UCSC NLP T6 at SemEval-2025 Task 1: Leveraging LLMs and VLMs for Idiomatic Understanding

Judith Clymo, Adam Zernik, Shubham Gaur
University of California, Santa Cruz
{jclymo, azernik, sgaur2}@ucsc.edu

Abstract

Idiomatic expressions pose a significant challenge for natural language models due to their non-compositional nature. In this work, we address Subtask 1 of the SemEval-2025 Task 1 (AdMIRe) (Pickard et al., 2025), which requires distinguishing between idiomatic and literal usages of phrases and identifying images that align with the relevant meaning. Our approach integrates large language models and vision-language models, and we show how different prompting techniques improve those models' ability to identify and explain the meaning of idiomatic language.

1 Introduction

Idiomatic expressions challenge natural language models as their meanings often defy the compositional rules of literal language. For example, *"a piece of cake"* (Figure 1) may literally refer to dessert but idiomatically means an easy task. Neural models struggle to differentiate these uses, as idiomaticity often requires contextual and cultural understanding. Linguistic theories suggest that idioms derive their meaning from real-world interactions, motivating multi-modal approaches that integrate text and images.

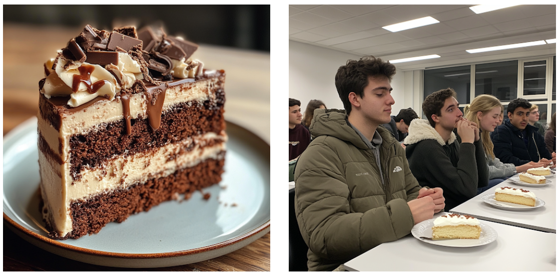

Figure 1: The result of prompting Midjourney with *"the dessert was a piece of cake"* and *"the exam was a piece of cake"*.

Understanding idiomaticity is vital for improving machine translation, sentiment analysis, and dialogue systems. SemEval AdMIRe task (Pickard et al., 2025) assesses idiomatic comprehension by ranking images based on how well they match the meaning of idiomatic or literal phrases

1.1 Task Details

Figure 2 illustrates the task setup (Pickard et al., 2025). We are given a **context sentence** containing a potentially idiomatic or literal **target phrase** along with five images. The task challenges us to rank these five images according to their semantic similarity to the meaning of the phrase in the context sentence.

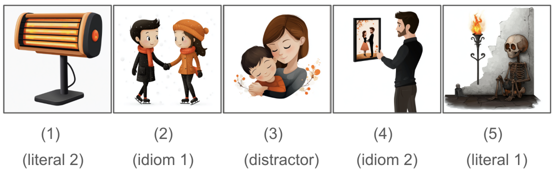

Figure 2: Image ranking based on semantic similarity to the target phrase *"old flame"* in the context sentence *"She ran into an old flame at the high school reunion"*. The correct order is *[4,2,1,5,3]*

The following metrics are used for evaluation:

- **Top-1 Accuracy:** Accuracy in selecting the correct highest-ranked image.
- **Discounted Cumulative Gain (DCG):** Weighted measure of ranking quality that discounts the impact of lower positions.

The task is available in both English and Portugese, however we focus only on the English version. We present a system that combines a large language model (LLM) for reasoning about idiomatic language with a vision-language model (VLM) for image ranking[1].

[1] GitHub: SemEval2025-Task1

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2103–2115
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 System description

2.1 Sentence Type Classification

A natural starting point is to identify whether the compound phrase in each context sentence was used in a literal or idiomatic sense. We use GPT-4 as our classifier, asking it to consider both potential meanings of the phrase while analyzing its usage in the given sentence.

2.2 Semantic Enhancement of Text Inputs

The use of carefully designed prompts to elicit targeted responses from large, pre-trained language models has shown promise in multiple domains (Liu et al., 2023). GPT-3 was among the first models to perform well with task-specific prompting in few-shot scenarios without requiring fine-tuning (Brown et al., 2020). More recently, query reformulation techniques have been shown to optimize inputs for pre-trained language models, with paraphrased inputs improving performance on downstream tasks (Haviv et al., 2021). Based on these insights, we designed an approach leveraging GPT-4 to produce context-aware definitions of the target phrase, testing a series of prompts that progressively layer prompting strategies, such as:

Manual Template Engineering Crafting a task-specific template using human introspection and domain knowledge has been shown to elicit contextually appropriate responses (Brown et al., 2020).

Prompt Augmentation Augmenting the prompt with few-shot examples improves performance by demonstrating the expected task behavior directly in the prompt (Liu et al., 2023).

Output Formatting Specifying a structured output format ensures consistency in responses and allows straightforward extraction of the final answer (Liu et al., 2023).

Chain-of-Thought This prompting technique (Wei et al., 2023) has been shown to improve performance on tasks requiring complex reasoning and contextual understanding, making this strategy particularly suited to idiomatic language.

Prompt Composition Complex tasks are decomposed in to smaller sub-tasks within a unified prompt (Liu et al., 2023).

The aim of our prompts is to obtain context-aware definitions of the compounds as they're used in the context sentences. The definitions are then used as the text inputs to the VLMs.

We present results below from two prompts used for generating definitions. Both prompts anchor GPT-4 in the role of a linguistics expert, include multiple examples and both require responses to follow a formal JSON schema specified in the prompt. Both prompts employ chain-of-thought reasoning but differ in their strategic approaches. Prompt 1 aims to find the ideal generalized definition of the idiom. GPT is asked to consider both literal and idiomatic definitions of the phrase before settling on a definition for the idiom that does not overlap with the literal meaning. Prompt 2 recognizes that even an ideal definition is not necessarily a useful image description, primarily because it might overgeneralize. Instead, it prompts the model to imagine five distinct scenes that depict the idiom then generate a single caption that is general enough to describe all five scenes.

Prompt 2 resulted in some interesting and descriptive outputs. For example, the Prompt 1 generated definition of *"graveyard shift"* is *"A late-night work schedule, often going until sunrise"*. While the definition is accurate, the inclusion of the word "often" points to generalization rather than specificity. In contrast, the result from Prompt 2 captures the idiom's meaning while describing a specific scene: *"Toiling through the night while the world sleeps"*. However sometimes the core meaning of the definition was diluted by this approach, for example, *"fancy dress"* received the definition *"Let your costume do the talking"*.

Both prompts are reproduced in the Appendix.

2.3 Image Alignment

To compare multi-modal inputs and find common themes we make use of models that are trained to match pairs of related text and images. Two encoders create embeddings for their inputs that are compared in batches using cosine similarity. Given n text and image inputs, we have a matrix of n^2 cosine similarities. In training, the categorical cross entropy loss is applied across both text and image dimensions. The image and text embedding spaces are therefore forced to have similar structure and the models are encouraged to extract similar information from the different modalities. Our experiments focus on the models CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and OpenCLIP (Ilharco et al., 2021).

CLIP has variants that use a ResNet or a vision transformer (ViT) for the image model. We experiment with both options, and with two different sized ViT encoders. ALIGN instead uses EfficientNet (Tan and Le, 2019) for the image encoder. Both have a transformer for the text encoder, however ALIGN uses the BERT architecture (bi-directional attention) and CLIP uses causally masked attention. The bigger difference between CLIP and ALIGN is in the data used to train them. CLIP is trained on about 400 million text-image pairs, and considerable effort was spent cleaning the data to ensure high quality. ALIGN used over 1 billion training examples but with less effort spent on data cleaning. Experiments on standard benchmarks suggest that the two models perform similarly well. The differences between CLIP and OpenCLIP are minimal, the latter being an open-source implementation of the former.

3 Results

The scores for our best performing model are shown in Table 1 and Table 2. We use GPT first to classify the usage type, then with Prompt 1 to define the meaning of the compounds that are used idiomatically. In the case of samples that are found to use their phrase literally, the compound is used directly without the context sentence or any definition. We used ALIGN to determine the final ranking of the five images from the provided text input.

	Top-1 Accuracy		
	Literal	**Idiomatic**	**All**
Test	0.86	0.88	0.87
Extended	0.74	0.61	0.68

Table 1: Performance results for Top-1 accuracy.

	DCG		
	Literal	**Idiomatic**	**All**
Test	3.34	3.47	3.41
Extended	3.29	3.11	3.20

Table 2: Performance results for DCG.

We show a comparison of different VLMs given this text input in Figures 3 and 4. ALIGN and OpenCLIP show stronger performance and greater consistency across the different metrics and data sets than the CLIP models tested, however none of the models was the clear winner across all settings.

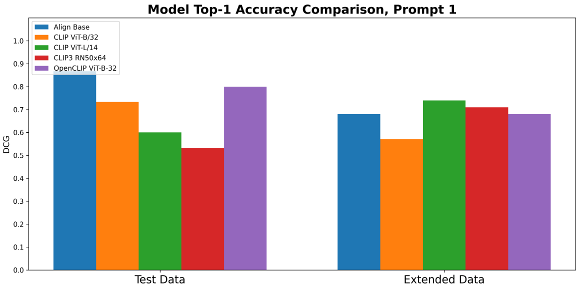

Figure 3: Top-1 accuracy for all models using GPT generated inputs with prompt 1.

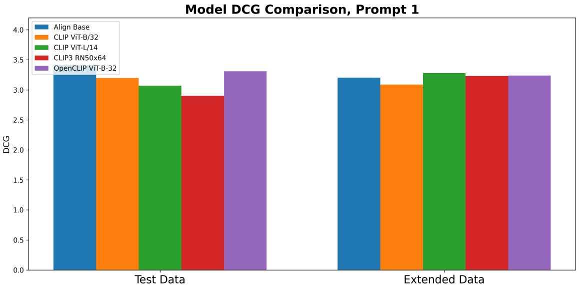

Figure 4: DCG for all models using GPT generated inputs with prompt 1.

3.1 Classification

The classification step using GPT is highly reliable. We benefited from an explicit classification step since this allowed us to use the compound phrase as input for the literal use samples, which consistently performed at least as well as using any generated definition. Further, we avoided the need to invoke the longer reasoning chain required to generate definitions for the literal samples.

		Predicted	
		Literal	**Idiomatic**
True	**Literal**	**52/7**	2/1
	Idiomatic	0/0	**46/8**

Table 3: Confusion matrix of classification results for literal and idiomatic expressions in extended/test sets.

3.2 Definition

Manually reviewing the definitions returned for the test set, we find that all of the definitions from

Prompt 1 correctly capture the meaning of the idiom. The weakest definition was for *"cold feet"*, *"Feeling scared or nervous before an important event"*. This is correct but misses the implication that a person with cold feet is thinking of not going ahead with the intended action or event.

A closer inspection of consistently low-scoring examples sheds light on where and why our system struggles. For instance, consider the phrase *"best man"*, used literally in the sentence, *"The best man means the quickest and most intelligent drive"*. Both the literal and idiomatic meanings of the phrase describe a man in a standout role, making it hard to visually separate one from the other. As a result, the model likely defaulted to the more familiar wedding-related meaning.

We see a similar failure case with *"eye candy"*. In the sentence, *"They gave me the impression that the development team has been focusing too much on eye candy rather than actual gameplay or level design,"* the idiomatic phrase criticizes style over substance. But our text input - *"something visually attractive but lacking depth"* - left too much room for literal interpretations. The literal images, including one showing colorful candies shaped like eyeballs, fit the figurative definition well enough to confuse the model.

4 Ablations

To demonstrate the value of the additional information provided by GPT, we test several other inputs to our VLMs.

4.1 Compound Only

In this experiment, only the compound phrase itself is given as the text input. This is insufficient information to complete the task, since the model cannot know whether the compound is meant to be understood literally or idiomatically. However, this serves to show the strength of the models' bias towards literal language. The performance in image ranking is already quite strong when the usage is literal, which demonstrates why in our final system we decided to use the compound as input when the usage has been classified as literal.

4.2 Sentence and Compound (Baseline)

The baseline experiment used the text input "{compound} in the context of {sentence}". All models continue to show much stronger performance on literal usage, which may also be because the images for literal use often incorporated other details from the sentence. Performance in the baseline experiments is summarized in Figure 5.

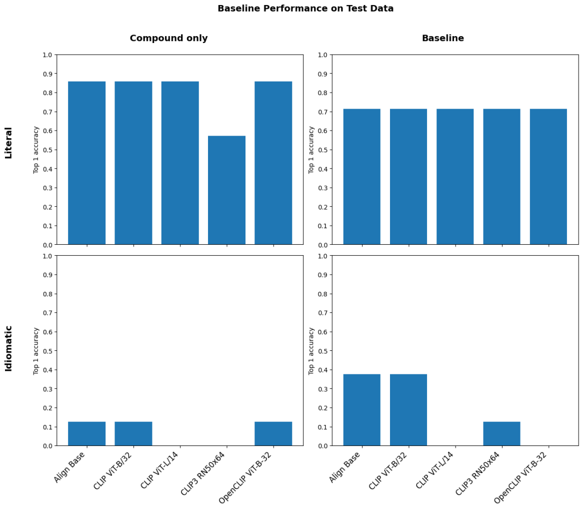

Figure 5: Top-1 accuracy for baseline experiments, test data.

4.3 Classification Only (Ablation 1)

We use GPT to classify the type of usage but do not generate any definition of the phrase. The text input for the VLMs is "{compound} in its {classification} sense". Although this input appears to remove some of the reasoning burden, it does not result in better ranking of the images compared to the baseline.

4.4 Zero Shot Prompt (Ablation 2)

We ask GPT for a definition of the compound in the simplest possible way, using the result as input for the VLMs. The prompt was "define {compound} as it's used in {sentence}". We see better performance on sentences with idiomatic use at the expense of reduced performance on literal use sentences.

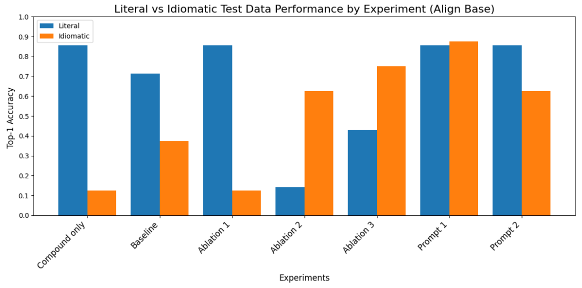

Figure 6: Top-1 accuracy for test data across all ablations, using ALIGN for text-image comparisons.

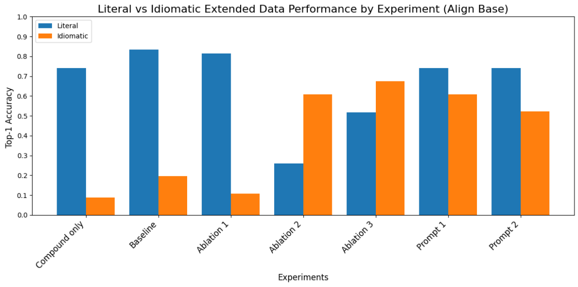

Figure 7: Top-1 accuracy for extended data across all ablations, using ALIGN for text-image comparisons.

4.5 Zero Shot Prompt with Classification (Ablation 3)

We ask GPT for a definition of the compound using a prior classification (also from GPT): “define {compound} in its {classification} sense”. Overall performance is improved, however it remains lower than the baseline for literal samples.

4.6 Ablation Results

Figure 6 and Figure 7 show the performance of ALIGN on the test and extended data sets respectively across the different ablation experiments described above and using the definitions generated by the two prompts described in Section 2.2. We noted similar trends in performance across all five VLMs tested.

We see that the definitions generated by Prompt 1 and Prompt 2 result in differing performance on the image ranking task, with Prompt 1 giving the better results, although this was not consistent across all models tested. For the idiomatic use samples it is unclear whether the definitions from these more complex prompts offer significant improvement compared to those from Ablation 3.

Full results of all ablation experiments are presented in the Appendix.

5 Discussion

We initially expected that a separate, fine-tuned model would be needed to classify the type of language, with queries to GPT only used to generate definitions for the idiomatic phrases. However, it turned out that GPT was highly effective for the classification task.

Despite impressive zero-shot performance on the literal inputs, none of the VLMs we tested was able to perform well for the idiomatic context sentences. It was surprising that given only the target phrase the models strongly preferred the literal use. Most of the phrases are very commonly used idiomatically and several seem unnatural in attempted literal usage. This is likely due to the specific purpose and training of the VLMs we used. Datasets built from tasks such as image captioning, for example, will tend to have a bias towards literal, descriptive language rather than poetic or abstract language. GPT instead preferred the idiomatic definition when prompted with only the phrase and was more likely to mistake literal usage for idiomatic, which better reflects the most common uses of such phrases.

Across our different experiments we saw that some samples were consistently easier for the VLMs to work with, regardless of the exact form of the text input. Figure 8 shows the combined top-1 accuracy for each sample across all experiments. The data points are colored according to the compound's usage, showing again that the literal use samples were in general easier than the idiomatic use samples.

A major limitation of our approach is that GPT does not respond in exactly the same way each time, even when given an identical prompt. It is difficult to fully understand the relationship between prompt and output, and this is further complicated in the present task by the output being then passed through another language model in the image ranking task. For the classification task, simple voting helps to mitigate this. A more complex voting algorithm could be used to combine multiple attempts at image ranking, for example finding a ranking whose total deviation from each of several individual ranks is minimized.

In addition, model updates will alter the behavior with respect to a prompt, sometimes in unexpected ways. To manage these changes over time, behavior on a training data set can be monitored and the prompts updated if average performance drops below some threshold. Language models can propose prompt modifications, so there is potential for these prompt updates to be applied automatically.

Many of the techniques we explored are applicable to a wide variety of tasks. Using language models to rephrase an input into a simple, direct and possibly structured form prior to further processing aids tool use. Chaining language model calls is also common in agentic frameworks.

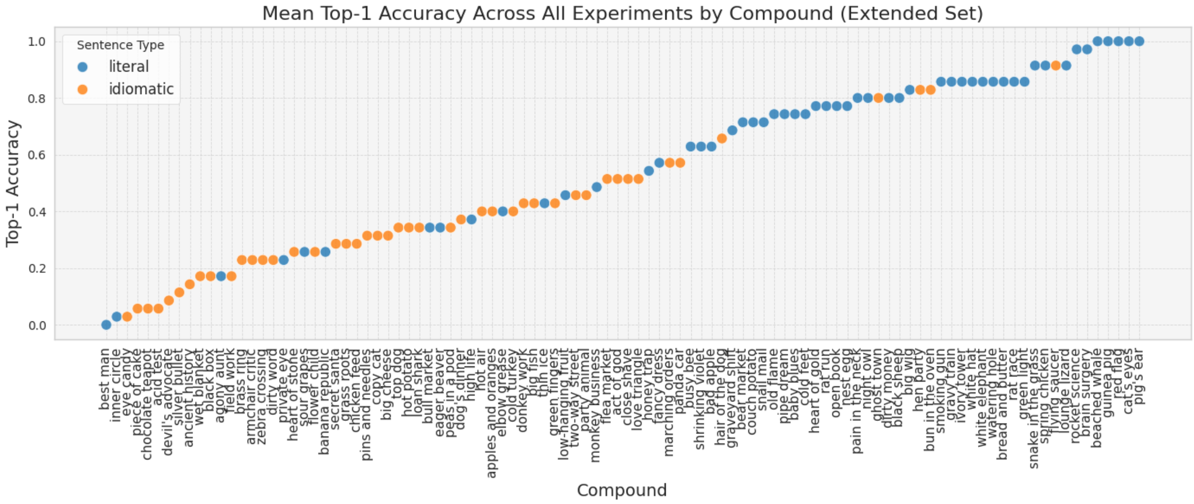

Figure 8: Average top-1 accuracy for every compound across all examples.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Amir Haviv, Reut Tsarfaty, and Hila Gonen. 2021. Bertese: Learning to speak to bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3592–3607. Association for Computational Linguistics.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, Carolina Scarton, and Marco Idiart. 2025. Semeval-2025 task 1: Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Proceedings of Machine Learning Research. 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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

A Prompt 1

Listing 1: The Python string template for Prompt 1.

```
prompt = """
You are a linguistics expert
    specializing in idioms. You will be
    given a set of idioms to process.
    For each one, do the following steps
     aloud (in writing):
1. Give a verbose explanation of the
    idiom, including what connotations
    it carries or undertones it evokes.
2. Give a definition of the *literal*
    meaning of the phrase. For noun
    phrases representing physical
    objects, focus on unambiguous visual
     descriptors.
```

```
3. Taking into consideration your
    response for #1 and #2, list three
    potential definitions, no longer
    than 20 words each, that capture the
     **core emotional or situational
    essence** conveyed by the idiom. Use
     **simple language that an average
    high-schooler would understand** and
     avoid figurative or overly abstract
     language. Focus on clear, visually
    interpretable descriptions that are
    distinct from the literal definition
    .
4. Choose the best definition.

---

Example outputs:
{{
  "data": [
    {{
      "target_phrase": "glass ceiling",
      "explanation": "Refers to an
          invisible barrier that
          prevents certain groups, often
           women or minorities, from
          advancing in their careers or
          social positions. Evokes
          frustration, inequality, and
          hidden obstacles. Frequently
          used in discussions of
          systemic discrimination.",
      "literal_definition": "A ceiling
          made of transparent glass.",
      "potential_definition_1": "A
          hidden obstacle that blocks
          people from reaching higher
          positions.",
      "potential_definition_2": "An
          unseen barrier that stops
          progress for qualified
          individuals.",
      "potential_definition_3": "A quiet
           limit that keeps certain
          groups from moving upward.",
      "result": "A hidden obstacle that
          blocks people from reaching
          higher positions."
    }},
    {{
      "target_phrase": "missing link",
      "explanation": "Suggests a crucial
           piece of information or
          evidence needed to bridge a
          gap in knowledge or
          understanding. Evokes the
          sense of an incomplete puzzle,
           emphasizing the importance of
           finding whats absent.",
      "literal_definition": "A link in a
           chain that is not present,
          creating a gap.",
      "potential_definition_1": "A key
          piece that completes an
          unfinished idea or puzzle.",
      "potential_definition_2": "
          Something crucial that holds
          everything together but is
          absent.",
      "potential_definition_3": "An
          important connecting factor
          that is missing or unknown.",
      "result": "A key piece that
          completes an unfinished idea
          or puzzle."
    }},
    {{
      "target_phrase": "paper tiger",
      "explanation": "Describes someone
          or something that appears
          threatening or powerful but is
           actually weak or ineffective.
           Connotes empty threats or
          superficial strength.",
      "literal_definition": "A tiger
          made of paper, such as origami
           or a paper figure.",
      "potential_definition_1": "
          Something that seems strong
          but has little real power.",
      "potential_definition_2": "A
          fragile threat that looks more
           dangerous than it is.",
      "potential_definition_3": "A force
           that seems scary but
          collapses under pressure.",
      "result": "Something that seems
          strong but has little real
          power."
    }}
    ...
  ]
}}

---

You must return a valid JSON object:
- Do not use double quotes inside your
    value strings.
- Do not include line breaks inside JSON
     values.
- Strictly follow the schema.

Schema:
{{
  "type": "object",
  "properties": {{
    "data": {{
      "type": "array",
      "items": {{
        "type": "object",
        "properties": {{
          "target_phrase": {{ "type": "
              string" }},
          "explanation": {{ "type": "
              string" }},
          "literal_definition": {{ "type
              ": "string" }},
          "potential_definition_1": {{ "
              type": "string" }},
          "potential_definition_2": {{ "
              type": "string" }},
          "potential_definition_3": {{ "
              type": "string" }},
          "result": {{ "type": "string"
              }}
        }},
        "required": ["target_phrase", "
            explanation", "
            potential_definition_1", "
            potential_definition_2", "
            potential_definition_3", "
            result"]
        }}
      }}
    }},
    "required": ["data"]
}}

Ensure the response is a valid JSON
    object with escaped quotes.

Here are the samples:
"""
```

B Prompt 2

Listing 2: The Python string template for Prompt 2.

```
prompt = """You are a linguistics and
    visual storytelling expert, with an
    expertise on differentiating
    idiomatic from literal language. For
     each sample idiom below, your task
    is to create visual and textual
    representations that align well with
     the idioms figurative meaning
    for use in matching with images.
    Follow these steps:

1. Identify the phrase: Give a concise
    definition of the phrase in its
    idiomatic sense.
2. Note the literal usage (briefly):
    Mention the plain or surface meaning
    , but clarify that you are focusing
    on the figurative interpretation for
     your examples.
3. Generate 5 distinct image ideas: For
    the given idiom, imagine 5 different
     scenes or situations that visually
    depict its figurative meaning.
    Describe each scene in 1-2 sentences
    , focusing on visual details.
4. Generalize the captions: Write a
    single caption that could apply to
    all 5 scenes. It should capture the
    essence of the idiom in a way that
    is broad enough to fit any of the
    scenes.
5. Refine: Reflect on how well your
    caption generalizes to all five
    scenes, then attempt to improve on
    it.
6. Consider which caption is best: Weigh
     the captions against each other,
    then pick the one that best fits all
     5 scenes.
7. Select the best caption: Repeat the
    caption you selected.

---

Example outputs:
{
  "data": [
    {
      "target_phrase": "glass ceiling",
      "explanation": "Refers to an
          invisible barrier that
          prevents certain groups (often
           women or minorities) from
          advancing to higher levels of
          power or responsibility.
          Implies a hidden form of
          discrimination that is not
          overtly acknowledged but still
           limits upward mobility.",
      "literal_definition": "A ceiling
          made of glass.",
      "image_ideas": [
        "A businesswoman standing just
            below a transparent barrier
            in a large corporate office,
             looking up at executives in
             the floor above.",
        "A group of female or minority
            employees reaching a fancy
            mezzanine level only to find
             an unseen barrier between
            them and the boardroom.",
        "A symbolic representation of
            cracks forming in a
            transparent barrier overhead
             as a woman holds a
            briefcase, showing
            determination to break
            through.",
        "A silhouette of a person
            pressed against a clear pane
            , with a hand raised as
            though trying to push past
            it.",
        "A visually layered office
            setting, where higher floors
             are accessible but
            separated by a nearly
            invisible division,
            highlighting the subtlety of
             the barrier."
      ],
      "generalized_caption_1": "Facing
          an unseen barrier to
          advancement.",
      "generalized_caption_2": "Pushing
          against a hidden boundary in
          pursuit of progress.",
      "thinking": "Both captions address
           the concept of a hidden
          obstruction. The second one, '
          Pushing against a hidden
          boundary in pursuit of
          progress,' suggests active
          resistance and forward motion,
           which suits the idioms
          connotation of striving to
          break through.",
      "result": "Pushing against a
          hidden boundary in pursuit of
          progress."
    },
    {
      "target_phrase": "paper tiger",
      "explanation": "Describes someone
          or something that appears
          threatening or powerful but is
           actually weak or ineffectual.
           Connotes false bravado or an
          overestimation of strength.",
      "literal_definition": "A tiger
          made out of paper.",
      "image_ideas": [
        "A large, menacing figure
            looming over a crowd, only
            to be revealed as hollow or
            easily torn.",
        "A roaring tiger image on a
            billboard that looks scary
            but is just thin paper
            peeling at the edges.",
        "A towering cardboard cutout of
            a tiger in a political rally
            , symbolizing empty threats
            or exaggerated power.",
        "A fierce-looking trophy made of
             paper mache, displayed in a
             spotlight to highlight its
            fragile nature.",
        "An intimidating sign with a
            tiger illustration in front
            of a building, but the sign
            is tattered and flapping in
            the wind, showing its
            vulnerability."
      ],
      "generalized_caption_1": "A
          formidable appearance that
          masks a fragile reality.",
      "generalized_caption_2": "
          Something that looks strong
          but lacks real power.",
      "thinking": "The second caption
          directly addresses the core
          meaning 'Something that
          looks strong but lacks real
          power.' It's concise and
          precise.",
      "result": "Something that looks
          strong but lacks real power."
    },
    {
      "target_phrase": "missing link",
      "explanation": "Refers to a
          crucial piece of information
          or element that helps connect
          different ideas, theories, or
          facts. Connotes something
          vital that completes a puzzle
          or fills a gap in
          understanding.",
      "literal_definition": "A link in a
           chain (like a ring or segment
          ) that is absent.",
      "image_ideas": [
        "A detective at a crime board
            tapping a blank space among
            photos and clues, indicating
             a vital piece of evidence
             thats not yet found.",
        "An evolutionary chart with a
            silhouette in the middle
            missing, leaving a gap in
            the progression from ape to
            human.",
        "A jigsaw puzzle nearly
            completed, except for a
            conspicuously empty spot in
            the center.",
```

```
        "A timeline pinned on a wall
            with a significant date
            missing, highlighting the
            gap in recorded history.",
        "A scientific lab setting where
            a researcher stands before a
             half-finished hypothesis,
            gazing at a large question
            mark on the board."
      ],
      "generalized_caption_1": "A
          crucial piece that completes
          the bigger picture.",
      "generalized_caption_2": "The
          vital connecting factor that
          brings everything together.",
      "thinking": "Between the two, 'A
          crucial piece that completes
          the bigger picture' fits the
          notion of something vital and
          absent, capturing the
          idiomatic essence succinctly
          .",
      "result": "A crucial piece that
          completes the bigger picture."
    }
    ...
    ...
    ...
  ]
}

---

You must return a valid JSON object:
- Do not use double quotes inside your
    value strings.
- Do not include line breaks inside JSON
     values.
- Strictly follow the schema.

Schema:
{
  "type": "object",
  "properties": {
    "data": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "target_phrase": { "type": "
              string" },
          "explanation": { "type": "
              string" },
          "literal_definition": { "type
              ": "string" },
          "image_ideas": { "type": "
              array", "items": { "type":
               "string" } },
          "generalized_caption_1": { "
              type": "string" },
          "generalized_caption_2": { "
              type": "string" },
          "thinking": { "type": "string"
               },
          "result": { "type": "string" }
        },
        "required": ["target_phrase", "
            image_ideas", "
            generalized_caption_1", "
```

```
                generalized_caption_2", "
                thinking", "result"]
            }
        }
    },
    "required": ["data"]
}

Ensure the response is a valid JSON
    object with properly escaped quotes.

Your turn. Here are the samples:
"""
```

C Experimental Results

In this section, we present the results across all 6 sets of experiments with different approaches. We evaluate the performance of various models using three key metrics: **Top-1 Accuracy**, **Spearman Correlation**, and **Discounted Cumulative Gain (DCG)**. For each metric, we provide results for both **idiomatic** and **literal** subsets of the data. The models evaluated include:

- **Align**: Align Base
- **CLIP1**: CLIP ViT-B/32
- **CLIP2**: CLIP ViT-L/14
- **CLIP3**: CLIP3 RN50x64
- **OpenClip**: OpenCLIP ViT-B-32

The results are divided into two sets: **Test** and **Extended**. Each set contains three tables, one for each metric. Below, we present the results in detail.

C.1 Test Set Results

Top-1 Accuracy (Table 4)

Spearman Correlation (Table 5)

Discounted Cumulative Gain (Table 6)

C.2 Extended Dataset Results

Top-1 Accuracy (Table 7)

Spearman Correlation (Table 8)

Table 3: Discounted Cumulative Gain (Table 9)

Experiment	**Model**	**All**	**Literal**	**Idiom**
Compound Only	Align	0.47	0.86	0.13
Compound Only	CLIP1	0.47	0.86	0.13
Compound Only	CLIP2	0.40	0.86	0.00
Compound Only	CLIP3	0.27	0.57	0.00
Compound Only	OpenClip	0.47	0.86	0.13
Baseline	Align	0.53	0.71	0.38
Baseline	CLIP1	0.53	0.71	0.38
Baseline	CLIP2	0.33	0.71	0.00
Baseline	CLIP3	0.40	0.71	0.13
Baseline	OpenClip	0.33	0.71	0.00
Ablation 1	Align	0.47	0.86	0.13
Ablation 1	CLIP1	0.40	0.86	0.00
Ablation 1	CLIP2	0.33	0.57	0.13
Ablation 1	CLIP3	0.27	0.57	0.00
Ablation 1	OpenClip	0.40	0.71	0.13
Ablation 2	Align	0.40	0.14	0.63
Ablation 2	CLIP1	0.53	0.43	0.63
Ablation 2	CLIP2	0.60	0.43	0.75
Ablation 2	CLIP3	0.47	0.29	0.63
Ablation 2	OpenClip	0.33	0.14	0.50
Ablation 3	Align	0.60	0.43	0.75
Ablation 3	CLIP1	0.53	0.43	0.63
Ablation 3	CLIP2	0.60	0.43	0.75
Ablation 3	CLIP3	0.47	0.29	0.63
Ablation 3	OpenClip	0.47	0.29	0.63
Prompt 1	Align	0.87	0.86	0.88
Prompt 1	CLIP1	0.73	0.86	0.63
Prompt 1	CLIP2	0.60	0.86	0.38
Prompt 1	CLIP3	0.53	0.57	0.50
Prompt 1	OpenClip	0.80	0.86	0.75
Prompt 2	Align	0.73	0.86	0.63
Prompt 2	CLIP1	0.67	0.86	0.50
Prompt 2	CLIP2	0.73	0.86	0.63
Prompt 2	CLIP3	0.60	0.57	0.63
Prompt 2	OpenClip	0.73	0.86	0.63

Table 4: Top-1 Accuracy results for the test dataset.

Experiment	**Model**	**All**	**Literal**	**Idiom**
Compound Only	Align	0.13	0.40	-0.10
Compound Only	CLIP1	-0.16	-0.24	-0.09
Compound Only	CLIP2	0.15	0.36	-0.04
Compound Only	CLIP3	0.09	0.21	-0.01
Compound Only	OpenClip	0.01	0.04	-0.03
Baseline	Align	0.34	0.46	0.24
Baseline	CLIP1	0.34	0.46	0.24
Baseline	CLIP2	0.17	0.46	-0.07
Baseline	CLIP3	0.23	0.27	0.19
Baseline	OpenClip	0.10	0.09	0.11
Ablation 1	Align	0.13	0.17	0.10
Ablation 1	CLIP1	0.09	0.10	0.09
Ablation 1	CLIP2	0.20	0.44	-0.01
Ablation 1	CLIP3	0.15	0.44	-0.11
Ablation 1	OpenClip	-0.02	-0.09	0.04
Ablation 2	Align	0.04	-0.06	0.12
Ablation 2	CLIP1	-0.09	-0.40	0.19
Ablation 2	CLIP2	0.29	0.09	0.48
Ablation 2	CLIP3	0.22	0.07	0.35
Ablation 2	OpenClip	0.13	-0.01	0.25
Ablation 3	Align	0.29	0.24	0.32
Ablation 3	CLIP1	0.20	0.23	0.18
Ablation 3	CLIP2	0.23	0.40	0.09
Ablation 3	CLIP3	0.16	0.00	0.30
Ablation 3	OpenClip	-0.03	-0.10	0.04
Prompt 1	Align	0.42	0.40	0.44
Prompt 1	CLIP1	-0.03	-0.24	0.16
Prompt 1	CLIP2	0.31	0.36	0.27
Prompt 1	CLIP3	0.25	0.21	0.27
Prompt 1	OpenClip	0.14	0.04	0.22
Prompt 2	Align	0.28	0.40	0.18
Prompt 2	CLIP1	-0.14	-0.24	-0.05
Prompt 2	CLIP2	0.29	0.36	0.23
Prompt 2	CLIP3	0.18	0.21	0.15
Prompt 2	OpenClip	-0.03	0.04	-0.10

Table 5: Spearman Correlation results for the test dataset.

Experiment	**Model**	**All**	**Literal**	**Idiom**
Compound Only	Align	2.71	3.34	2.15
Compound Only	CLIP1	2.67	3.34	2.07
Compound Only	CLIP2	2.63	3.38	1.98
Compound Only	CLIP3	2.41	2.93	1.96
Compound Only	OpenClip	2.68	3.33	2.12
Baseline	Align	2.91	3.32	2.55
Baseline	CLIP1	2.91	3.32	2.55
Baseline	CLIP2	2.58	3.29	1.95
Baseline	CLIP3	2.68	3.27	2.17
Baseline	OpenClip	2.59	3.30	1.98
Ablation 1	Align	2.70	3.34	2.15
Ablation 1	CLIP1	2.62	3.32	2.01
Ablation 1	CLIP2	2.60	3.12	2.16
Ablation 1	CLIP3	2.51	3.04	2.04
Ablation 1	OpenClip	2.63	3.17	2.16
Ablation 2	Align	2.73	2.31	3.10
Ablation 2	CLIP1	3.02	2.85	3.17
Ablation 2	CLIP2	3.00	2.74	3.23
Ablation 2	CLIP3	2.85	2.51	3.14
Ablation 2	OpenClip	2.71	2.31	3.07
Ablation 3	Align	3.10	2.81	3.35
Ablation 3	CLIP1	3.01	2.80	3.21
Ablation 3	CLIP2	3.11	2.88	3.32
Ablation 3	CLIP3	2.88	2.70	3.03
Ablation 3	OpenClip	2.93	2.66	3.18
Prompt 1	Align	3.41	3.34	3.47
Prompt 1	CLIP1	3.20	3.34	3.07
Prompt 1	CLIP2	3.07	3.38	2.80
Prompt 1	CLIP3	2.90	2.93	2.87
Prompt 1	OpenClip	3.31	3.33	3.30
Prompt 2	Align	3.17	3.34	3.03
Prompt 2	CLIP1	3.03	3.34	2.76
Prompt 2	CLIP2	3.16	3.38	2.97
Prompt 2	CLIP3	2.96	2.93	2.99
Prompt 2	OpenClip	3.14	3.33	2.97

Table 6: Discounted Cumulative Gain results for the test dataset.

Experiment	Model	All	Literal	Idiom
Compound Only	Align	0.44	0.74	0.09
Compound Only	CLIP1	0.46	0.76	0.11
Compound Only	CLIP2	0.49	0.83	0.09
Compound Only	CLIP3	0.52	0.85	0.13
Compound Only	OpenClip	0.48	0.81	0.09
Baseline	Align	0.54	0.83	0.20
Baseline	CLIP1	0.54	0.83	0.20
Baseline	CLIP2	0.51	0.80	0.17
Baseline	CLIP3	0.54	0.80	0.24
Baseline	OpenClip	0.52	0.87	0.11
Ablation 1	Align	0.49	0.81	0.11
Ablation 1	CLIP1	0.48	0.76	0.15
Ablation 1	CLIP2	0.51	0.81	0.15
Ablation 1	CLIP3	0.54	0.87	0.15
Ablation 1	OpenClip	0.48	0.81	0.09
Ablation 2	Align	0.42	0.26	0.61
Ablation 2	CLIP1	0.34	0.28	0.41
Ablation 2	CLIP2	0.45	0.37	0.54
Ablation 2	CLIP3	0.46	0.35	0.59
Ablation 2	OpenClip	0.40	0.22	0.61
Ablation 3	Align	0.59	0.52	0.67
Ablation 3	CLIP1	0.51	0.48	0.54
Ablation 3	CLIP2	0.51	0.48	0.54
Ablation 3	CLIP3	0.56	0.44	0.70
Ablation 3	OpenClip	0.56	0.50	0.63
Prompt 1	Align	0.68	0.74	0.61
Prompt 1	CLIP1	0.57	0.76	0.35
Prompt 1	CLIP2	0.74	0.83	0.63
Prompt 1	CLIP3	0.71	0.85	0.54
Prompt 1	OpenClip	0.68	0.81	0.52
Prompt 2	Align	0.64	0.74	0.52
Prompt 2	CLIP1	0.59	0.76	0.39
Prompt 2	CLIP2	0.63	0.83	0.39
Prompt 2	CLIP3	0.66	0.85	0.43
Prompt 2	OpenClip	0.63	0.81	0.41

Table 7: Top-1 Accuracy results for the extended dataset

Experiment	Model	All	Literal	Idiom
Compound Only	Align	0.24	0.41	0.03
Compound Only	CLIP1	0.16	0.32	-0.03
Compound Only	CLIP2	0.13	0.31	-0.08
Compound Only	CLIP3	0.21	0.39	-0.01
Compound Only	OpenClip	0.10	0.30	-0.14
Baseline	Align	0.28	0.52	0.00
Baseline	CLIP1	0.28	0.52	0.00
Baseline	CLIP2	0.22	0.40	0.01
Baseline	CLIP3	0.26	0.40	0.10
Baseline	OpenClip	0.13	0.34	-0.11
Ablation 1	Align	0.19	0.42	-0.08
Ablation 1	CLIP1	0.13	0.31	-0.08
Ablation 1	CLIP2	0.11	0.34	-0.15
Ablation 1	CLIP3	0.20	0.43	-0.07
Ablation 1	OpenClip	0.11	0.27	-0.07
Ablation 2	Align	0.18	0.16	0.20
Ablation 2	CLIP1	0.04	0.05	0.02
Ablation 2	CLIP2	0.14	0.09	0.19
Ablation 2	CLIP3	0.14	0.16	0.12
Ablation 2	OpenClip	0.08	0.00	0.17
Ablation 3	Align	0.20	0.14	0.26
Ablation 3	CLIP1	0.09	0.04	0.16
Ablation 3	CLIP2	0.08	0.10	0.05
Ablation 3	CLIP3	0.17	0.14	0.20
Ablation 3	OpenClip	0.11	0.05	0.18
Prompt 1	Align	0.30	0.41	0.17
Prompt 1	CLIP1	0.25	0.32	0.18
Prompt 1	CLIP2	0.22	0.31	0.11
Prompt 1	CLIP3	0.26	0.39	0.10
Prompt 1	OpenClip	0.25	0.30	0.20
Prompt 2	Align	0.28	0.41	0.11
Prompt 2	CLIP1	0.25	0.32	0.17
Prompt 2	CLIP2	0.22	0.31	0.12
Prompt 2	CLIP3	0.32	0.39	0.23
Prompt 2	OpenClip	0.19	0.30	0.06

Table 8: Spearman Correlation results for the extended dataset

Experiment	**Model**	**All**	**Literal**	**Idiom**
Compound Only	Align	2.70	3.29	2.02
Compound Only	CLIP1	2.74	3.31	2.06
Compound Only	CLIP2	2.80	3.43	2.07
Compound Only	CLIP3	2.83	3.44	2.10
Compound Only	OpenClip	2.76	3.41	2.00
Baseline	Align	2.90	3.43	2.28
Baseline	CLIP1	2.90	3.43	2.28
Baseline	CLIP2	2.86	3.41	2.22
Baseline	CLIP3	2.91	3.39	2.35
Baseline	OpenClip	2.86	3.46	2.15
Ablation 1	Align	2.77	3.36	2.07
Ablation 1	CLIP1	2.79	3.32	2.16
Ablation 1	CLIP2	2.83	3.40	2.16
Ablation 1	CLIP3	2.86	3.44	2.18
Ablation 1	OpenClip	2.79	3.43	2.04
Ablation 2	Align	2.73	2.40	3.12
Ablation 2	CLIP1	2.62	2.43	2.84
Ablation 2	CLIP2	2.74	2.53	2.98
Ablation 2	CLIP3	2.78	2.59	3.01
Ablation 2	OpenClip	2.71	2.32	3.16
Ablation 3	Align	3.00	2.85	3.18
Ablation 3	CLIP1	2.91	2.83	3.01
Ablation 3	CLIP2	2.91	2.84	2.99
Ablation 3	CLIP3	2.97	2.83	3.14
Ablation 3	OpenClip	3.01	2.87	3.17
Prompt 1	CLIP1	3.09	3.31	2.83
Prompt 1	CLIP2	3.28	3.43	3.10
Prompt 1	CLIP3	3.23	3.44	2.98
Prompt 1	OpenClip	3.24	3.41	3.03
Prompt 2	Align	3.14	3.29	2.98
Prompt 2	CLIP1	3.03	3.31	2.71
Prompt 2	CLIP2	3.14	3.43	2.79
Prompt 2	CLIP3	3.14	3.44	2.79
Prompt 2	OpenClip	3.14	3.41	2.83

Table 9: Discounted Cumulative Gain results for the extended dataset