Preprint:
Pickard, T. orcid.org/0000-0002-2523-2364, Villavicencio, A. orcid.org/0000-0002-3731-9168, Mi, M. et al. (3 more authors) (Submitted: 2025) SemEval-2025 Task 1: AdMIRe -- Advancing Multimodal Idiomaticity Representation. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arxiv.2503.15358

SemEval-2025 Task 1: AdMIRe - Advancing Multimodal Idiomaticity Representation

Thomas Pickard[1], Aline Villavicencio[1,2] Maggie Mi[1], Wei He[2], Dylan Phelps[1] and **Marco Idiart[3]**

[1] University of Sheffield, UK
[2] University of Exeter, UK
[3] Federal University of Rio Grande do Sul, Brazil

{tmrpickard1, zmi1, drsphelps1}@sheffield.ac.uk

{a.villavicencio, w.he}@exeter.ac.uk, marco.idiart@gmail.com

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 (LLMs), idiomaticity remains a significant obstacle to robust semantic representation. We present datasets and tasks for SemEval-2025 Task 1: AdMIRe (Advancing Multimodal Idiomaticity Representation), which challenges the community to assess and improve models' ability to interpret idiomatic expressions in multimodal contexts and in multiple languages. Participants competed in two subtasks: ranking images based on their alignment with idiomatic or literal meanings, and predicting the next image in a sequence. The most effective methods achieved human-level performance by leveraging pretrained LLMs and vision-language models in mixture-of-experts settings, with multiple queries used to smooth over the weaknesses in these models' representations of idiomaticity.

1 Introduction

Idioms are a class of multi-word expression (MWE) which pose a challenge for current state-of-the-art models because their meanings are often not predictable from the individual words that compose them (Dankers et al., 2022; Villavicencio et al., 2005). For instance, "eager beaver" is unlikely to refer to a passionate muskrat; rather, it typically describes a person who is keen and enthusiastic. These expressions may also generate ambiguity between the literal, surface meaning arising from their component words and their idiomatic meaning (He et al., 2024b). These, among other characteristics, make them a valuable testing ground for examining how NLP models capture meaning.

Advances in language modeling of such phenomena have been made in recent years (Zeng and Bhat, 2022; Zeng et al., 2023; He et al., 2024a), but large language models (LLMs) which perform well on general benchmarks (even for well-resourced languages such as English) fail to consistently exhibit good understanding of figurative language (Mi et al., 2024; Phelps et al., 2024). This has an impact on their application in natural language processing activities such as sentiment analysis (Williams et al., 2015; Spasić et al., 2020), understanding and inference tasks and machine translation (Yazdani et al., 2015; Syahrir and Hartina, 2021). For example, due to poor automatic translation of an idiom, the Israeli PM appeared to call the winner of Eurovision 2018 a 'real cow' instead of a 'real darling'![1].

Several benchmark datasets exist for 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., 2024) but concerns have been raised that these tasks do not necessarily require that language models possess good representations of idiom meaning (Boisson et al., 2023; He et al., 2024b). More recent datasets (Yosef et al., 2023; Saakyan et al., 2025) introduce a visual modality to idiom processing tasks alongside text, and their findings indicate that this task is indeed more difficult for vision-language models (VLMs) to perform.

The task presented here builds on previous SemEval tasks exploring the evaluation of compositional models (Marelli et al., 2014), paraphrases and interpretation of noun compounds (Hendrickx et al., 2013; Butnariu et al., 2009) and idiomaticity detection (Tayyar Madabushi et al., 2022). We incorporate visual (§3.1) and visual-temporal (§3.2) modalities across two subtasks 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), and we focus on nominal com-

[1] metro.co.uk

pounds which have interpretations in literal and idiomatic senses which are both plausible and imageable.

The AdMIRe datasets are available from doi.org/10.15131/shef.data.28436600.v1 under a CC-BY-4.0 license.

2 Dataset Construction

2.1 Target compound selection

The potentially idiomatic noun compounds used in this study were sourced from existing datasets including NCTTI (Tayyar Madabushi et al., 2021), FLUTE (Chakrabarty et al., 2022) and MAGPIE (Haagsma et al., 2020), or identified by the researchers during the project. For the purposes of this task, we filtered out compositional expressions (e.g. *olive oil*), focusing exclusively on those that exhibit duality — i.e., expressions that can be interpreted either literally or figuratively depending on the context (e.g. *silver bullet*). The candidate lists covered expressions in both English (EN) and Brazilian Portuguese (PT-BR).

The candidate expressions were then used to create two data subsets.

2.2 Static Images (Subtask A)

Native speakers of the target language were asked to write a short sentence describing a visual scene depicting each target expression in the following contexts:

- strongly figurative
- mildly figurative
- mildly literal
- strongly literal

In addition, a 'distractor' prompt was also requested - this was something unrelated to either the figurative or literal meaning of the expression. Where the annotators were unable to construct any of these scenes, the item was excluded as a candidate. For instance, it is difficult to capture the semantics of an idiomatic *kangaroo court* in an image. Candidate items therefore needed to have plausible and imageable literal and idiomatic interpretations.

For each item selected, the annotators also provided two context sentences; one in which the expression is used literally, and one idiomatic. These sentences were obtained from existing corpora (Tayyar Madabushi et al., 2021; Jakubíček et al., 2013) or were written specifically for AdMIRe. For Portuguese, we observed that many adjective-noun compounds would normally be distinguished in writing between literal and idiomatic cases since they would be hyphenated in the idiomatic instance. For instance, *ovelha-negra* means "black sheep" (with the same idiomatic sense as English), but *ovelha negra* would be a literal black sheep. To avoid creating a shortcut for the models, we removed these hyphens from the idiomatic context sentences, and a native speaker reviewed them to confirm that they still appeared to be natural.

In total, we obtained 100 English and 55 Portuguese potentially idiomatic expressions, with literal and idiomatic context sentences and visual scene descriptions for each item.

Table 1: Data Sources for Static Images.

Source	EN	PT-BR
NCTTI	54	54
MAGPIE	11	-
Crowdsourced	31	1
FLUTE	4	-
Total	100	55

2.3 Image Sequences (Subtask B)

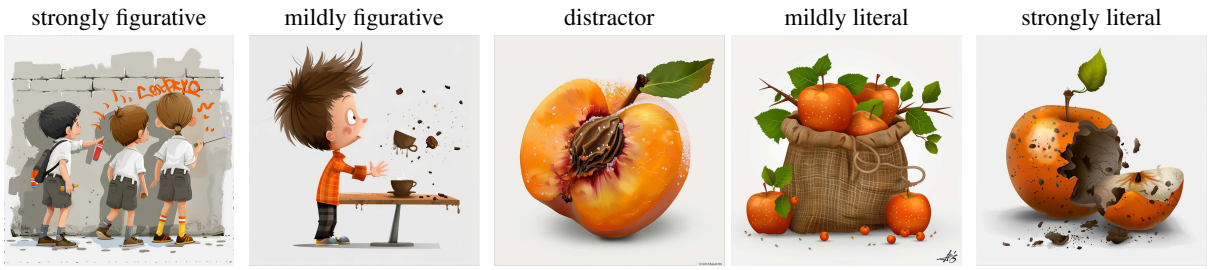
The semantics of many idiomatic expressions are difficult to capture in a single, static image, as they incorporate aspects of ongoing actions or changes over time. For instance, a *brain drain* is not an instantaneous event but something which takes place over a period of time. In order to represent some of these items in the AdMIRe dataset, we also incorporated a visual-temporal modality.

For the image sequences dataset, native speakers of English were asked to write short descriptions of three visual scenes which form a sequence representing either the literal or idiomatic sense of a given expression, akin to a three-panel comic strip. For each of the sequences, they also wrote two alternatives to the final image in the sequence. These alternatives included elements relating to the previous panels but were incompatible with the target expression, and were intended to increase the difficulty of identifying the correct completion for the sequence based solely on image similarity without using the semantic information.

A total of 30 items were collected in English. Note that 10 items appear in both of the English data subsets.

2.4 Image Generation

For each visual scene created by the annotators, we used a commercial text-to-image diffusion model



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

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

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

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

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

Figure 1: Subtask A data example for *bad apple*. Images generated using Midjourney. Captions are displayed partially.

(Midjourney[2]) to generate a corresponding image. This choice of tool provided the fine-grained human control needed to produce high-quality, domain-specific images tailored to the task, guided by human expert supervision to ensure alignment with literal and idiomatic meanings. A consistent 'style reference' image was used, guiding the model to produce images with a consistent, cartoon-like appearance. For image sequences where the same character(s) were needed, we also employed a set of character reference images in the same style.

2.5 Caption Generation

In order to support participation in the shared task by teams who wish to work only with text models (which lowers complexity and computational costs), we generated descriptive captions for each image. These were obtained by using the *LLaVA-HF/v1.6-mistral-7b-hf*[3] (LLaVA) model, a large vision-language model specifically designed for tasks requiring multimodal reasoning (Liu et al., 2024). LLaVA integrates a vision encoder to extract semantic features from images and a large language model to process these features and generate text. By employing the prompt "*What is shown in this image?*", the model generates captions that describe the content of the input images. The workflow ensures that the visual and textual components of the model work in harmony to produce accurate and contextually relevant descriptions. To ensure the quality of the generated captions, all outputs were reviewed and verified by human evaluators. For items in Portuguese, we provide captions in both English and Portuguese.

[2] https://www.midjourney.com/

[3] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

3 Task Description

3.1 Task A: Multiple Image Choice

Subtask A uses the static image portion of the AdMIRe dataset (§2.2). Given a context sentence containing a potentially idiomatic nominal compound (NC) and a set of five images, the task is to **rank** the images based on how accurately they depict the meaning of the NC used in that sentence.

A variation of task also allows for monomodal settings, where given a sentence and five text captions (each describing the content of one of the images, as described in Section 2.5) the goal is to rank the image captions on how they capture the meaning of the NC.

Figure 1 provides an example of the Subtask A data for the expression *bad apple*.

The dataset is split into training, development and test sets, with each compound present in only one subset and one, randomly-selected, sense (literal or idiomatic). In addition, we provide an extended evaluation set, which uses all 100 compounds (and therefore overlaps with the other subsets), but selects the other sense of the NC. For example, *elbow grease* appears with its idiomatic sense in the training dataset, and literally in the extended evaluation set.

The English dataset for Subtask A includes 70 training items (350 image-caption pairs), 15 development items, and 15 test items, while the Portuguese dataset comprises 32 training items (160 image-caption pairs), 10 development items, and 13 test items.

First two of the sequence:

(a) The image shows a classroom scene with five animated characters, likely children, sitting at desks...

(b) The image shows an animated character, a young boy with red hair, sitting at a desk with a laptop...

Candidates:

(c) The image shows a group of animated characters, likely children, standing in front of a television screen that displays various cartoon animals...

(d) The image shows an animated character, a young man with brown hair, standing next to a trash can...

(e) The image shows a cartoon illustration of a young girl with red hair, wearing a white sweater with a red bow and a plaid skirt...

(f) The image shows a group of animated characters that appear to be in a state of distress or chaos...

Figure 2: Subtask B data example for *bad apple*. Images (a) and (b) form the initial part of the sequence, while images (c) through (f) serve as completion candidates. In this instance, the intended sense of *bad apple* is idiomatic.

3.2 Subtask B: Image Sequences (Next Image Prediction)

Subtask B uses the image sequence portion of the AdMIRe dataset (§2.3).

For each target expression, the first two images in either the literal or idiomatic sequence are provided, along with a set of four candidate images for completion of the sequence. The candidates consist of the intended completion, the two associated alternatives and the completion for the other sense (literal/idiomatic) of the NC. The task is to correctly identify the intended completion image, while also determining whether the depicted sense of the nominal compound (NC) is idiomatic or literal. Examples are shown in Figure 2.

As with Subtask A, we also offer two settings for Subtask B, with descriptive text replacing the images in the 'caption' setting. In the Subtask B dataset, the English set includes 20 examples for training, 5 for development, and 5 for testing. All 30 items are also included in an extended evaluation set, with initial sequences and completion candidates appropriate to the NC sense not used in the primary data.

3.3 Evaluation

3.3.1 Subtask A

For subtask A, we set an expected rank ordering of the 5 images which depends on the sense in which the expression is used in the context sentence. The image strongly associated with the target sense should be ranked first, followed by the mildly associated one. The images for the other sense follow, while the 'distractor' image is always expected to be least relevant. For instance, if *bad apple* is used idiomatically in the context sentence then the expected ranking for the images would be: strongly figurative, miildly figurative, mildly literal, strongly literal, distractor. For the images in Figure 1, this would produce $[a, b, d, e, c]$.

Performance for Subtask A is assessed with two key metrics: a) Top Image Accuracy, which measures only the correct identification of the **most** representative image and b) Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002), an established information retrieval metric that not only captures the fraction of retrieved relevant information but also takes into account their correct ordering.

Discounted Cumulative Gain (DCG) is defined as

$$\text{DCG}_n = \sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)},$$

where rel_i is the relevance score of the i-th item, and n is the number of items considered.

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 a weighting 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 DCG score obtainable is therefore 3.631.

Competition rankings for Subtask A are based on top image accuracy, with DCG breaking ties.

3.3.2 Subtask B

This subtask assesses the model's ability to complete a sequence of images that narratively represent an idiomatic expression, along with distinguishing between idiomatic and literal meanings.

Evaluation metrics for subtask B are a) Image completion accuracy, which measures the correctness of the selected image to complete the narrative and b) Sentence type accuracy, measuring the effectiveness in identifying idiomatic versus literal expressions.

Subtask	Language	Track	
		Text & Images	Text-Only
A	English	16	5
	Portuguese	11	1
B	English	3	1

Table 2: Number of submissions made to each subtask

4 Participating Systems and Results

The AdMIRe shared task competitions were configured using the Codabench platform (Xu et al., 2022), with the main benchmark (Subtask A, images & text) attracting 198 registered participants[4]. Users were allowed to make multiple submissions during the competition, and were able to select their best result for evaluation. Submissions during the test phase (which determined the final leaderboard position) were limited to 5 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 outlinging 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 20 official team submissions were received. The number of submitted entries for each combination of subtask, track and language is summarised in Table 2.

4.1 Results

Results for each combination of subtask, language and track are presented in tables 3 - 8.

4.2 Popular Approaches

Model Types Participating teams employed a variety of approaches to the AdMIRe subtasks. All teams opted to use prompting methods with large, generative language and vision-language models and/or to fine-tune smaller pretrained models on the task. Popular model families included GPT-4 (OpenAI, 2024); QWEN (Bai et al., 2023); SBERT (Reimers and Gurevych, 2019); RoBERTa (Liu et al., 2019); CLIP (Radford et al., 2021) and its variants such as CLIP-ViLT (Wang et al., 2024), AltCLIP (Chen et al., 2022) and ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022). There was also some exploration of the recent DeepSeek model family (DeepSeek-AI, 2025).

Pipeline components Most (13/20) teams broke the task down into multiple steps - typically, a binary classification of the context sentence as idiomatic or literal, followed by the image ranking. Many teams introduced variations in processing for literal and idiomatic instances, with alterations to LLM prompts, input data augmentation (particularly for text passed to CLIP) and in some cases models finetuned for each sentence type.

Mixtures of Experts Four teams used a mixture-of-experts approach to one or more elements of the task, employing a variety of model types and sizes or prompt variation to smooth out inconsistencies in the model outputs.

Challenges of LLMs Several teams reported that they took steps to try to ensure that the outputs of generative LLMs were useful, including by prompting to constrain generation and by post-processing to parse the text. Three teams working with LLMs observed a bias in these models towards treating potentially idiomatic expressions as idiomatic regardless of the context in which they occur. Other work has described similar challenges when using LLMs for idiomaticity detection (Phelps et al., 2024; Mi et al., 2024; He et al., 2024b). See also Khoshtab et al. (2025) for discussion of prompting methods for LLM figurative language processing.

Data Augmentation Seven teams describe some kind of data augmentation step, including paraphrasing and backtranslation to generate variations for finetuning, the addition of information about the target compound (by distillation from LLMs) and image variations. Two teams (FJWU_Squad and dutir914) generated additional training instances automatically using generative model pipelines. We hope that these datasets will be made public by the authors in support of future research efforts.

4.3 Most Effective Approaches

4.3.1 Subtask A

The system submitted by AlexUNLP introduced a transformation step between the sentence classifica-

[4]We encountered a number of automated registrations which may have inflated this number somewhat. It is unclear what anyone hoped to achieve by generating spam registrations for the benchmark.

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Top 1 Acc	DCG Score	Rank	Top 1 Acc	DCG Score
PALI-NLP	1	0.93	3.52	1	0.83	3.43
dutir914	2	0.93	3.46	3	0.79	3.28
AlexUNLP-NB	3	0.93	3.45	5	0.72	3.22
AIMA	4	0.87	3.44	10	0.48	2.90
daalft	5	0.87	3.43	2	0.81	3.35
PoliTo	6	0.87	3.381	4	0.75	3.20
UCSC NLP T6	7	0.87	3.36	-	-	-
Zhoumou	8	0.73	3.20	6	0.69	3.21
HiTZ-Ixa	9	0.73	3.13	7	0.58	3.00
TueCL	10	0.67	3.16	-	-	-
Howard University-AI4PC	11	0.67	3.13	-	-	-
UoR-NCL	12	0.67	3.10	8	0.57	2.96
FJWU_Squad	13	0.60	2.90	11	0.47	2.85
JNLP	14	0.53	3.14	9	0.55	3.13
Modgenix	15	0.53	2.82	-	-	-
YNU-HPCC	16	0.47	2.85	-	-	-
UMUTeam	17	0.40	2.68	12	0.24	2.52

Table 3: Leaderboard results - **Subtask A, English**, Text & Images

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Top 1 Acc	DCG Score	Rank	Top 1 Acc	DCG Score
CTYUN-AI	1	0.87	3.51	1	0.64	3.10
daalft	2	0.67	3.07	4	0.33	2.61
JNLP	3	0.67	3.04	3	0.51	2.86
Transformer25	4	0.47	2.82	2	0.54	3.04
ChuenSumi	5	0.40	2.89	5	0.29	2.68

Table 4: Leaderboard results - **Subtask A, English**, Text Only

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Top 1 Acc	DCG Score	Rank	Top 1 Acc	DCG Score
HiTZ-Ixa	1	1.00	3.51	7	0.45	2.82
dutir914	2	0.92	3.43	2	0.69	3.06
Zhoumou	3	0.85	3.33	3	0.67	3.10
daalft	4	0.77	3.31	4	0.56	2.95
PALI-NLP	5	0.69	3.21	1	0.76	3.23
AlexUNLP-NB	6	0.62	3.09	6	0.51	2.91
UoR-NCL	7	0.54	3.05	5	0.56	2.90
YNU-HPCC	8	0.38	2.92	-	-	-
UMUTeam	9	0.38	2.57	8	0.18	2.33
Howard University-AI4PC	10	0.23	2.64	-	-	-
Modgenix	11	0.23	2.57	-	-	-

Table 5: Leaderboard results - **Subtask A, Portuguese**, Text & Images

tion and image ranking stages, in which compounds detected as being idiomatic are replaced with compositional synonyms (*dirty money* becomes *illegal money*). This step is intended to bypass the VLM's tendency to favour literal interpretations of compounds. The USCS NLP T6 team also comment on this phenomenon, and share our hypothesis that this likely stems from the use of image-caption datasets for model training; it seems likely that humans employ idiomatic language less frequently when captioning images[5]. AlexUNLP also evaluated a large number of different language and vision model combinations, with an ensemble of several models yielding the best results. For Portuguese, multimodal models outperformed translating into English.

The second-placed system from dutir914 also used an ensemble of outputs from several QWEN2.5 models, with chain-of-thought prompt-

[5]Very frequent idiomatic expressions and things which are often photographed may be exceptions - we recommend asking an image generation model to picture a *hen party*.

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Top 1 Acc	DCG Score	Rank	Top 1 Acc	DCG Score
CTYUN-AI	1	0.92	3.43	1	0.56	2.97

Table 6: Leaderboard results - **Subtask A, Portuguese**, Text Only

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Image Accuracy	Sentence Type	Rank	Image Accuracy	Sentence Type
daalft	1	0.60	1.00	2	0.23	0.77
PALI-NLP	2	0.60	0.80	1	0.93	1.00
Modgenix	3	0.60	0.60	-	-	-

Table 7: Leaderboard results - **Subtask B, English**, Text & Images

ing influencing the model generations. The team also generated additional training data, which was used to fine-tune CLIP. This data was produced by using DeepSeek to generate explanations and representative sentences for idiomatic expressions, which were passed to Flux.1-dev[6] to generate images.

PALI-NLP obtained the highest overall performance for Subtask A. In particular, their methodology was the most successful on the extended evaluation sets for both English and Portuguese. Among other detailed adjustments to the prompts used to interact with LLMs, they introduce an interesting adjustment to counter the models' bias towards idiomatic interpretations. By first asking the LLM to produce examples of the expression used literally before providing the context sentence, they improve classification accuracy from 91.4 to 98.6%. Examplars are also incorporated into the instruction prompts, with challenging ones purposefully selected. Multiple prompt variations are used, with the final output derived from aggregating the corresponding outputs (Wang et al., 2023).

4.3.2 Subtask B

Subtask B (image sequence completion; §3.2) attracted few submissions, and no teams participated only in this subtask. Due to the limited size of the development and test datasets, measured system performance was somewhat volatile. The highest-ranked system (daalft) reported a substantial drop in accuracy on the extended test set. PALI-NLP's approach again exhibited greater stability here. Their methodology involved prompting an LLM to generate a continuation of the narrative represented in the two initial images, then selecting from the candidate images based on their appropriateness for this generated continuation. As with subtask A, they achieved improvements of around 10% by aggregating across the output of several prompt variations (Wang et al., 2023).

4.3.3 Text-Only Methods

While most participating teams focused on the text & image configuration, a few also submitted versions of their systems which used only the provided image captions (§2) as input. These teams generally reported that the image data was beneficial to their overall performance. Several teams reported that (automatically) translating the captions into Portuguese improved the classification and ranking performance of their multimodal language models. Some teams, especially those working with smaller models (e.g. SBERT; Reimers and Gurevych, 2019), found that they needed to shorten the supplied captions through truncation or summarisation.

Two teams participated only in the text-only tracks. The Transformer25 team employed smaller, fine-tuned SBERT models for the image ranking task, but augmented the captions with information about the target item generated by a large pretrained GPT-4 model. The most successful text-only approach (CTYUN-AI) also employed data augmentation using synonym replacement and backtranslation, and this team also finetuned QWEN (Bai et al., 2023) models to the AdMIRe task. Interestingly, they report that their experiments with knowledge distillation from GPT-4 did not provide performance uplift, and that the largest QWEN2.5-72B model was no more effective than its 32B-parameter version.

[6] https://flux1ai.com/dev

Team	Test Set Metrics			Extended Evaluation Metrics		
	Rank	Image Accuracy	Sentence Type	Rank	Image Accuracy	Sentence Type
daalft	1	1.00	1.00	2	0.60	0.77
PALI-NLP	2	0.80	0.60	1	0.60	0.90

Table 8: Leaderboard results - **Subtask B, English**, Text Only

5 Human Evaluation

In order to provide a comparison point with the model-driven systems, we presented subtask A (using the English extended evaluation dataset) to human annotators. Twelve (self-described) fluent speakers of English were recruited from among the staff and postgraduate research students of a UK University, and were compensated with a voucher to the value of GBP15 for their time. The annotations were collected using a survey deployed on the Qualtrics online platform[7]. For each item, annotators were asked whether they are familiar with its idiomatic meaning. If they answered 'Yes', they were then asked to use drag-and-drop to rank the candidate images according to how well they represent the meaning of the expression as it is used in the context sentence.

Each annotator saw between 50 and 75 of the 100 items in the extended evaluation dataset, selected at random, and each item was annotated 7-9 times. Table 9 shows the mean top 1 accuracy and DCG score across the human annotators on the dataset. We also include the metrics for the best-performing individual annotator and the results obtained by treating all of the annotators for each item as a pool of experts, ranking the images according to the mean rank assigned by the annotators.

Table 9: Human annotator performance on the English extended evaluation set

	Top 1 Acc	DCG Score
Annotator average	0.71	3.22
Best individual	0.86	3.41
Pool of experts	0.83	3.39

These results would place the average annotator in 5th position against the systems submitted to the shared task on a leaderboard based on the extended evaluation set scores. The best individual annotator would outperform the model-driven systems, and the pool of experts approach would tie with the most performant system.

[7] https://www.qualtrics.com/

6 Discussion

Subtask B We received few entries for subtask B. This may have been influenced by the smaller quantity of training data available, or perhaps the more complex semantics of the compounds meant participants opted to focus on subtask A in the first instance. The best-performing system for subtask B obtained impressive results, which suggests that the task may not be more difficult in practice.

Extended evaluation The extended evaluation datasets appear to have yielded more robust results than the smaller test sets; models which were tuned on the training data did not always hold up very well on the larger test set, suggesting that some overfitting may have occurred. In this instance, there may be an advantage to have overlap between the compounds included in training and test sets, especially when they are used in different senses.

Languages We were pleased to see the Portuguese datasets receiving attention from most of the participating teams. There was a smaller difference in performance between the English and Portuguese portions of the dataset than we were anticipating (Phelps et al., 2024), suggesting that multilingual LLMs' understanding of figurative elements of languages other than English may be improving.

LLMs Participating teams, including the best-performing system overall, were able to get good results from large-scale LMs. However, this required substantial editing and refinement of both input prompts and generated output, and often the use of several parallel queries to smooth out model variance. These, presumably, came with corresponding costs in terms of human effort, money and computation (with its associated environmental impact). Most of the LLMs used by participants were also closed-source, making them difficult to examine in depth.

Human performance The results of our human evaluation (§5) suggest that the task is by no means trivial for humans who are familiar with the expres-

sion in question, but that the expected order of the images we generated is reasonably well-aligned with human understanding.

Mixtures of 'experts' Mixture-of-experts approaches proved useful to increase overall accuracy across various model scales. This suggests that individual language models may exhibit good 'understanding' of particular idiomatic expressions, or be able to handle them in specific senses and contexts, but that no one model has a complete grasp on the phenomenon of idiomaticity. Generative LLMs also appeared to be inconsistent in their outputs, varying in response to how inputs are formulated and/or with their stochastic outputs. To some extent, these observations also hold for our human annotators, with no individual's responses perfectly matching the expected answers.

7 Conclusion

The AdMIRe shared task has encouraged participants to push the boundaries of idiomatic language representation by leveraging multimodal approaches incorporating both static and temporal visual cues. By expanding on previous idiomaticity-related tasks, AdMIRe introduces a dataset of nominal compounds that are both plausible and imageable in literal and figurative contexts, covering both English and Brazilian Portuguese. This challenge has provided valuable insights into the capabilities and limitations of contemporary vision-language models in processing idiomatic expressions.

While top-performing models achieved human-level performance, this success required significant computational resources and fine-tuning, which points to the limitations of these models. There remains considerable room for improvement in both the quality of idiomatic understanding in language models and their efficiency in processing these expressions. Future iterations of the dataset and task could drive further advancements by incorporating more languages, diverse figurative expressions, and refined methodologies that minimize dataset artifacts (Boisson et al., 2023).

Ultimately, the AdMIRe challenge contributes to the ongoing effort to bridge the gap between human and machine language comprehension. This fosters progress in NLP applications such as machine translation, sentiment analysis, and the broader goal of understanding language.

Limitations

Dataset size The datasets for both subtasks are small in the context of contemporary machine learning data, which limited how much they could be used to train or fine-tune models. However, we believe that the manual curation of target items, context sentences, image prompts and generated images by native speakers means that the data quality is high. Individual idiomatic expressions are encountered infrequently, even in well-resourced languages such as English, yet are generally well-understood by human readers; we consider the AdMIRe dataset to represent a realistic yet achievable challenge for NLP systems.

Language variety The AdMIRe dataset contains items in only two languages, both of which are relatively well-resourced. Expanding the dataset to a greater diversity of languages could enable better evaluation of idiomaticity representation.

Cultural background The datasets were constructed by creators who are primarily middle-class and work in academic settings, and who are speakers of Brazilian Portuguese and/or British English. The same applies to most of the human annotators whose responses we used for evaluation. 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 Similarly, the datasets are likely to reflect biases and limitations present in the tools we used to construct them, especially the image generation and captioning models. While we made efforts to introduce diversity in the people depicted, we encountered some challenges in this regard. For instance, we were unable to identify prompts which would successfully generate images depicting a literal *one-armed bandit*[8]. Expressions like *blue blood* also presented difficulty, and we note that the depictions of female characters in particular in the dataset tend to be 'conventionally attractive' and have somewhat exaggerated facial features.

[8] Idiomatically, a slot machine.

Acknowledgments

The authors would like to acknowledge the contributions of our Brazilian collaborators who made this work possible. In particular, our thanks go to Rozane Rebechi, Juliana Carvalho, Eduardo Victor, César Rennó Costa, Marina Ribeiro and their colleagues and students.

We would also like to thank the members of the Multi-Word Expressions Group for their continued feedback and support.

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

This work also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

This work was supported by the EQUATE project.

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

Mohamed Badran, Youssof Nawar, and Nagwa El-Makky. 2025. AlexUNLP-NB at SemEval-2025 Task 1: A pipeline for idiom disambiguation and visual representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. 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.

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.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. *arXiv preprint*. ArXiv:2205.12404 [cs].

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *Preprint*, arXiv:2211.06679.

Judith Clymo, Adam Zernik, and Shubham Gaur. 2025. UCSC NLP T6 at SemEval-2025 Task 1: Leveraging LLMs and VLMs for idiomatic understanding. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. 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.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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

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.

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, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024a. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.

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

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, volume 2013, pages 125–127. Valladolid.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Saurav K. Aryal and Lawal Abdulmujeeb. 2025. Howard University-AI4PC at SemEval-2025 Task 1: Using GPT-4o and CLIP-ViLT to decode figurative language across text and images. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Maira Khatoon, Arooj Kiyani, Tehmina Doltana Farid, and Sadaf Abdul Rauf. 2025. FJWU_Squad at SemEval-2025 Task 1: An idiom visual understanding dataset for idiom learning. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. 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. *arXiv preprint*. ArXiv:2410.16461 [cs].

Liu Lei, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2025. YNU-HPCC at SemEval-2025 Task 1: Enhancing multimodal idiomaticity representation via LoRA and hybrid loss optimization. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Ann Maria Piho, Lara Eulenpesch, Julio Julio, Victoria Lohner, and Wiebke Petersen. 2025. Transformer25 at SemEval-2025 Task 1: A similarity-based approach. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Thanet Markchom, Tong Wu, Liting Huang, and Huizhi Liang. 2025. UoR-NCL at SemEval-2025 Task 1: Using generative LLMs and CLIP models for multilingual multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Blake Matheny, Phuong Minh Nguyen, and Minh Le Nguyen. 2025. JNLP at SemEval-2025 Task 1: Multimodal idiomaticity representation with large language models. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. *arXiv preprint*.

Joydeb Mondal and Pramir Sarkar. 2025. Modgenix at SemEval-2025 Task 1: Context aware vision language ranking (CAViLR) for multimodal idiomaticity understanding. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

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

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

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.

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.

Arash Rasouli, Erfan Sadraiye, Omid Ghahroodi, Hamid Reza Rabiee, and Ehsaneddin Asgari. 2025. AIMA at SemEval-2025 Task 1: Bridging text and image for idiomatic knowledge extraction via mixture of experts. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

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

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding Figurative Meaning through Explainable Visual Entailment. *arXiv preprint*. ArXiv:2405.01474 [cs].

Irena Spasić, Lowri Williams, and Andreas Buerki. 2020. Idiom-Based Features in Sentiment Analysis: Cutting the Gordian Knot. *IEEE Transactions on Affective Computing*, 11(2):189–199.

Syahrir Syahrir and St Hartina. 2021. The Analysis of Short Story Translation Errors (A Case Study of Types and Causes). *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 9(1). Number: 1.

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. *arXiv preprint*. ArXiv:2204.10050 [cs].

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.

Sumiko Teng and Chuen Shin Yong. 2025. ChuenSumi at SemEval-2025 Task 1: Sentence transformer models and processing idiomacity. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Lorenzo Vaiani, Davide Napolitano, and Luca Cagliero. 2025. PoliTo at SemEval-2025 Task 1: Beyond literal meaning: A chain-of-though approach for multimodal idiomacity understanding. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. 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.

Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Zehua Hao, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. 2024. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5390–5400.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yanan Wang, Dailin Li, Yicen Tian, Bo Zhang, Wang Jian, and Liang Yang. 2025. dutir914 at SemEval-2025 Task 1: An integrated approach for multimodal idiomaticity representations. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

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.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal. Association for Computational Linguistics.

Anar Yeginbergen, Elisa Sanchez, Andrea Jaunarena, and Ander Salaberria. 2025. HiTZ-Ixa at SemEval-2025 Task 1: Multimodal idiomatic language understanding. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

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, Mengyuan Zhou, XiaoLong Hou, and Lianxin Jiang. 2025. Pali-nlp at semeval-2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Yue Yu, Jiarong Tang, and Ruitong Liu. 2025. TueCL at SemEval-2025 Task 1: Admire: Advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2022. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.

Yingzhou Zhao, Bowen Guan, Liang Yang, and Hongfei Lin. 2025. Zhoumou at SemEval-2025 Task 1: Leveraging multimodal data augmentation and large language models for enhanced idiom understanding. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*. Association for Computational Linguistics.