLMs and LVLMs, multi-query prompting, and reranking strategies, yet performance still varies considerably across languages, idiom types and sense (literal vs idiomatic), indicating that robust multimodal idiom grounding remains an open challenge.

Participant papers from AdMIRe and related multimodal shared tasks broadly converge on a similar pattern: rather than training models from scratch, systems typically start from powerful proprietary or open-source LVLMs and focus on designing task-specific prompts, scoring functions and ensembling schemes for each benchmark (You et al., 2025). This line of work demonstrates that careful prompt engineering can substantially improve idiom-related performance, but it also highlights the engineering cost and limited generality of heavily task-tuned pipelines. Based on these studies, our work investigates how far a purely zero-shot LVLM-based system can go on AdMIRe 2, using efficient prompting strategies to address both the text-only and image+text tracks. By directly comparing performance across tracks in a shared multilingual PIE setting, we provide complementary evidence on the strengths and remaining limitations of LVLMs for grounded idiomaticity.

3 Methodology

3.1 Proposed Architecture

We propose a zero-shot inference pipeline designed to rank associated images based on the literal or idiomatic usage of PIEs. Our approach utilizes the reasoning capabilities of the state-of-the-art LVLMs through a structured chain-of-thought (CoT) prompting strategy.

As depicted in Figure 1, our architecture leverages both candidate images and their captions to support two tracks: **text-only** and **multimodal** (image + text). In the text-only track, captions proxy visual content, whereas the multimodal track additionally allows reasoning over visual cues absent from text.

3.2 Chain-of-Thought Prompting

Our system employs a five-step CoT prompting strategy[2] that guides the model through explicit reasoning stages before producing a final image ranking.

Step 1: Usage Type Classification First, we analyze the context sentence to determine whether the PIE is used literally or idiomatically. To achieve this, we provide explicit definitions for both categories: literal usage describes physical, photographable scenarios, whereas idiomatic us-

[2]Complete prompt templates for each step are provided in Appendix A.

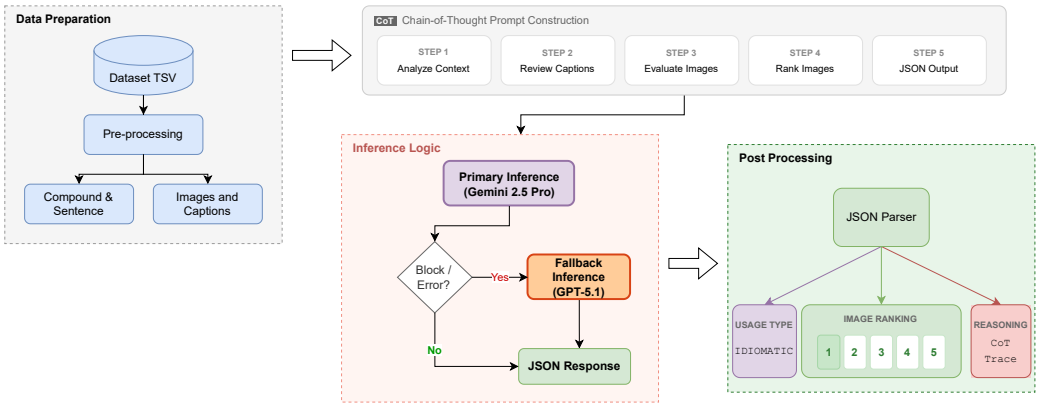

Figure 1: Overview of the proposed system. The system takes a PIE, context sentence, and five candidate images (with captions) as input. A chain-of-thought prompting strategy guides the model through usage classification, image evaluation, and ranking stages to produce the final output.

age conveys figurative meaning distinct from the surface-level interpretation.

Step 2: Reviewing Images and Captions The model examines the five candidate images to understand what each represents. For the text-only track, we analyze the provided image captions to interpret the scenery. In the multimodal track, both the raw images and their corresponding captions are passed to the model.

Step 3: Image Evaluation Following the review of images and captions, the model evaluates how well each of the five candidate images represents the usage type (literal or idiomatic) identified in Step 1. A critical design choice is to prioritize evaluation of how well each image represents the expression's literal or idiomatic meaning, rather than how closely it matches the context sentence. For instance, if the expression "night owl" is used idiomatically in the context sentence, the model is instructed to prefer images that depict the figurative meaning over images that match the literal description of the PIE. For each candidate image, the model generates a quality rating, supported by a rationale that quotes the cues and information utilized for the inference.

Step 4: Image Ranking Based on the evaluations from Step 3, the model ranks all five candidate images from best to worst according to how well they represent the identified meaning of the PIE. We instruct the model to compare lower-ranked candidates with higher-ranked ones, using the ratings from Step 3, to ensure that the ranking is grounded in the preceding analysis rather than generated arbitrarily.

Step 5: Output Generation Finally, the model produces a structured JSON output containing all components of its analysis. The output includes the identified usage type, the reasoning behind this classification, evaluations for each candidate image, and the final ranking of the candidate images with supporting statements.

3.3 Model Inference Logic

Our inference pipeline implements a hybrid LVLM architecture that routes instances between two models via a primary-fallback mechanism. We use Gemini 2.5 Pro (Comanici et al., 2025) as our primary model and GPT-5.1 (OpenAI, 2025) as the fallback.

For each PIE in the dataset, we construct the CoT prompt and send it to the primary model. A major challenge while processing PIEs across different languages is that certain expressions can be flagged by the API's safety filters, particularly idioms involving sensitive terms. We monitor API responses for refusal signals and route to a fallback model when the primary model refuses. This ensures valid rankings for every instance without manual intervention.

3.4 Post Processing

In this stage, we parse the model's JSON output directly, with a regex-based fallback to extract rankings when parsing fails due to syntax errors. This ensures valid rankings even from malformed outputs.

3.5 Evaluation

The system's performance on both tracks was evaluated using two official metrics: Top-1 Accuracy, measuring whether the gold-standard best image is ranked first, and NDCG@5 with relevance weights [3, 1, 0, 0, 0], capturing overall ranking quality. Leaderboard rankings were determined by average Top-1 Accuracy across languages.

4 Experiments

4.1 Dataset

The AdMIRe 2 shared task provides a dataset covering PIEs across 15 diverse languages, listed in Table 2. The construction and statistics of this multilingual resource are described in the accompanying dataset paper (Torunoğlu-Selamet et al., 2026). Each data instance consists of a PIE, a context sentence in which the PIE is used, and five candidate images along with their generated captions.

4.2 Text-only Track

For the text-only track, Table 1 presents our system's performance across all 15 languages. Our approach achieves 55.9% average Top-1 accuracy, with a NDCG@5 of 0.831. Brazilian Portuguese leads with 78.9% accuracy, followed by Slovenian 72.5% and Russian 65.0%, while Ecuadorian Spanish presents the most challenging case at only 25.0% accuracy.

4.3 Multimodal Track

The right half of Table 1 presents our hybrid system's performance on the multimodal track for all languages. Incorporating images leads to clear gains: average Top-1 Accuracy improves to 60.1% and NDCG@5 increases to 0.849. Overall, our system shows 4.2% better performance on multimodal track.

The benefit of visual information is especially observed in several languages. Turkish shows the largest improvement, from 50.5% to 67.8%, followed by Brazilian Portuguese, Serbian, Norwegian, Slovak, and Slovenian. Even for the hardest language, Ecuadorian Spanish, multimodality yields a modest gain.

4.4 Literal vs. Idiomatic Asymmetry

Across both tracks, we observe a consistent asymmetry between literal and idiomatic usages. In the text-only setting, our system attains 61.1% average accuracy on literal uses, compared to 51.3% on idiomatic ones, indicating that idiomatic readings are substantially harder to capture. When images are added, overall performance improves for both usage types: average accuracy increases to 66.9% for literal cases and 54.8% for idiomatic ones. However, the literal–idiomatic gap does not disappear; in fact, it slightly widens, suggesting that current LVLMs leverage visual cues more effectively for concrete, photographable meanings than for abstract figurative interpretations. This pattern is particularly evident in languages such as Greek and Norwegian, where literal cases consistently dominate idiomatic ones, and aligns with prior findings that idioms remain challenging even for strong multimodal models.

4.5 Comparison between Text-only and Multimodal Track

Comparing columns in Table 1, 14 out of 15 languages benefit from adding images, confirming that visual grounding generally helps the model align PIE interpretations with the correct image. Notably, Igbo is the only exception where text-only setting outperforms the multimodal setting by 4%, which may be due to noisier captions or weaker image-text alignment for this language. This suggests that while multimodal information is beneficial in most cases, its impact can be uneven across languages and data conditions.

4.6 Model Comparison

To understand the contribution and performance of the underlying LVLMs, we analyze the performance of our primary model against the fallback model. Detailed breakdowns are provided in Table 3 (text-only) and Table 4 (text + image) in the Appendix.

Gemini 2.5 Pro outperforms GPT-5.1 across both tracks, achieving 55.9% vs. 55.0% on text-only and 59.8% vs. 57.3% on the multimodal track. In the text-only track, the two models perform similarly, with less than one percentage point difference, demonstrating the strong linguistic capabilities of both models. However, the gap widens considerably when images are included, showing Gemini 2.5 Pro's stronger vision-language capacity. Gemini gains 3.9% points when moving from text-only to text+image, compared to GPT-5.1's 2.3% improvement. This disparity is most pronounced in Turkish, where Gemini's accuracy jumps from 50.5% to 67.6% while GPT-5.1 shows a more modest increase from 52.7% to 58.2%.

	Text-Only Track						Multimodal Track					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	.458	.556	.378	.774	.806	.749	**.497**	**.568**	**.439**	**.799**	**.811**	**.789**
Georgian	.513	.600	.444	.791	.828	.762	**.531**	**.620**	**.460**	**.808**	**.840**	**.783**
Greek	.591	.702	.481	.856	.917	.795	**.639**	**.740**	**.538**	**.874**	**.925**	**.822**
Igbo	**.478**	**.606**	**.427**	**.784**	.836	**.764**	.435	.545	.390	.777	**.843**	.751
Kazakh	.603	.593	**.608**	.838	.849	**.833**	**.609**	**.667**	.578	**.844**	**.879**	.825
Norwegian	.614	.780	.451	.853	.909	.799	**.673**	**.790**	**.559**	**.881**	**.922**	**.841**
Portuguese (Brazil)	.789	.860	.719	.917	.951	.884	**.855**	**.868**	**.842**	**.939**	**.956**	**.922**
Portuguese (Portugal)	.618	.670	**.570**	.863	.872	**.855**	**.641**	**.717**	**.570**	**.865**	**.882**	.848
Russian	.650	.742	**.577**	.871	.903	**.846**	**.686**	**.855**	.551	**.891**	**.947**	**.846**
Serbian	.551	.599	**.505**	.819	.849	**.792**	**.617**	**.740**	.500	**.837**	**.897**	.779
Slovak	.543	.691	.422	.836	.889	.793	**.596**	**.721**	**.494**	**.854**	**.902**	**.815**
Slovenian	.725	.792	.658	.893	.925	.860	**.779**	**.833**	**.725**	**.911**	**.940**	**.882**
Spanish (Ecuador)	.250	.045	**.423**	.723	**.658**	.777	**.271**	**.091**	**.423**	**.726**	.650	**.791**
Turkish	.505	.514	.500	.823	.826	.821	**.676**	**.722**	**.645**	**.895**	**.912**	**.884**
Uzbek	.500	.417	**.536**	.816	.811	.818	**.517**	**.556**	.500	**.834**	**.855**	**.824**
Average	**.559**	**.611**	**.513**	**.831**	**.855**	**.810**	**.601**	**.669**	**.548**	**.849**	**.877**	**.827**

Table 1: Results across both tracks for 15 languages. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id). The multimodal track outperforms text-only by 4.2% in overall accuracy.

5 Conclusion

In this paper, we presented the solution developed by the ITUNLP group for the AdMIRe 2 shared task on multilingual multimodal idiom understanding. The proposed approach tackles the task in a purely zero-shot setting. Our system yields promising results, achieving 1^{st} place on the official leaderboard in terms of average performance across both tracks.

For future work, we plan to evaluate open-source LVLMs within our proposed system to gain a broader perspective on model performance. Furthermore, since the AdMIRe 2 dataset currently covers only 15 languages, we aim to apply our system to additional languages and datasets, further expanding its applicability and testing its robustness.

References

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, and Joakim Nivre. 2025. Using LLMs to advance idiom corpus construction. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

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.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23782–23801, Suzhou, China. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710, Suzhou, China. Association for Computational Linguistics.

Frances Laureano De Leon, Harish Tayyar Madabushi, and Mark G. Lee. 2025. Evaluating large language models on multiword expressions in multilingual and code-switched contexts. *Preprint*, arXiv:2504.20051.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *Preprint*, arXiv:2109.04413.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5.1 instant and gpt-5.1 thinking system card. `https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf`. Technical report.

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.

Geetanjali Rakshit and Jeffrey Flanigan. 2022. FigurativeQA: A test benchmark for figurativeness comprehension for question answering. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 160–166, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. *Preprint*, arXiv:2405.01474.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. *Preprint*, arXiv:2006.09479.

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, Dublin, Ireland. 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, Seattle, United States. Association for Computational Linguistics.

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, Seattle, United States. Association for Computational Linguistics.

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.

Özge Umut, Atakan Site, Doğukan Arslan, and Gülşen Eryiğit. 2025. Exploring Turkish idiomaticity with large language models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 533–538.

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, Singapore. Association for Computational Linguistics.

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.

Appendix

A Generation Prompts

A.1 Task Setup and Context

Research Context & Task

You are an expert linguist conducting academic research on potentially idiomatic expressions (PIEs) for a linguistic analysis study. This is a scholarly research task analyzing language usage patterns.
Research Context: Academic linguistic analysis of expression usage patterns.
Expression: `{compound}`
Context Sentence: `{sentence}`
Research Task:

1. Determine if `{compound}` is used LITERALLY or IDIOMATICALLY in this sentence.
2. Rank 5 images by how well they represent the LITERAL MEANING or IDIOMATIC MEANING of `{compound}` (based on Step 1).

CRITICAL: Rank by how well images represent the literal/idiomatic MEANING of the expression, NOT by how well they match the specific context sentence. If it's literal, rank images closest to the literal meaning highest. If it's idiomatic, rank images closest to the idiomatic meaning highest. You must work through this problem step-by-step. Complete each step fully before moving to the next.

A.2 Chain-of-Thought Reasoning Steps

Step 1: Analyze Context and Determine Usage Type

Determine whether `{compound}` is used **LITERALLY** or **IDIOMATICALLY** in this sentence.

- **LITERAL:** Words describe their actual physical meaning; you could photograph exactly what they describe (e.g., "night owl" = actual owl).
- **IDIOMATIC:** Expression has figurative meaning different from literal interpretation; describes abstract concepts (e.g., "night owl" = person who stays up late).

Output: State "LITERAL" or "IDIOMATIC" with 2-3 sentence reasoning referencing specific context clues.

Step 2: Review Image Captions

Below are the 5 image captions. Examine the caption text to understand what each image represents:
[System Note: Image captions inserted here dynamically]

- **Image 1 (`{img_name}`):** Caption: "`{caption}`"
- ...

Note what each caption describes and how it relates to literal vs figurative representation.

Step 3: Evaluate Each Image Against Usage Type

For each image (based on its caption), evaluate how well it represents the LITERAL MEANING or IDIOMATIC MEANING of `{compound}` (based on Step 1).
CRITICAL: Evaluate how well the image (based on its caption) represents the literal/idiomatic MEANING of the expression, NOT how well it matches the specific context sentence.
Evaluation criteria:

- **If LITERAL:** Does the caption suggest the image shows the physical/literal meaning of `{compound}`? Rank images that best represent what `{compound}` literally means (the actual physical thing/action), regardless of whether they match the specific context scene.
- **If IDIOMATIC:** Does the caption suggest the image represents the figurative/idiomatic meaning of `{compound}`? Rank images that best represent the idiomatic meaning (the abstract concept), regardless of whether they match the specific context.

For each image (4-5 sentences):

- Quote specific caption phrases
- Explain how well the caption suggests the image represents the literal/idiomatic MEANING of `{compound}`
- Do NOT evaluate based on context matching - only evaluate meaning representation
- Match quality: EXCELLENT / GOOD / MODERATE / POOR / VERY POOR

Step 4: Rank Images

Rank all 5 images from BEST to WORST based on how well their captions suggest they represent the literal/idiomatic MEANING of `{compound}`.
CRITICAL RANKING RULE: Rank by meaning representation, NOT context matching:

- **If LITERAL:** Rank images that best represent the LITERAL MEANING of `{compound}` highest (captions suggesting the actual physical thing/action). The top 2 images should be closest to the literal meaning, regardless of context.
- **If IDIOMATIC:** Rank images that best represent the IDIOMATIC MEANING of `{compound}` highest (captions suggesting the figurative/abstract concept). The top 2 images should be closest to the idiomatic meaning, regardless of context.

For #1 image (5-6 sentences):

- Start by restating WHY the expression is literal/idiomatic (from Step 1)
- Quote caption phrases
- Explain why this image's caption best represents the literal/idiomatic MEANING of `{compound}`
- Focus on meaning representation, NOT context matching
- Briefly compare to lower-ranked images

A.3 Final Output Specification

Step 5: Final Output (JSON)

Provide your complete analysis as VALID JSON. Ensure all 5 images are included in the ranking. **CRITICAL:** Use the ACTUAL image filenames listed above, NOT placeholder names like "image1.png".

JSON structure (use actual filenames from the images listed above):

```
{
  "usage_type": "literal" or "idiomatic",
  "usage_reasoning": "2-3 sentences with context clues",
  "image_evaluations": {
      "{image_names[0]}": "EXCELLENT - quote caption...",
      "{image_names[1]}": "GOOD - quote caption...",
      "{image_names[2]}": "MODERATE - quote caption",
      "{image_names[3]}": "POOR - quote caption",
      "{image_names[4]}": "VERY POOR - quote caption"
  },
  "reasoning": "5-6 sentences: restate WHY literal/idiomatic...",
  "ranking": ["{image_names[0]}", "{image_names[1]}", "{image_names[2]}",
              "{image_names[3]}", "{image_names[4]}"]
}
```

Requirements:

- Use the EXACT image filenames from Step 2 (e.g., `{image_names[0]}`)
- Quote captions, connect to usage type
- Escape quotes as
  " in JSON strings
- Output ONLY JSON.

B Language Codes

Table 2: List of languages and their corresponding codes.

Language	Code	Language	Code
Chinese	zh	Russian	ru
Georgian	ka	Serbian	sr
Greek	el	Slovak	sk
Igbo	ig	Slovenian	sl
Kazakh	kk	Spanish (Ecuador)	es-EC
Norwegian	no	Turkish	tr
Portuguese (Brazil)	pt-BR	Uzbek	uz
Portuguese (Port.)	pt-PT		

C Detailed Per-Language Results on Text-Only and Multimodal Tracks

D Inference Parameters

	Gemini 2.5 Pro						GPT 5.1					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	**.458**	**.556**	.378	.774	**.806**	.749	.436	.444	**.429**	**.776**	.774	**.778**
Georgian	.513	.600	.444	.791	.828	.762	**.531**	.600	**.476**	**.804**	**.834**	**.781**
Greek	**.591**	**.702**	.481	**.856**	**.917**	.795	**.591**	.673	**.510**	.848	.896	**.800**
Igbo	**.478**	**.606**	**.427**	**.784**	**.836**	**.764**	.391	.454	.365	.766	.824	.742
Kazakh	**.603**	.593	**.608**	.838	.849	.833	.577	**.630**	.549	**.850**	**.880**	**.834**
Norwegian	**.614**	**.780**	**.451**	**.853**	**.909**	**.799**	.554	.700	.412	.838	.892	.785
Portuguese (Brazil)	**.789**	**.860**	**.719**	**.917**	**.951**	**.884**	.741	.816	.667	.905	.946	.865
Portuguese (Portugal)	**.618**	.670	**.570**	**.863**	.872	**.855**	.595	**.717**	.482	.849	**.902**	.800
Russian	.650	.742	.577	.871	.903	.846	**.664**	**.758**	**.590**	**.876**	**.913**	**.848**
Serbian	**.551**	**.599**	**.505**	**.819**	**.849**	**.792**	.543	.588	.500	.809	.842	.777
Slovak	**.543**	**.691**	.422	**.836**	**.889**	.793	.530	.574	**.494**	.824	.840	**.811**
Slovenian	**.725**	**.792**	.658	.893	**.925**	.860	.704	.742	**.667**	**.894**	.915	**.873**
Spanish (Ecuador)	.250	.045	.423	.723	.658	.777	**.354**	**.090**	**.576**	**.739**	**.661**	**.806**
Turkish	.505	.514	.500	.823	.826	.821	**.527**	**.556**	**.509**	**.826**	**.831**	**.823**
Uzbek	.500	.417	.536	.816	**.811**	.818	**.508**	.417	**.548**	.816	.803	**.822**
Average	**.559**	**.611**	.513	**.831**	**.855**	**.810**	.550	.584	**.518**	.829	**.855**	.817

Table 3: Detailed performance comparison of Gemini 2.5 Pro and GPT 5.1 on the **text only track**. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id).

	Gemini 2.5 Pro						GPT 5.1					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	.480	**.617**	.367	.789	**.825**	.759	**.497**	.568	**.439**	**.799**	.811	**.789**
Georgian	**.531**	**.620**	**.460**	**.808**	**.840**	**.783**	.469	.560	.397	.792	.835	.757
Greek	**.639**	.740	**.538**	**.874**	.925	**.822**	.635	**.760**	.510	.868	**.930**	.807
Igbo	.426	.485	**.402**	.777	.828	**.757**	**.435**	**.545**	.390	.777	**.843**	.751
Kazakh	**.609**	.667	**.578**	**.844**	.879	**.825**	.564	**.722**	.480	.838	**.915**	.797
Norwegian	**.673**	**.790**	**.559**	**.881**	**.922**	.841	.658	.770	.549	.874	.906	**.844**
Portuguese (Brazil)	**.855**	.868	**.842**	**.939**	.956	**.922**	**.855**	**.921**	.789	.938	**.969**	.907
Portuguese (Portugal)	**.641**	**.717**	**.570**	**.865**	.882	**.848**	.614	.708	.526	.856	**.893**	.822
Russian	**.686**	**.855**	**.551**	**.891**	**.947**	.846	.650	.790	.538	.871	.931	.823
Serbian	.595	.661	**.532**	.829	.867	**.792**	**.617**	**.740**	.500	**.837**	**.897**	.779
Slovak	**.596**	**.721**	.494	**.854**	**.902**	**.815**	.583	.676	**.506**	.844	.897	.800
Slovenian	**.779**	**.833**	**.725**	**.911**	**.940**	**.882**	.758	.817	.700	.907	.933	.880
Spanish (Ecuador)	**.271**	**.091**	**.423**	**.726**	**.650**	**.791**	.208	.045	.346	.721	.649	.782
Turkish	**.676**	**.722**	**.645**	**.895**	**.912**	**.884**	.582	.681	.518	.835	.874	.810
Uzbek	**.517**	**.556**	**.500**	**.834**	**.855**	**.824**	.475	.444	.488	.815	.824	.811
Average	**.598**	**.663**	**.539**	**.848**	**.875**	**.826**	.573	.650	.512	.838	.874	.811

Table 4: Detailed performance comparison of Gemini 2.5 Pro and GPT 5.1 on the **text+image track**. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id).

Model	**Max Output Tokens**	**Temperature**	**Reasoning Effort**
Gemini 2.5 Pro	8,192	1.0	Default
GPT-5.1 (fallback)	2,500	Default	Medium

Table 5: Inference parameters used for the primary and fallback models.