DCSN-NLP at MWE-2026 AdMIRe 2: Bridging Literal and Figurative Meaning Through Hierarchical Multimodal Reasoning

David Cotigă* **and Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
cotigadavid@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

This paper presents our system for the MWE-2026 AdMIRe 2.0 shared task, which aimed to advance multimodal idiomatic understanding across 15 languages. We address the task of selecting, from a set of five images, the one that best represents either the literal or idiomatic meaning of a given compound in context. Our approach follows a multi-step pipeline: a large language model (LLM) first determines whether the compound is used literally or idiomatically and generates auxiliary text, consisting of an idiomatic meaning explanation and a visual description of the literal meaning. An ensemble of three CLIP models then identifies the two images most semantically similar to the appropriate generated text via a voting mechanism. Finally, the LLM selects the best image from these two candidates.

1 Introduction

Idiomatic expressions pose a fundamental challenge for language understanding systems because their meanings are not compositionally derivable from their constituent words (Dankers et al., 2022). Rather than being inferable from surface lexical cues, idioms often encode abstract, culturally grounded semantics that require contextual reasoning and access to shared world knowledge (Sag et al., 2002).

For large language models (LLMs), this difficulty is further amplified by inherent ambiguity and distributional bias in training data. Because many idioms occur predominantly in their figurative sense, models tend to strongly associate a given compound with its idiomatic meaning, often underrepresenting or overlooking its literal interpretation.

For example, when prompting GPT-5 (Singh et al., 2025) with the sentence "Some cat's eyes were installed on the newly built road for the safety of the drivers" and asking whether "cat's eyes" is used literally or idiomatically, the model incorrectly interprets it as literal, despite the term's metaphorical usage referring to reflective road studs.

The AdMIRe 2.0 task is designed to test these limitations by evaluating models' ability to understand idiomatic expressions across both linguistic and visual modalities. Given an idiomatic compound, a contextual sentence, and a set of candidate images, systems must identify the image that best matches the intended interpretation of the expression (Arslan et al., 2026).

To address this challenge, we propose a multi-step multimodal disambiguation pipeline that combines linguistic reasoning with structured visual grounding. At a high level, our approach first determines whether an expression is used idiomatically or literally in context, then leverages contrastive text–image representations to progressively narrow the space of candidate interpretations before making a final decision.

2 Task and Dataset

The first edition of the task (Pickard et al., 2025) evaluated systems on a dataset comprising two languages: English and Brazilian Portuguese. The second edition significantly expands this scope, introducing a substantially larger multilingual dataset covering 15 different languages (Torunoğlu-Selamet et al., 2026) which are composed as shown in Appendix C.

Our pipeline was originally developed and validated using the English subset from the first edition. Additionally, at the time of writing, ground-truth annotations for the second edition dataset have not yet been released. For these reasons, we restrict our experiments to the English dataset and focus exclusively on this language for the remainder of the paper.

*Corresponding authors.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 217–225

The English dataset is divided into training, development, test, and extended evaluation subsets, as shown in Table 1. The extended evaluation subset is formed by concatenating the 3 subsets previously mentioned, using different contextual sentences.

Data	# instances
English Train	70
English Dev	15
English Test	15
English Extended	100

Table 1: Dataset statistics for the English language

For each item, the system receives an idiomatic expression (e.g. "bad apple"), a contextual sentence (e.g."The team's efforts were spoiled by one bad apple") and a set of five images (see Figure 1). The system must choose the image which best represents the literal or the idiomatic meaning, depending on which is used in the contextual sentence. In this example, the expected response is the first image. Each set of five images is constructed according to a fixed blueprint: one image that strongly reflects the idiomatic meaning, one that weakly reflects the idiomatic meaning, one image that strongly reflects the literal meaning, one that weakly reflects the literal meaning, and one image that is unrelated to either interpretation. In what follows, these images are referred to as strong idiomatic, weak idiomatic, strong literal, weak literal, and distractor, respectively.

3 Pipeline Overview

This section provides a high-level overview of the proposed multimodal disambiguation pipeline. First, a LLM is provided with a compound expression and its surrounding sentence and tasked with classifying the usage as either idiomatic or literal. Next, for each compound, two textual representations are generated: one explicitly describing the idiomatic meaning, and another describing the literal meaning in visual terms (e.g., for "love triangle": "three hearts connected at the corners, forming a triangle").

These textual descriptions, together with the set of five candidate images, are then processed by three distinct CLIP-based models (Radford et al., 2021) to produce joint text–image embeddings. For each image, similarity scores with respect to both textual descriptions are computed, and an ensemble voting mechanism selects the two most semantically aligned candidates. Finally, the LLM performs a comparison between these two finalists and makes the ultimate disambiguation decision (see Figure 2).

Conceptually, our system can be viewed as a hierarchy in which the LLM serves as the final decision-maker, while the preceding multimodal stages act as a structured filtering process that progressively reduces the ambiguity space to a minimal set of competing interpretations.

4 System Description

4.1 Baselines

First, it is important to acknowledge both the capabilities and limitations of large language models in the context of multimodal idiomatic understanding. When prompted (see Appendix A) to directly select the correct image from a set of five candidates, OpenAI's GPT-4o (OpenAI et al., 2024) achieves an accuracy of approximately 75%. These findings align with results presented in the first edition of the task (Alfter, 2025), where an accuracy of up to 81% was obtained using an LLM-only approach augmented with additional idiom detection and explanation steps.

When it comes to human evaluation, the paper describing the first edition of the task mentions the performance and evaluation method of human annotators, who averaged a precision of 71%, with the best scoring individual obtaining a score of 86%.

4.2 Binary Classification

As a first step, we use the LLM to classify each expression-context pair as either idiomatic or literal. We employ a reasoning-oriented prompting strategy that provides the model with specific evaluation criteria to ground its decision, while strictly constraining the generated output to a single binary label. (The exact prompts are presented in Appendix A). This step is critical, as the resulting classification determines which textual representation is used in the subsequent multimodal similarity stages.

Table 2 reports the classification accuracies obtained using two large language models, GPT-5 and GPT-4o, under two prompting strategies: simple prompting and the reasoning-oriented prompting strategy. For the stronger model (GPT-5), this prompting strategy yields only a modest improvement, as the model already performs near ceiling

Strongly idiomatic | Weakly idiomatic | Distractor | Weakly literal | Strongly literal

Figure 1: Data example for *bad apple*

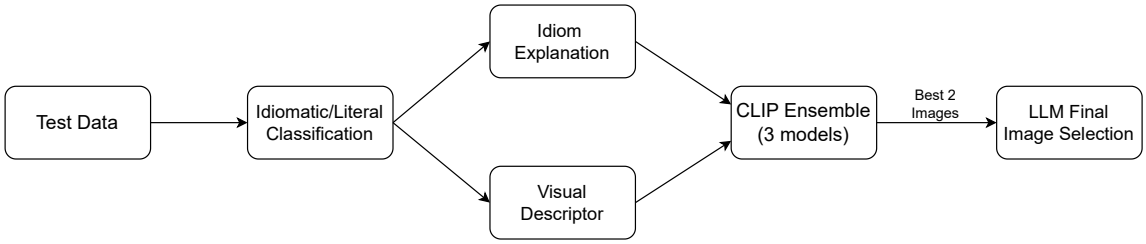

Figure 2: Overview of the proposed multimodal pipeline.

with simple prompting (96%), reaching 99% when guided reasoning is encouraged. In contrast, for the less capable GPT-4o model, the same prompting strategy produces a substantial performance gain, increasing accuracy from 91% to 99%. This effectively brings GPT-4o to near-perfect performance and largely closes the performance gap to GPT-5.

Model	Reasoning	Simple
gpt-5	99%	96%
gpt-4o	99%	91%

Table 2: Performance comparison between Reasoning-Oriented and Simple prompting.

4.3 Auxiliary Text

Certain compounds (e.g., "love triangle", "eye candy") are overwhelmingly used in idiomatic contexts and rarely occur with a literal interpretation in either natural language or visual data. Consequently, multimodal models such as CLIP, which are trained on large-scale web corpora, tend to encode these expressions primarily according to their idiomatic meaning. To ease the burden on the encoding models and encourage more interpretable representations, we generate two auxiliary textual descriptions: one explaining the idiomatic meaning and another providing a visual description of the literal interpretation. In what follows, we will refer to these as "synonym" and "visual", respectively. Also, we will refer to the original expression as "compound". Such examples are presented in the appendix (Table 6).

Table 3 illustrates the issue discussed above using the love triangle example (see Figure 3): the three encoding models assign a higher average similarity score to the strongly idiomatic image rather than to the strongly literal one. The visual descriptor used for the literal interpretation: "three hearts connected at the corners, forming a triangle" does not exactly match the image; nevertheless, it guides the models toward a more faithful representation of the intended literal meaning.

Image Category	Compound	Synonym	Visual
Strong Literal	0.169	0.207	**0.187**
Distractor	0.072	0.137	0.015
Weak Idiomatic	0.135	0.211	0.086
Weak Literal	0.107	0.097	0.091
Strong Idiomatic	**0.188**	**0.256**	0.079

Table 3: Average cosine similarity scores between the three text types (Compound, Synonym, Visual) and the five image categories. The highest score per column is bolded.

However, the outputs of the LLM when used with the default temperature vary considerably. Lower decoding temperatures reduce variability by producing more concentrated token distributions and more deterministic outputs. While this can improve consistency, it also limits exploration of alternative reasoning paths, making the generation more sensitive to early token choices and increasing the

Setting	1st run	2nd run	3rd run
Temp 1.0	82%	90%	87%
Temp 0.3	84%	83%	84%
MBR 0.7	88%	90%	91%
MBR 1.0	90%	88%	89%

Table 4: Performance comparison across text generating strategies.

likelihood that initial errors influence subsequent reasoning.

Minimum Bayes Risk (MBR) decoding explicitly explores multiple plausible generations and selects the solution that is consistent across samples. This allows MBR to recover from individual generation errors and better approximate the model's underlying posterior, resulting in improved performance despite higher computational cost. Recent work has shown that MBR-style decoding can substantially improve robustness by selecting outputs that minimize disagreement across multiple generations rather than relying on a single sample (Heineman et al., 2024).

In our implementation, we apply MBR decoding by generating 20 independent responses at a temperature of 1.0 and 0.7, respectively, for each input. Each response is embedded using a Sentence-Transformers model (all-mpnet-base-v2) (Reimers and Gurevych, 2019). The final prediction is selected as the one whose embedding exhibits minimal average distance to the others, effectively choosing the most representative or consensus solution among the candidates. As shown in Table 4, this strategy consistently outperforms both default-temperature decoding and low-temperature decoding across all evaluated metrics, confirming that increased diversity combined with consensus-based selection yields better performance than increased determinism alone.

4.4 CLIP Ensemble

To encode both images and text, we use an ensemble of vision–language models. Table 5 reports results for seven models released by LAION (Schuhmann et al., 2022), Google (Zhai et al., 2023), and OpenAI (Radford et al., 2021), evaluated under ground-truth idiomaticity classification, thereby isolating the image–text matching performance from upstream classification errors. The three selected models are highlighted in the table.

We report performance using six metrics. Top-1 (T1) and Top-2 (T2) accuracy measure whether the correct image is ranked first or among the top two candidates, respectively. Evaluation is conducted over three complementary sub-tasks: Syn, which assesses idiomatic cases by matching images against the synonym-based generated text; Lit, which evaluates literal cases using the original compound expression; and Vis, which measures performance on literal cases using the visual descriptor of the compound.

As part of the candidate selection process, we introduce an exclusion step in which the image whose embedding is closest to the visual descriptor is removed from consideration. This is intended to prevent this image from being selected as a finalist. Without this exclusion, the system achieves an average accuracy of 83%.

The performance gain introduced by this step is observed exclusively in idiomatic cases, as literal interpretations are generally easier for the models to detect. Applying the same exclusion strategy to literal cases—by removing the most idiomatically aligned image—leads instead to a decrease in overall accuracy.

4.5 LLM Final Choice

For literal and idiomatic interpretations, the distinction between the image that best represents the meaning and the one that more vaguely resembles it is often subtle and can exceed the discriminative capacity of vision–language embedding models alone. The following example illustrates this, as the synonym text generated for "piece of cake" ("Something that is very easy to do or accomplish, often used to describe a task or activity that requires little effort or skill.") produced an embedding more similar to the weak idiomatic, instead of the expected image (see Table 7).

In such cases, similarity-based ranking may fail to reliably separate near-identical candidates. To address this limitation, we introduce a final LLM-based decision stage.

After the CLIP ensemble reduces the candidate pool to the two most semantically aligned images, the LLM is prompted to perform a direct comparison between these finalists, conditioned on the original compound, its context sentence, and the relevant auxiliary text.

Empirically, this final decision step yields a substantial improvement in accuracy, increasing performance from 74% achieved by the CLIP ensemble alone, to 90%. This result highlights the comple-

Model	Syn T1	Syn T2	Lit T1	Lit T2	Vis T1	Vis T2
google/siglip-so400m-patch14-384	61	82	77	92	88	98
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	68	83	81	94	83	96
openai/clip-vit-large-patch14	70	85	83	98	77	96
laion/CLIP-ViT-g-14-laion2B-s12B-b42K	72	87	81	98	84	100
google/siglip-large-patch16-384	4	32	26	51	22	52
laion/CLIP-ViT-B-32-laion2B-s34B-b79K	55	78	81	96	81	96
google/siglip-base-patch16-256	25	44	34	62	30	54

Table 5: Performance comparison of CLIP and SigLIP models.

mentary strengths of embedding-based retrieval and generative reasoning: while the former effectively narrows the search space, the latter excels at fine-grained semantic discrimination.

5 Results

The model accurately predicted the correct image in 53% of tests across all 15 languages, achieving second place in the overall ranking for the image and text subtrack. The exact results for each language are displayed in the appendix. On the English dataset, which was not part of the official evaluation, it achieved an accuracy of around 90% on average. The observed discrepancy between performance on the English subset and the multilingual dataset may be attributed to several factors, including differences in data availability, language-specific idiomatic usage, and the predominantly English-centric training and prompting of the underlying models, as well as a possible difference in the quality of the data.

Considering the variant that was used in the official evaluation and which is described in this paper: there are sources of error at each step of the pipeline, from the binary classification to the final LLM choice. The testing data, once augmented and published, will be a valuable asset that could further improve the system, especially by enabling development focused on multiple languages.

6 Conclusion

In this paper, we presented a multi-step, multimodal system for the MWE-2026 AdMIRe 2.0 shared task. Our approach combines auxiliary text generation, a CLIP model ensemble, and a final LLM-based decision stage to address the challenges of multimodal idiomatic understanding.

Our analysis shows that auxiliary textual representations significantly improve image–text alignment, that Minimum Bayes Risk decoding yields more robust generations than temperature-based control, and that LLMs are most effective when used as high-level decision-makers rather than direct classifiers. Despite the increased computational cost, the resulting gains justify the design choices in settings where accuracy is critical. Because the system relies heavily on LLM-based reasoning, model selection directly impacts both performance and cost. Stronger models improve accuracy but introduce significant computational and economic overhead, necessitating a trade-off between effectiveness and efficiency.

Future work includes extending the system to additional languages, exploring fine-tuned multimodal encoders, and investigating more principled risk functions for MBR decoding. We hope our findings contribute to a deeper understanding of idiomaticity in multimodal language processing.

7 Limitations

We identify several limitations that warrant further investigation:

- The experiments were conducted exclusively on the English language; therefore, the results are strongly influenced by the quality of the models in multilingual settings.
- A more extensive ablation study would be necessary to fully understand the contribution of each component; at present, each component yields a direct and gradual increase in accuracy.
- A more in-depth analysis of the models' familiarity with Brazilian Portuguese data is needed, as it may indicate potential data contamination from the first edition of the task.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). This research is also supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

David Alfter. 2025. daalft at SemEval-2025 task 1: Multi-step zero-shot multimodal idiomaticity ranking. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 127–140, Vienna, Austria. 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.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum Bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. pages 1–15.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

A LLM Prompts

Baseline Prompt
You are given 5 images, the compound '{compound}' and the context: {sentence}. If the compound is used in the context idiomatically, choose the image that best represents the IDIOMATIC MEANING, and if it is used literally, choose the image that best represents the LITERAL MEANING. Respond with ONLY "1", "2", "3", "4" or "5" - nothing else.

Reasoning-Oriented Prompt for Classification
Here is the idiom: {compound} Although this idiom is usually used idiomatically, it can be used, in rare cases, literally. Think about such an example, then look at this sentence: {sentence} Think about both possible interpretations: literal and idiomatic. Then decide which meaning matches the compound in this sentence. Your response must be exactly one word: literal OR idiomatic

Simple Prompt for Classification
In the following sentence, is the compound {compound} used idiomatically or literally: {sentence} Your response must be exactly one word: literal OR idiomatic.

Prompt for Synonym Text Generation
Define this idiom in 30-40 words, as it would appear in a dictionary: {compound} Write ONLY the definition itself in english. Do NOT include the idiom phrase in your response. Start directly with the meaning, like: "Dishonest or mischievous behavior..." not "Monkey business means..."

Prompt for Visual Description Text Generation
Generate a generic, SHORT visual description for the literal meaning of: {compound} Write ONLY the visual description itself in english. Do NOT include the compound phrase in your response. Start directly with the meaning, like: "Multiple hens having a party..." not "Women having a party..." for hen party and "Very angry aunt..." NOT "Caring woman sending letters..." for "agony aunt" REMEMBER: I WANT THE LITERAL MEANING, NOT IDIOMATIC OR METAPHORICAL

Prompt for Final 2-Candidate Choice
Choose which image better shows the {LITERAL/IDIOMATIC} meaning of: {compound} Respond only with 1 or 2.

B Examples

Type	Generated Text
	Compound: Ghost town
Synonym	A deserted town or area that was once populated or active, often characterized by abandoned buildings and a lack of human activity, typically resulting from economic decline or natural disasters.
Visual	A spectral figure wandering through deserted streets and empty buildings.

Table 6: Examples of generated auxiliary text for the compound "Ghost town".

Image Category	Avg. Similarity
Strong Idiomatic	0.155
Weak Idiomatic	**0.197**
Distractor	0.095
Weak Literal	0.101
Strong Literal	0.093

Table 7: Average cosine similarity scores between the 5 images and the synonym text, for the "piece of cake" example

Strongly idiomatic | Weakly idiomatic | Distractor | Weakly literal | Strongly literal

Figure 3: Data example for *love triangle*

Strongly idiomatic | Weakly idiomatic

Figure 4: Data example for *piece of cake*

C Test Data

Language	# Records	# Compounds
Chinese	179	57
Georgian	113	32
Greek	208	52
Igbo	115	42
Kazakh	156	51
Norwegian	202	51
Portuguese (Brazil)	228	57
Portuguese (Portugal)	220	58
Russian	140	39
Serbian	363	95
Slovak	151	42
Slovenian	240	60
Spanish (Ecuador)	48	13
Turkish	180	54
Uzbek	120	42

Table 8: Dataset composition by language. #Records denotes the total number of instances (rows) available for each language, while #Compounds indicates the number of distinct idiomatic expressions.

Language	**Acc_{all}**	**Acc_{lit}**	**Acc_{id}**	**$Corr_{all}$**	**$Corr_{lit}$**	**$Corr_{id}$**	**DCG_{all}**	**DCG_{lit}**	**DCG_{id}**
Brazilian Portuguese	0.80	0.89	0.71	0.19	0.23	0.15	0.91	0.95	0.87
Russian	0.68	0.90	0.50	0.24	0.59	-0.03	0.87	0.95	0.82
Slovenian	0.67	0.77	0.58	0.23	0.27	0.20	0.87	0.91	0.83
Turkish	0.62	0.75	0.53	0.15	0.24	0.08	0.84	0.90	0.81
European Portuguese	0.57	0.68	0.46	0.12	0.17	0.07	0.81	0.86	0.77
Greek	0.57	0.67	0.46	0.20	0.29	0.10	0.84	0.89	0.80
Kazakh	0.53	0.70	0.44	0.04	0.21	-0.05	0.80	0.89	0.75
Norwegian	0.52	0.72	0.32	0.13	0.30	-0.04	0.80	0.89	0.71
Slovak	0.48	0.54	0.42	0.19	0.21	0.17	0.78	0.80	0.77
Georgian	0.47	0.50	0.44	0.18	0.27	0.10	0.75	0.77	0.74
Chinese	0.45	0.43	0.47	0.11	0.21	0.03	0.76	0.74	0.77
Serbian	0.45	0.55	0.36	0.10	0.14	0.06	0.76	0.81	0.72
Ecuadorian Spanish	0.42	0.18	0.62	0.16	0.16	0.17	0.81	0.78	0.84
Igbo	0.39	0.55	0.33	-0.01	0.00	-0.01	0.74	0.80	0.72
Uzbek	0.33	0.47	0.26	-0.00	0.25	-0.11	0.74	0.80	0.71

Table 9: Per-language evaluation results for AdMIRe 2.0 Subtask A, sorted by decreasing overall accuracy (Acc_{all}).