MWE-2026 Shared Task: AdMIRe 2 Advancing Multimodal Idiomaticity Representation

Doğukan Arslan[1], Rodrigo Wilkens[2], Wei He[2], Dilara Torunoğlu Selamet[1], Thomas Pickard[3], Aline Villavicencio[2,3], Adriana Pagano[4], Gülşen Eryiğit[1]
[1] Istanbul Technical University, Türkiye
[2] University of Exeter, UK
[3] University of Sheffield, UK
[4] Federal University of Minas Gerais, Brazil
{arslan.dogukan, torunoglud, gulsen.cebiroglu}@itu.edu.tr, tmrpickard1@sheffield.ac.uk
{w.he, r.wilkens, a.villavicencio}@exeter.ac.uk, apagano@ufmg.br

Abstract

Idiomatic expressions present a unique challenge in NLP, as their meanings are often not directly inferable from their constituent words. Despite recent advancements in large language models, idiomaticity remains a significant obstacle to robust semantic representation. We present datasets and task results for MWE-2026 Shared Task 2: Advancing Multimodal Idiomaticity Representation 2 (AdMIRe 2), which challenges the community to assess and improve models' ability to interpret idiomatic expressions in multimodal contexts across multiple languages. Participants competed in an image ranking task in which, for each item, systems receive a context sentence containing a potentially idiomatic expression (PIE) and five candidate images. Participating systems are required to predict the sentence type (i.e., idiomatic vs. literal) for the given context and rank the images by how well they depict the intended meaning in that context. Among the participating systems the most effective methods include pipelines utilizing closed-source commercial models such as Gemini 2.5 and GPT-5, and employing chain-of-thought reasoning strategies. Methods to mitigate language models' bias towards literal interpretations and ensembles to smooth out variance were common.

1 Introduction

Idioms constitute a class of multiword expressions (MWEs) that remains challenging for state-of-the-art language models, as their meanings are often not predictable from the meanings of their constituent words (Dankers et al., 2022; Villavicencio et al., 2005). For instance, the expression *devil's advocate* is not typically used with literal denotation derived from its component words, but rather construes the meaning of someone who presents a contentious opinion in order to test an opposing argument or provoke debate. Idiomatic expressions may further give rise to ambiguity between a literal, compositional interpretation and an idiomatic, non-compositional one (He et al., 2024). These characteristics make idioms a particularly informative testbed for investigating how current language models represent and process meaning. While large language models (LLMs) perform well on general benchmarks, it is still unclear to what extent they consistently exhibit good understanding of figurative language (Mi et al., 2025; Phelps et al., 2024), even for well-resourced languages such as English.

These challenges have recently been highlighted in shared evaluations. For instance, the first edition of this task (SemEval-2025 Task 1: AdMIRe; Pickard et al., 2025) focused on two languages, English and Portuguese, to assess models' ability to interpret idiomatic expressions in multimodal contexts and the PARSEME 2.0 shared task (Scholivet et al., 2026) proposed two multilingual challenges targetting MWEs: (a) their identification and (b) their paraphrasing. In addition, there are several benchmark datasets dedicated to the processing of idiomatic expressions in text (e.g. Chakrabarty et al., 2022; Haagsma et al., 2020; Tedeschi et al., 2022; Tayyar Madabushi et al., 2021; Garcia et al., 2021; Mi et al., 2025; Arslan et al., 2025). While current models may display competitive performance on some of these datasets, it is unclear to what extent they actually require that language models possess good representations of idiom meaning (Boisson et al., 2023; He et al., 2024), or whether models are benefiting from other artifacts to address these tasks.

Moreover, even if the addition of a visual modality (alongside text) to idiom processing could lead to more informative clues being available to disambiguate and interpret potentially idiomatic expressions, it is not certain whether models benefit from the additional information. Indeed, performance on datasets like IRFL (Yosef et al., 2023) and V-FLUTE (Saakyan et al., 2025) indicates that idiomaticity processing is more difficult for vision-language models (VLMs) to perform.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 276–287

In this edition of the AdMIRe shared task, in an attempt to determine the multilingual coverage and generalizability of the results obtained by available models, we expand the number of languages, adding new evaluation instances for Chinese, Georgian, Greek, Igbo, Kazakh, Norwegian, Portuguese (Portugal), Portuguese (Brazil), Russian, Serbian, Slovak, Slovenian, Spanish (Ecuador), Turkish, and Uzbek to the existing English and Portuguese (Brazil) training data. These languages provide variation in terms of language families and scripts, and also in terms of NLP resources. Two variants of one of the languages are also included: Brazilian Portuguese and European Portuguese. We incorporate visual (§2) modalities for all 15 languages in an effort to promote the construction of higher-quality semantic representations of idioms. Our dataset incorporates items in both English (EN) and Brazilian Portuguese (PT-BR) as part of the training data, while the other languages are available only at test time, as unseen items. As in AdMIRe 1, we use nominal compounds and verbal idioms having interpretations in literal and idiomatic senses which are both plausible and imageable. This paper presents the task (§3), participating systems and results (§4) and finishes with discussions (§5) conclusions, limitations and future work (§6).

2 Dataset

Following the first edition of the AdMIRe shared task, Torunoğlu-Selamet et al. (2026) recently introduced a cross-lingual benchmark for multimodal idiomaticity understanding, as an initiative under the UniDive COST Action (Savary et al., 2024). The paper followed the same data creation strategy from Pickard et al. (2025) and introduced data for a large number of languages.[1]

For each language, the dataset contains around 60 potentially idiomatic expressions (PIEs), expressions that can be interpreted idiomatically, whether or not they are used that way in context. Annotators select a subset of the English PIEs from the dataset used in first version of the AdMIRe shared task and provide their counterparts in each target language. For the text modality, each of these is provided with at least two context sentences where the PIE is used with either its idiomatic or literal meaning. This allows verifying how well and how consistently models can distinguish these two uses for each of the target languages. The context sentences originate from diverse sources, including naturally occurring corpus data and sentences produced through expert construction or large language models. In addition, for the visual part, 5 images were machine-generated using manually-written prompts and validated by the language experts. These images cover a spectrum from fully literal to fully idiomatic interpretations of the expression, along with a semantically unrelated distractor (i.e., strongly figurative, mildly figurative, mildly literal, strongly literal, and distractor). Figure 1 provides an example set of images for the expression *green fingers*. Additionally, auto-generated captions are provided for each image in the text-only track and as part of the textual information available to the models. That means that for each language and for each PIE the models have available:

- 2 manually validated context sentences (one literal, one idiomatic)
- 5 automatically generated then manually validated images
- 5 automatically generated captions

3 Task Description

Given a context sentence containing a PIE and a set of five images, the task is to rank the images based on how well they depict the meaning of the PIE used in that sentence. A variation of the task (i.e., text-only) also allows for unimodal settings, where given a sentence and five text captions (each describing the content of one of the images) the goal is to rank the image captions on how accurately they capture the meaning of the PIE.

Publicly available training data from the first edition of the task was provided to shared-task participants for English and Brazilian Portuguese only, while no training data was released for the remaining languages. AdMIRe 2 excluded English from the set of test languages and introduced newly-created test sets for Portuguese, enabling evaluation across two language variants: a new unseen set for Brazilian Portuguese and a new set for European Portuguese.

3.1 Evaluation

We set an expected rank ordering of the 5 images following the sense in which the expression is used

[1] The full resource contains data for 34 different languages; however, due to the strict and unified formatting requirements for the shared task, only 15 languages were fully prepared at the time of the AdMIRe 2 shared task data release and were included in the evaluation.

(a) The image depicts an elderly man engaged in gardening activities. He is crouching down next to a small garden bed...

(b) The image depicts an elderly man with gray hair and a mustache, wearing a light-colored shirt and dark pants...

(c) The image depicts a close-up of a human hand with the palm facing upwards and fingers slightly curled...

(d) The image depicts a cartoon-style illustration of a young girl with brown hair tied back in a ponytail...

(e) The image depicts an elderly man playing the piano. He is seated on a black stool with orange accents...

Figure 1: Data example for *green fingers*. Images generated using Midjourney. Captions are displayed partially. (Torunoğlu-Selamet et al., 2026)

in the context sentence. The image strongly associated with the target sense is expected to be ranked first, followed by the mildly associated one. The images for the other senses and the 'distractor' image can follow interchangeably. For instance, for an idiomatic use of *green fingers* in a context sentence, strongly figurative and mildly figurative are expected to be ranked first. For the images in Figure 1, this would produce, for instance, $[a, b, d, e, c]$.

Performance for the task is assessed with two key metrics: a) Top-1 Image Accuracy, which measures only the correct identification of the **most** representative image and b) Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002), which was also adopted in the first edition of AdMIRe, as it is an established information retrieval metric that not only captures the fraction of retrieved relevant information but also takes into account their correct ordering.

Normalized Discounted Cumulative Gain (nDCG) is defined as

$$\text{nDCG} = \frac{DCG_n}{iDCG_n} = \frac{\sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)}}{\sum_{i'=1}^{n} \frac{rel^*_{i'}}{\log_2(i'+1)}},$$

where n is the number of items considered, rel_i is the relevance score (gain) of the i-th item in the system's ranking, $rel^*_{i'}$ is the relevance score of the i'-th item in the ideal ranking, and $iDCG$ is DCG of the ideal ordering of results.

Because our expected order of images is somewhat arbitrary (for a literal instance of a given expression, the idiomatic depictions are essentially no more relevant than the distractor), after experimentation we adopt relevance scores ($rel^*_{i'}$) of $[3, 1, 0, 0, 0]$ for the five image positions; this allows the metric to capture some of the relevant semantics beyond the top image accuracy without penalising systems which permute the order of the low-relevance images. The maximum (ideal) DCG score obtainable is therefore 3.631 and nDCG is bounded between 0 and 1, with higher values reflecting better ranking quality.

Competition rankings for the task are based on top image accuracy, with nDCG breaking ties.

4 Participating Systems and Results

The AdMIRe shared task competitions[2] were configured using the Codabench platform (Xu et al., 2022), attracting 27 registered participants in the images & text track and 22 registered participants in the text-only track. Users were allowed to submit multiple times during the competition, and their best result was used for evaluation. Submissions during the test phase (which determined the final leaderboard position) were limited to 10 in order to discourage 'gaming' the system while allowing participants to evaluate more than one approach if desired.

Once the competition ended, teams were asked to complete a brief questionnaire outlining their approach and enabling us to link CodaBench usernames with team names in their system description papers. Only teams who submitted a system description paper are included in the official task leaderboards. A total of 10 official team submissions were received.

[2] Available at https://www.codabench.org/competitions/10547/ and https://www.codabench.org/competitions/10548/

Team	Rank	Top-1 Accuracy			nDCG Score		
		Overall	Literal	Idiomatic	Overall	Literal	Idiomatic
ITUNLP	1	0.60 ± 0.1	0.67 ± 0.2	0.55 ± 0.1	0.85 ± 0.1	0.88 ± 0.1	0.83 ± 0.0
DCSN-NLP	2	0.53 ± 0.1	0.62 ± 0.2	0.46 ± 0.1	0.81 ± 0.0	0.85 ± 0.1	0.77 ± 0.0
ITUNLP2	3	0.52 ± 0.3	0.53 ± 0.3	0.52 ± 0.3	0.70 ± 0.4	0.70 ± 0.4	0.70 ± 0.4
tiberiucarp	4	0.50 ± 0.1	0.54 ± 0.2	0.46 ± 0.1	0.80 ± 0.0	0.84 ± 0.1	0.78 ± 0.0
PolyFrame	5	0.35 ± 0.1	0.57 ± 0.1	0.16 ± 0.0	0.73 ± 0.0	0.85 ± 0.0	0.62 ± 0.0
VisAffect	6	0.33 ± 0.0	0.13 ± 0.1	0.47 ± 0.1	0.72 ± 0.0	0.59 ± 0.0	0.81 ± 0.0
IdiomRanker-X	7	0.30 ± 0.2	0.48 ± 0.3	0.13 ± 0.1	0.58 ± 0.3	0.69 ± 0.4	0.49 ± 0.3
3K2T	8	0.13 ± 0.2	0.13 ± 0.2	0.13 ± 0.2	0.21 ± 0.4	0.21 ± 0.4	0.21 ± 0.4

Table 1: Leaderboard results for the image and text track. Macro-averaged Top-1 Accuracy and nDCG scores are reported overall and separately for literal and idiomatic sentences, together with their standard deviations. Teams are ranked by overall Top-1 Accuracy.

Team	Rank	Top-1 Accuracy			nDCG Score		
		Overall	Literal	Idiomatic	Overall	Literal	Idiomatic
ITUNLP	1	0.56 ± 0.1	0.61 ± 0.2	0.51 ± 0.1	0.83 ± 0.0	0.86 ± 0.1	0.81 ± 0.0
LST	2	0.41 ± 0.1	0.58 ± 0.1	0.28 ± 0.1	0.76 ± 0.0	0.85 ± 0.1	0.68 ± 0.0
alexandru412	3	0.32 ± 0.2	0.33 ± 0.2	0.29 ± 0.3	0.59 ± 0.3	0.60 ± 0.3	0.57 ± 0.3
PolyFrame	4	0.32 ± 0.1	0.48 ± 0.1	0.19 ± 0.1	0.71 ± 0.0	0.81 ± 0.1	0.63 ± 0.0

Table 2: Leaderboard results for the text-only track. Macro-averaged Top-1 Accuracy and nDCG scores are reported overall and separately for literal and idiomatic sentences, together with their standard deviations. Teams are ranked by overall Top-1 Accuracy.

4.1 Results

The team's ranking is shown in Tables 1 and 2, where the former includes both modality types and the latter only text. The tables report the macro-averaged mean and standard deviation of accuracy and nDCG scores for literal and idiomatic items, as well as the overall performance. Systems that do not support a given language are assigned a score of zero for that language when computing the macro-average.

In this version of the AdMIRe shared task, teams were challenged with 15 different languages. Most teams tested their solutions in 15 languages, except ITUNLP2 (Umut and Şenceylan, 2026), IdiomRanker-X (Çolak, 2026), and alexandru412 (Alexandru-Marian, 2026) (12 languages), and 3K2T (Kömürcü and Temel, 2026) (3 languages). Although we observe performance variation across the different languages, in general, performance within each language is fairly consistent, as illustrated in Figure 2, which shows the average performance and standard deviation for each language[3]. Detailed performance figures by language can be seen in the Appendix.

[3]ISO 639-1 codes have been used to represent languages.

Finally, the overall results obtained by the participating teams consistently display better accuracy for literal than for idiomatic items (Tables 1 and 2). The exception is the system by VisAffect (Bilen et al., 2026), which got better accuracy for idiomatic items in both Image and Text and Text only tracks.

4.2 Popular Approaches

Model Types Participating teams employed a variety of approaches to the AdMIRe 2 task, predominantly relying on large generative language models (LLMs) and vision-language models (VLMs). For text generation and reasoning, teams frequently utilized the GPT series (specifically GPT-4o and GPT-5; OpenAI, 2024, 2025), Qwen (versions 2.5 and 3) (Bai et al., 2023), DeepSeek models (DeepSeek-AI, 2025) and Gemini 2.5 Pro (Google, 2025) and 3 Pro Preview (DeepMind, 2025). For embeddings and vision-language alignment, there was a shift towards newer architectures; while standard CLIP (Radford et al., 2021) variants remained popular (used by DCSN-NLP (Cotigă and Nisioi, 2026) and tiberiucarp (Carp, 2026)), some teams also adopted SigLIP2 (Tschannen et al., 2025) and Jina-CLIP-v2 (Koukounas et al., 2024); i.e. PolyFrame (Hosseini-

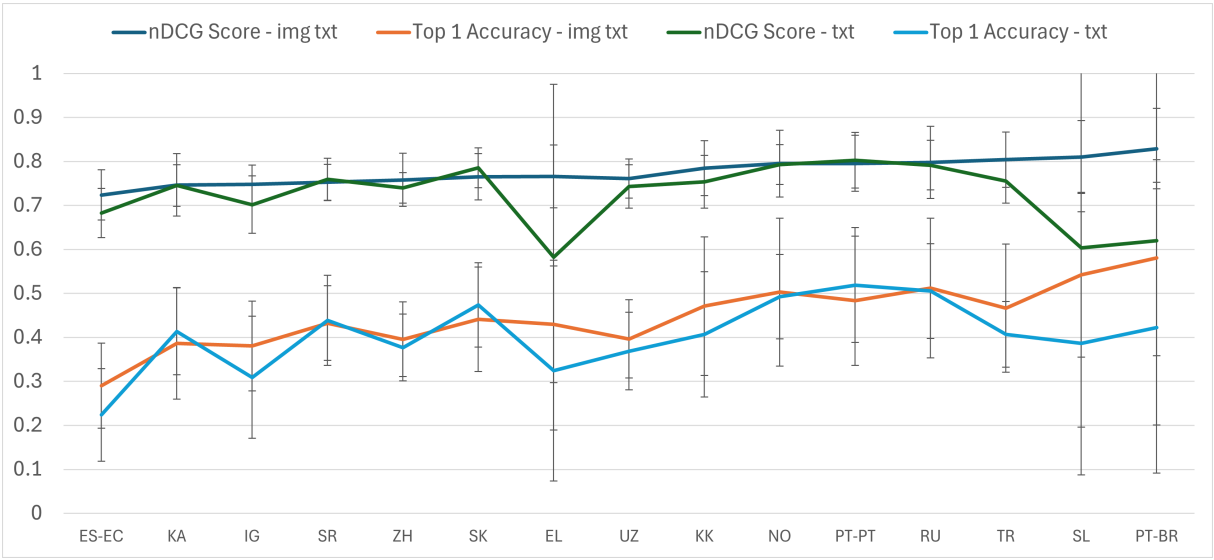

Figure 2: Average (and standard deviation) of evaluation metrics by language

Kivanani, 2026) and VisAffect (Bilen et al., 2026) respectively. Dense retrieval models like BGE-M3 (Chen et al., 2024) and multilingual encoders like XLM-RoBERTa (Conneau et al., 2020) were also widely used for text-specific ranking components.

Pipeline components Most teams implemented multi-stage pipelines rather than end-to-end solutions. A dominant pattern involved an initial binary classification of the context sentence as literal or idiomatic, triggering specialized downstream processing. To bridge the modality gap, several teams used LLMs to generate auxiliary text—such as semantic glosses or visual descriptions—to better guide the vision models, while others employed hybrid architectures that fused scores from direct vision-language matching and text-caption retrieval.

Ensembles and Fusion Robustness was often achieved through ensemble techniques. DCSN-NLP utilized a voting mechanism across an ensemble of three CLIP models (ViT-H-14, ViT-L-14, ViT-g-14). Polyframe adopted weighted Borda rank aggregation to fuse outputs from vision and text streams. LST (QIU et al., 2026) employed a large-model ensemble, aggregating outputs from GPT-4, Qwen-Plus, and DeepSeek-V3 prompting strategies.

Bias Mitigation Teams actively addressed the "literal bias" of LLMs (Phelps et al., 2024; Mi et al., 2025). PolyFrame implemented "idiom synonym replacement" for idiomatic instances, replacing the target expression with a non-figurative synonym to prevent the model from grounding visual features in the literal constituent words. Alexandru412 introduced stochastic option shuffling during inference to mitigate positional bias in multiple-choice ranking.

Data Augmentation and Cross-Lingual Strategies Zero-shot transfer was critical for handling the 15 target languages. alexandru412 utilized a test-time translation strategy (converting context sentences to English) to leverage an English-fine-tuned Qwen model. IdiomRanker-X employed dynamic prompting with "focus markers" to guide attention. PolyFrame relied on the inherent multilingual capabilities of SigLIP2 and BGE-M3 for zero-shot ranking without language-specific fine-tuning.

4.3 Most Effective Approaches

4.3.1 Text & Images Methods

The top-performing system from ITUNLP (Site et al., 2026) achieved the highest accuracy across both the Multimodal (Text + Image) and Text-Only leaderboards. Their approach leveraged a "hybrid LVLM pipeline" that combined the reasoning capabilities of GPT-5.1 with the multimodal understanding of Gemini 2.5 Pro. By delegating the initial semantic analysis to a strong reasoning model and the visual grounding to a specialized VLM, they effectively mitigated the noise often seen in end-to-end zero-shot inference.

The second-placed team, DCSN-NLP , introduced a "Hierarchical Multimodal Reasoning"

strategy to align abstract idioms with visual features. Instead of matching images directly to the idiomatic expression, their pipeline first used an LLM (such as GPT-5 or GPT-4o) to generate auxiliary text—specifically, a visual description of the literal meaning and an explanation of the idiomatic meaning. These generated descriptions were then used to query an ensemble of three CLIP models. While a voting mechanism was used to narrow the selection to the top two candidates, the final selection was performed by the LLM, which compared the finalists to make the ultimate decision.

As mentioned in the previous section, to address the "literal bias" present in vision-language models, PolyFrame (rank 5) implemented a targeted transformation step. For sentences classified as idiomatic, they replaced the target expression with a non-figurative synonym before passing the text to the vision encoder. This simple yet effective substitution prevents the VLM (in their case, SigLIP2) from grounding the literal objects and focuses the ranking on the semantic payload of the expression.

4.3.2 Text-Only Methods

The Text-Only track demonstrated that strong language models can rival multimodal systems by exploiting caption semantics. LST secured second place (Top-1 Accuracy: 0.41) using an ensemble of GPT-4, Qwen-Plus, and DeepSeek, effectively reasoning over captions without accessing pixel data. For teams participating in both tracks, the contribution of the visual modality varied. The winning team, ITUNLP, saw a drop in accuracy (0.60 to 0.56) when removing images, confirming the value of their VLM pipeline. Furthermore, PolyFrame observed a minimal performance gap (0.35 vs. 0.32) as well.

5 Discussion

We were pleased to see that most teams covered all 15 languages in the shared task. The results obtained confirm that idiomatic processing is still challenging for models, and that they are still more accurate when processing literal than idiomatic instances. Moreover, visual data seems to be helping disambiguation for both literal and idiomatic items. However, more in-depth analyses of the specific causes of error in each of the languages is still needed, and will be left for future work.

6 Conclusions

The AdMIRe tasks provide a particularly original approach to assessing models' idiom understanding by grounding figurative meaning in both textual and visual contexts. AdMIRe 2 establishes a challenging and carefully designed benchmark for multilingual and multimodal idiomaticity understanding. By combining textual contexts with visually grounded representations that span the idiomatic–literal continuum, the task enables a fine-grained evaluation of models' ability to disambiguate figurative language across languages and modalities. This shared task paves the way for further cross-lingual analyses and provides a valuable benchmark for systematically assessing the capabilities of large language models and vision–language models in idiom understanding. While the top-performing system attains an nDCG score of 85%, the task remains challenging for today's systems, leaving clear room for improvement.

Limitations

Zero-shot setting In AdMIRe 2, while training data was provided for English and Portuguese (inherited from the previous iteration; Pickard et al., 2025), the shared task introduced additional languages for which no labelled training examples were released. This experimental design forced systems to rely on zero-shot cross-lingual transfer or static pretrained knowledge rather than learning from task-specific examples. Consequently, models could not be fine-tuned to capture the specific cultural and linguistic nuances of idioms in these new set of languages, making performance heavily dependent on the coverage and biases of the underlying LLMs or VLMs rather than their ability to adapt to the specific task distribution.

Cultural background The datasets were constructed by creators who work in academic settings, and who are native speakers of the language that they study. The second edition replicates some of the limitations of the first edition, simply by means of adopting the same protocol. The language experts worked independently on their languages, and the examples they selected are also impacted by the constraints of the shared task and the parallel effort. Although the language specific examples have been carefully curated by these language experts, a subsequent independent crosslingual validation would require native speaker knowledge of the target lan-

guages and is left for future work. Our backgrounds and experiences will certainly have influenced the idiomatic expressions and context sentences we selected, the visual representations we favoured and so on.

AI tools used As for the first edition, the datasets are likely to reflect biases and limitations present in the tools used to construct them, especially the image generation and captioning models. For instance, efforts to introduce diversity in the images depicted depended on the quality of the image generation tools employed, and were not always successfully achieved.

Acknowledgments

The authors would like to acknowledge the contributions of the language leader collaborators who made this work possible. We would also like to thank the MWE 2026 Workshop Chairs, the members of the UniDive COST Action (Savary et al., 2024) and the SIGLEX-MWE Special Interest Group for their continued feedback and support. We dedicate this work to the memory of Federico Sangati and Silvio Ricardo Cordeiro.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). This work was also partly supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications [UKRI grant number EP/S023062/1]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Cristea Alexandru-Marian. 2026. alexandru412 at MWE-2026 AdMIRe 2: Dominating text-only idiom understanding via cross-lingual transfer and augmentation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Barış Bilen, Ali Azmoudeh, Hazım Kemal Ekenel, and Hatice Kose. 2026. VisAffect at MWE-2026 AdMIRe 2: IMMCAN idiom multimodal cross-attention network. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Andrei Tiberiu Carp. 2026. tiberiucarp at MWE-2026 AdMIRe 2: GLIMMER-Gloss-based image multiword meaning expression ranker. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David Cotigă and Sergiu Nisioi. 2026. DCSN-NLP at MWE-2026 AdMIRe 2: Bridging literal and figurative meaning through hierarchical multimodal reasoning. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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.

Google DeepMind. 2025. Gemini 3 system card. Technical report, Google.

DeepSeek-AI. 2025. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, pages 279–287, Marseille, France. European Language Resources Association.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2024. Investigating idiomaticity in word representations. *Computational Linguistics*, pages 1–48.

Nina Hosseini-Kivanani. 2026. Polyframe at MWE-2026 AdMIRe 2: When words are not enough: Multimodal idiom disambiguation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Kubilay Kağan Kömürcü and Tuğçe Temel. 2026. 3K2T at MWE-2026 AdMIRe 2: CARIM– category-aware reasoning for idiomatic multimodality. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

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. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI.

Dylan Phelps, Thomas M. R. 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, Torino, Italia. ELRA and ICCL.

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.

Le QIU, Yu-Yin Hsu, and Emmanuele Chersoni. 2026. Lst at AdMIRe 2: Advancing multimodal idiomaticity representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Atakan Site, Oğuz Ali Arslan, and Gülşen Eryiğit. 2026. ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM pipeline for multimodal idiom understanding and ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. 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, Carlos Manuel Hidalgo-Ternero, Chaya Liebeskind, Danka Jokić, Diego Alves, Eleni Triantafyllidi, Erik Velldal, Fred Philippy, Giedre Valunaite Oleskeviciene, Ieva Rizgeliene, Inguna Skadina, Irina Lobzhanidze, Isabell Stinessen Haugen, Jauza Akbar Krito, Jelena M. Marković, Johanna Monti, Josue Alejandro Sauca, Kaja Dobrovoljc, Kingsley O. Ugwuanyi, Laura Rituma, Lilja Øvrelid, Maha Tufail Agro, Manzura Abjalova, Maria Chatzigrigoriou, María del Mar Sánchez Ramos, Marija Pendevska, Masoumeh Seyyedrezaei, Mehrnoush Shamsfard, Momina Ahsan, Muhammad Ahsan Riaz Khan, Nathalie Carmen Hau Norman, Nilay Erdem Ayyıldız, Nina Hosseini-Kivanani, Noémi Ligeti-Nagy, Numaan Naeem, Olha Kanishcheva, Olha Yatsyshyna, Daniil Orel, Petra Giommarelli, Petya Osenova, Radovan Garabik, Regina E. Semou, Rozane Rebechi, Salsabila Zahirah Pranida, Samia Touileb, Sanni Nimb, Sarfraz Ahmad, Sarvinoz Nematkhonova, Shahar Golan, Shaoxiong Ji, Sopuruchi Christian Aboh, Srdjan Sucur, Stella Markantonatou, Sussi Olsen, Vahide Tajalli, Veronika Lipp, Voula Giouli, Yelda Yeşildal Eraydın, Zahra Saaberi, and Zhuohan Xie. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Özge Umut and Bora Şenceylan. 2026. ITUNLP2 at MWE-2026 AdMIRe 2: Modular zero-shot pipelines for multimodal idiom grounding and ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377. Special issue on Multiword Expression.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

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.

Mehmet Utku Çolak. 2026. Idiomranker-x at AdMIRe 2: Multilingual idiom-image alignment via low-rank adaptation of cross-encoders. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Appendix

A Image & Text Results

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.53	0.81
ITUNLP	2	0.50	0.80
DCSN-NLP	3	0.45	0.76
tiberiucarp	4	0.44	0.77
PolyFrame	5	0.35	0.72
VisAffect	6	0.30	0.71
IdiomRanker-X	7	0.28	0.70

Table 3: Chinese (ZH) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.56	0.82
ITUNLP	2	0.53	0.81
tiberiucarp	3	0.50	0.79
DCSN-NLP	4	0.47	0.75
VisAffect	5	0.34	0.72
PolyFrame	6	0.27	0.69
IdiomRanker-X	7	0.27	0.70

Table 4: Georgian (KA) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.64	0.87
DCSN-NLP	2	0.57	0.84
tiberiucarp	3	0.54	0.83
PolyFrame	4	0.36	0.72
IdiomRanker-X	5	0.34	0.72
VisAffect	6	0.27	0.69

Table 5: Greek (EL) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.56	0.84
ITUNLP	2	0.43	0.78
3K2T	3	0.41	0.76
DCSN-NLP	4	0.39	0.74
tiberiucarp	5	0.37	0.74
PolyFrame	6	0.33	0.73
VisAffect	7	0.30	0.73
IdiomRanker-X	8	0.22	0.69

Table 6: Igbo (IG) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.70	0.89
ITUNLP	2	0.61	0.84
3K2T	3	0.56	0.84
DCSN-NLP	4	0.53	0.80
tiberiucarp	5	0.51	0.81
VisAffect	6	0.40	0.74
PolyFrame	7	0.33	0.74
IdiomRanker-X	8	0.28	0.71

Table 7: Kazakh (KK) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.82	0.94
ITUNLP	2	0.67	0.88
DCSN-NLP	3	0.52	0.80
tiberiucarp	4	0.51	0.80
PolyFrame	5	0.42	0.75
IdiomRanker-X	6	0.38	0.74
VisAffect	7	0.29	0.70

Table 8: Norwegian (NO) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.88	0.96
ITUNLP	2	0.86	0.94
DCSN-NLP	3	0.80	0.91
tiberiucarp	4	0.67	0.88
PolyFrame	5	0.46	0.77
IdiomRanker-X	6	0.34	0.73
VisAffect	7	0.30	0.71

Table 9: Braz. Portuguese (PT-BR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.72	0.91
ITUNLP	2	0.64	0.86
DCSN-NLP	3	0.57	0.81
tiberiucarp	4	0.55	0.83
PolyFrame	5	0.43	0.76
VisAffect	6	0.30	0.71
IdiomRanker-X	7	0.30	0.72

Table 10: Euro. Portuguese (PT-PT) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.71	0.91
ITUNLP	2	0.69	0.89
DCSN-NLP	3	0.68	0.87
tiberiucarp	4	0.63	0.85
PolyFrame	5	0.40	0.73
VisAffect	6	0.36	0.73
IdiomRanker-X	7	0.35	0.74

Table 11: RU – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.62	0.84
tiberiucarp	2	0.48	0.78
DCSN-NLP	3	0.45	0.76
PolyFrame	4	0.39	0.74
VisAffect	5	0.36	0.72
IdiomRanker-X	6	0.31	0.71

Table 12: Serbian (SR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.60	0.85
tiberiucarp	2	0.51	0.82
DCSN-NLP	3	0.48	0.78
PolyFrame	4	0.38	0.73
VisAffect	5	0.34	0.72
IdiomRanker-X	6	0.31	0.72

Table 13: Slovak (SK) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.82	0.94
ITUNLP	2	0.78	0.91
DCSN-NLP	3	0.67	0.87
tiberiucarp	4	0.59	0.84
PolyFrame	5	0.41	0.75
IdiomRanker-X	6	0.36	0.75
VisAffect	7	0.28	0.70

Table 14: Slovenian (SL) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
DCSN-NLP	1	0.42	0.81
ITUNLP2	2	0.40	0.79
tiberiucarp	3	0.33	0.74
VisAffect	4	0.33	0.69
ITUNLP	5	0.27	0.73
IdiomRanker-X	6	0.23	0.69
PolyFrame	7	0.17	0.66

Table 15: Ecuadorian Spanish (ES-EC) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.68	0.90
ITUNLP2	2	0.65	0.88
DCSN-NLP	3	0.62	0.84
3K2T	4	0.54	0.83
tiberiucarp	5	0.48	0.80
PolyFrame	6	0.34	0.71
VisAffect	7	0.31	0.72
IdiomRanker-X	8	0.29	0.70

Table 16: Turkish (TR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.52	0.83
ITUNLP2	2	0.52	0.83
tiberiucarp	3	0.42	0.77
VisAffect	4	0.42	0.75
3K2T	5	0.41	0.78
DCSN-NLP	6	0.33	0.74
PolyFrame	7	0.32	0.72
IdiomRanker-X	8	0.31	0.71

Table 17: Uzbek (UZ) – Image and Text

B Text-only Results

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.46	0.77
alexandru412	2	0.41	0.76
LST	3	0.36	0.74
PolyFrame	4	0.28	0.69

Table 18: Chinese (ZH) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.51	0.79
alexandru412	2	0.46	0.77
LST	3	0.40	0.74
PolyFrame	4	0.28	0.68

Table 19: Georgian (KA) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.59	0.86
LST	2	0.43	0.76
PolyFrame	3	0.31	0.71

Table 20: Greek (EL) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.48	0.78
LST	2	0.33	0.71
PolyFrame	3	0.29	0.69
alexandru412	4	0.14	0.63

Table 21: Igbo (IG) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.60	0.84
LST	2	0.42	0.76
alexandru412	3	0.33	0.71
PolyFrame	4	0.28	0.72

Table 22: Kazakh (KK) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.61	0.85
alexandru412	2	0.52	0.80
LST	3	0.43	0.77
PolyFrame	4	0.41	0.75

Table 23: Norwegian (NO) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.79	0.92
LST	2	0.53	0.81
PolyFrame	3	0.37	0.75

Table 24: Brazilian Portuguese (PT-BR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
alexandru412	1	0.64	0.85
ITUNLP	2	0.62	0.86
LST	3	0.45	0.77
PolyFrame	4	0.37	0.73

Table 25: European Portuguese (PT-PT) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.65	0.87
LST	2	0.51	0.79
alexandru412	3	0.47	0.77
PolyFrame	4	0.39	0.74

Table 26: RU – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.55	0.82
alexandru412	2	0.48	0.77
LST	3	0.40	0.74
PolyFrame	4	0.31	0.71

Table 27: Serbian (SR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
alexandru412	1	0.56	0.80
ITUNLP	2	0.54	0.84
LST	3	0.44	0.78
PolyFrame	4	0.35	0.73

Table 28: Slovak (SK) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.72	0.89
LST	2	0.45	0.78
PolyFrame	3	0.37	0.74

Table 29: Slovenian (SL) – Text Only

Team	Rank	Top-1 Acc.	nDCG
LST	1	0.35	0.73
ITUNLP	2	0.25	0.72
PolyFrame	3	0.19	0.67
alexandru412	4	0.10	0.61

Table 30: Ecuadorian Spanish (ES-EC) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.51	0.82
LST	2	0.40	0.74
alexandru412	3	0.40	0.75
PolyFrame	4	0.32	0.70

Table 31: Turkish (TR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.50	0.82
alexandru412	2	0.34	0.71
LST	3	0.32	0.73
PolyFrame	4	0.32	0.71

Table 32: Uzbek (UZ) – Text Only