I Came, I Saw, I Explained: Benchmarking Multimodal LLMs on Figurative Meaning in Memes

Shijia Zhou **Saif M. Mohammad** **Barbara Plank** **Diego Frassinelli**
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
National Research Council Canada, Ottawa, Canada
{zhou.shijia, b.plank, diego.frassinelli}@lmu.de

Abstract

Internet memes represent a popular form of multimodal online communication and often use figurative elements to convey layered meaning through the combination of text and images. However, it remains largely unclear how multimodal large language models (MLLMs) combine and interpret visual and textual information to identify figurative meaning in memes. To address this gap, we evaluate eight state-of-the-art generative MLLMs across three datasets on their ability to detect and explain six types of figurative meaning. In addition, we conduct a human evaluation of the explanations generated by these MLLMs, assessing whether the provided reasoning supports the predicted label and whether it remains faithful to the original meme content. Our findings indicate that all models exhibit a strong bias to associate a meme with figurative meaning, even when no such meaning is present. Qualitative analysis further shows that correct predictions are not always accompanied by faithful explanations.

Keywords: figurative meaning, internet memes, multimodal large language model

1. Introduction

On social media such as Reddit and X (Twitter), people actively use memes, which combine visual and textual components, to express a wide range of opinions. Memes are posted in debate on global issues, such as climate change (Zhou et al., 2025), as well as in playful reflections on daily life's small triumphs and troubles (Hwang and Shwartz, 2023). Many of these memes rely on multimodal figurative expressions, to convey meanings in creative, often non-literal ways. For example, Figure 1 shows a meme from MET-MEME (Xu et al., 2022), which humorously conveys a witty or self-deprecating attitude. It parodies Julius Caesar's famous quote "I came, I saw, I conquered" by keeping the initial "I came, I saw", which sets up the expectation of a powerful declaration, but replaces the last part with "I complained". Together with the accompanying grumpy cat, this subverts the expectation and creates a humorous contrast.

The presence of figurative meaning in memes represents a challenging aspect of multimodal communication,[1] and has become a phenomenon of growing interest in NLP. Liu et al. (2022a) first introduced FIGMEMES, a dataset for classifying six types of figurative expressions in politically opinionated memes, covering diverse themes and visual styles. Metaphor processing attracts attention in meme interpretation tasks (Xu et al., 2022; Hwang and Shwartz, 2023; Xu et al., 2024), and several studies have linked offensive memes with sarcastic expressions (Sharma et al., 2020; Pramanick et al., 2022; Kumari et al., 2024).

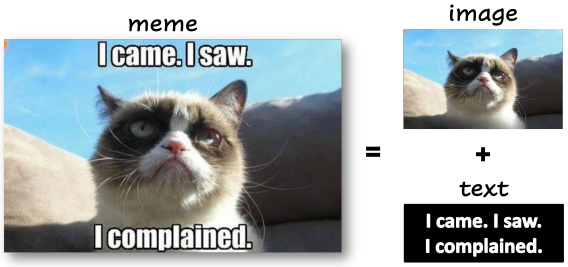

Figure 1: A meme from MET-MEME (Xu et al., 2022) playfully adapts Caesar's famous quote "I came, I saw, I conquered". In our work, we investigate the respective contributions of image and text to models' prediction of figurative meaning in memes.

However, except for FIGMEMES (Liu et al., 2022a), figurative meaning in memes remains understudied. Little work has investigated to what extent generative multimodal large language models (MLLMs) can detect figurative meaning in memes, and how this ability varies across different types, such as allusion, irony, and metaphor. There is also a lack of systematic investigation into how controlled manipulations of information from different modalities affect figurative meaning predictions on memes.

[1] In this paper, we use the term *figurative meaning* instead of the more commonly used *figurative language* to emphasize the multimodal nature of the phenomenon we are analyzing.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9460–9477
11–16 May 2026.

Recently, natural language explanations have become an important additional signal for studying how a model processes an input, e.g., for interpreting figurative meaning (Saakyan et al., 2025). However, there is a lack of well-defined criteria for evaluating the quality of model-generated explanations of figurative meaning.

To address these gaps, we define a combined *detection and explanation* task for memes, requiring models not only to identify one or more types of figurative meaning, but also provide a corresponding explanation that justifies their predictions. Our contributions are:

- We introduce a detection and explanation task for figurative meaning in memes, and evaluate the performance of eight models on both *binary* and *multi-label* classification tasks.
- We systematically investigate the impact and contribution of information from *different modalities* on model behavior through controlled input manipulations, and two prompting setups.
- We conduct a human study to assess how coherently model-provided explanations are perceived for their respective predicted labels, and how well they are grounded in the meme's visual and textual components.

Our benchmarking of eight state-of-the-art MLLMs reveals three core findings regarding their interpretation of meme figurativeness. First, we identify a pervasive bias where models consistently over-assign figurative meanings to memes that are intended to be literal. Second, model performance and modality reliance are highly type-specific: while irony and sarcasm are heavily visually grounded, metaphor detection uniquely requires a complex interplay between textual and visual cues. Finally, our human evaluation uncovers a significant *faithfulness gap*: even when models predict correct labels, their explanations often suffer from visual hallucinations, over-interpretation, or an inability to grasp the underlying socio-psychological nuances of human behavior. Our code and evaluation are publicly available at `https://github.com/mainlp/Figurative-Meaning-in-Memes`.

2. Related Work

Figurative Language in NLP Figurative language is characterized by the use of words in a non-literal manner to create abstract meanings that go beyond surface-level interpretation (Lakoff and Johnson, 2008). This may facilitate and intensify indirect communication and rhetorical effects, serving persuasive and humorous discourse, or audience engagement (Fussell and Moss, 2014; Burgers et al., 2016).

In NLP, figurative language has traditionally been studied through a text-centric lens (Chakrabarty et al., 2022a; Stowe et al., 2022; Lai et al., 2023). With the advance of MLLMs, research has shifted toward systematic benchmarking of their ability to process these non-literal expressions. This includes broad investigations into the models' general reasoning capabilities and linguistic nuances (Liu et al., 2022b; Jang et al., 2023; Yerukola et al., 2024; Jang and Frassinelli, 2024), as well as their performance in specialized interactive contexts such as dialogue systems (Jhamtani et al., 2021).

Despite the emphasis on text, figurative meaning is inherently a multimodal phenomenon (Chakrabarty et al., 2022b; Akula et al., 2023; Kulkarni et al., 2024). With the increasing availability of multimodal generative models, researchers started investigating non-literal meaning across combined modalities, with work on humor understanding in comics (Hessel et al., 2023), on generating visualizations of textual metaphors (Zhang et al., 2024), and on captioning visual figurative meanings (Saakyan et al., 2025).

Multimodal Meme Understanding Internet memes are a distinctive form of multimodal communication, where visual components shape the interpretation of text through context, priming, or template-based expectations (Shifman, 2013; Nissenbaum and Shifman, 2017; Wiggins, 2019). The NLP and vision-language communities have shown a growing interest in memes, with research highlighting how visual templates provide a contextual framework that shapes the interpretation and structure of the associated text (Zhou et al., 2024; Bates et al., 2025). Various tasks have been explored, including sentiment analysis (Hossain et al., 2022), hateful or harmful meme detection (Kiela et al., 2020; Cao et al., 2022; Liu et al., 2025), emotion classification (Sharma et al., 2020), caption generation (Hwang and Shwartz, 2023), style generalization (Nandy et al., 2024), and VQA tasks (Nguyen et al., 2025).

Although memes often contain a wide range of figurative meaning, few studies examine how well models handle them. Liu et al. (2022a) laid groundwork for multimodal pre-LLM figurative meaning analysis and introduced FigMemes, a collection of politically-opinionated memes.

To the best of our knowledge, our study is the first to systematically evaluate multimodal models in zero-shot settings by explicitly separating visual and textual inputs, enabling us to precisely assess the individual contribution of each modality. Moreover, we move beyond label accuracy to evaluate the quality of model-generated explanations, providing a more comprehensive benchmark for the interpretation of figurative meaning.

	FigMemes								Memotion 2					MET-Meme		
Type	None	Allus.	Exag.	Irony	Anthrop.	Met	Contr.	*Total*	NS	LS	VS	ES	*Total*	Lit	Met	*Total*
Count	495	265	265	320	131	286	171	1372	804	388	246	62	1500	1054	1054	2108

Table 1: Statistics of our evaluation data. Underlined types indicate negative cases (e.g. non figurative) in the binary classification task of each dataset. *Total* shows the number of evaluated memes in each dataset; in FigMemes, a meme can span multiple types of figurative meaning. FigMemes: Allus. = Allusion, Exag. = Exaggeration, Anthrop. = Anthropomorphism, Met. = Metaphor, Contr. = Contrast. Memotion 2: NS, LS, VS, ES indicate increasing levels of sarcasm (Not, Little, Very, Extremely). MET-Meme: Lit = Literal, Met = Metaphoric.

3. Experimental Setup

3.1. Datasets

We analyze three publicly available meme datasets concerning figurative expressions: FigMemes (Liu et al., 2022a), Memotion 2 (Ramamoorthy et al., 2022), and MET-Meme (Xu et al., 2022). To ensure modality comparability, only memes with embedded text are evaluated.

- **FigMemes** is multi-label and the first and only dataset which annotates six types of figurative language in memes: `allusion`, `exaggeration/hyperbole`, `irony/sarcasm`, `anthropomorphism/zoomorphism`, `metaphor/simile` and `contrast`. They collect memes from 4chan /pol/ (the politically incorrect) board,[2] a platform known for its high prevalence of hateful and offensive material.
- **Memotion 2** annotates memes from Google Images with multiple sentiment and emotion labels, including funny, sarcasm, offensive, and motivational. Here we focus on `sarcasm` only as it is not only associated with the emotion of contempt, but also characterized by indirectness and irony, which are central to figurative expressions.
- **MET-Meme** includes memes from Twitter, Weibo, Google, and Baidu images. It focuses on metaphor occurrences in memes, but also contains manual annotations of sentiment categories, intentions, and offensiveness degrees. While the dataset contains both Chinese and English data, in this work, we focus only on the English memes, and specifically on the binary labels indicating whether a meme is `metaphorical` or not.

Label Distribution For evaluation, we use the FigMemes test set, covering six figurative types in political memes, and the Memotion 2 validation set for sarcasm analysis.[3] For MET-Meme, we take all metaphoric items and randomly sample an equal number of literal items. This selection provides a diverse range of non-literal meanings across different meme domains. The statistics of our evaluation data are shown in Table 1.

All three datasets were originally annotated by multiple annotators. We rely on these labels which are summarized below. FigMemes takes the majority vote of three annotators, who are also the authors of the paper; Memotion 2 takes the majority vote of three workers on Amazon Mechanical Turk; and each meme in MET-Meme is annotated by at least three graduate students and two research assistants with relevant professional knowledge.[4]

Preprocessing In previous work (Liu et al., 2022a; Hwang and Shwartz, 2023), the whole meme – with its embedded text – is usually treated as visual input. Here, to explore the contributions of different modalities to model processing, we create an *image*-only input condition by removing textual information embedded in the original meme, as illustrated in Figure 1. We apply PaddleOCR (Cui et al., 2025) to detect text and mask it, and then use LaMa (Suvorov et al., 2021) for inpainting. The quality of the inpainted images is manually verified and found to be high (further details about this in Appendix A).

3.2. Task Setup

We explore the extent to which i) MLLMs uncover non-literal meanings in memes and ii) provide faithful explanations of predictions.

Task and Prompt Unlike prior work (Liu et al., 2022a; Pramanick et al., 2022; Xu et al., 2022), to the best of our knowledge, we are the first to setup *detection and explanation* tasks: models must not only determine whether an input contains one or more types of figurative meanings, but also provide a corresponding explanation. For detection, we follow two different strategies according to the information available in the different resources. For

[2] `http://boards.4chan.org/pol/`

[3] In Memotion 2 (Sharma et al., 2020; Pramanick et al., 2022), *little*, *very*, and *extremely sarcastic* are considered positive cases; *not sarcastic* is considered negative. The test set is not publicly available.

[4] FigMemes labels achieved a Fleiss' κ of 0.42, whereas equivalent agreement metrics for metaphor and sarcasm are not reported for the other datasets.

binary classification, we detect `sarcasm` in MEMOTION 2, `metaphor` in MET-MEME, and each type of figurative meaning independently in FIGMEMES. Additionally, since each meme in FIGMEMES can be associated with multiple types of figurative meaning, we define a *multi-label* classification task exclusive to FIGMEMES.

We focus on a zero-shot evaluation and design two prompting strategies to validate the results. In the first strategy, we simulate the annotation style of the source datasets: we provide definitions of each figurative type, following FIGMEMES (Liu et al., 2022a), and prompt the model to return the binary label for specific type of figurative meaning, and its explanation to justify the label:

```
System:
You are an expert linguistic annotator.
User:
You are asked to annotate whether the input
   contains a {FIGURATIVE TYPE} expression
   or not.
Definition:
   {FIGURATIVE TYPE}: {DEFINITION}
{INPUT}
Please explain why you are assigning this
   label. Your explanation should clearly
   justify your choice and reference the
   relevant visual and/or textual compon-
   ents in the input.
Return your answer ONLY as a JSON object
   in this exact format:
{
  "label":
    {FIGURATIVE TYPE}: "Yes" or "No",
  "explanation": {generated_explanation}
}
ONLY return a valid JSON object in the
   exact format above.
```

For multi-label classification, we incorporate the definitions of all six figurative types into a unified prompt, enabling the model to perform simultaneous inference for every category in a single query.

In the second strategy, we prompt the model to provide a continuous, probability-like value between 0 and 1 to indicate the degree to which the input expresses a specific type of figurative meaning. Except for replacing the "`label`" field with a continuous "`score`", the prompt template is identical to the first setting.

Input Modality To investigate how information from each modality influences model behavior, we evaluate three input conditions: the original *meme*, as well as two ablation settings: *text*-only, which inputs only the embedded text from the meme, and *image*-only, which removes the embedded text while retaining the visual content (see Section 3.1).

Labels per Meme	0	1	2	3	4
Meme Count	495	715	277	51	4

Table 2: Count of labels per meme in FIGMEMES. On average, each meme has 0.93 labels. 0 means no figurativeness is present.

To minimize potential modality bias arising from prompt phrasing, we deliberately exclude the term "meme" from the prompt template.

3.3. Evaluation Setup

Models are required to generate both labels and corresponding explanations. We evaluate model performance along three aspects:

1. model label(s) vs. gold label(s);
2. model label(s) vs. model explanation;
3. model explanation vs. original meme content.

The first aspect is assessed automatically, while the latter two are evaluated via human assessment.

Automatic Evaluation For binary classification, we report F1 for each figurative type in the text.

For multi-label classification on FIGMEMES, we adopt micro-averaging because the gold annotations in FIGMEMES are highly sparse, with most memes containing only a single type of figurative meaning. The count of labels across memes is reported in Table 2. In this setting, micro metrics provide a more stable and reliable estimate of overall performance. We report micro-accuracy, micro-precision, micro-recall and micro-F1. Macro scores are provided in Appendix B.1.

All results are averaged over 5 runs; standard deviations are reported in Appendix B.2.

Human Evaluation To conduct the human evaluation, we randomly sample 90 memes, 30 from each dataset; for FIGMEMES, we focus on memes containing a single type of figurative meaning. We evaluate explanations along two aspects with seven criteria. First, we focus on how well the explanation generated by the model supports its generated label (i.e. evaluating model explanation vs. label):

1. **Relevance:** The explanation highlights the key features of the label and shows why it applies.
2. **Consistency:** The explanation aligns with the label and contains no internal contradictions.

Second, we evaluate how well the explanation generated by the model reflects the content of the original meme. These criteria are taken from Hwang and Shwartz (2023):

3. **Correctness:** The explanation accurately conveys the meaning intended by the person who posted the meme.
4. **Appropriate Length:** The explanation length is appropriate for conveying the meaning (i.e. it is not too verbose).
5. **Visual Completeness:** The explanation describes all the important components in the image.
6. **Textual Completeness:** The explanation describes all the important components in the text embedded in the meme.
7. **Faithfulness:** All components of the explanation are supported by either the visual or text content (i.e. there are no made-up components).

The first author of this paper and a bachelor student majoring in computational linguistics hired as student assistant rated the model explanations on a 5-point Likert scale, ranging from *strongly disagree* (score 0) to *strongly agree* (score 4). After annotating the first 15 memes, the two annotators discussed their evaluations to align their understanding and ensure consistency in subsequent ratings. The Cohen's κ between the two annotators is 0.79 across all evaluated items, indicating substantial agreement. We report the averaged score of two annotators in Section 4.

3.4. Models

We evaluate eight models from three families of state-of-the-art MLLMs of different sizes: Aya-Vision (`Aya`, Üstün et al. 2024; 8B and 32B), Gemma 3 (`Gemma`, Team et al. 2025; 4B, 12B, and 27B), and Qwen2.5-VL (`Qwen`, Bai et al. 2025; 7B, 32B, and 72B). All models are evaluated in zero-shot settings with a unified prompt template (as shown in Section 3.2) across all model sizes. Implementation details are provided in Appendix B.3.

4. Results and Discussion

We begin by focusing on the first prompt setup to discuss about the results of model performance on detecting figurative meaning in Section 4.1 and on explaining it in Section 4.2. The influence of prompt settings will be discussed in Section 4.3.

4.1. Automatic Evaluation of Detection

Q1: How does model performance vary in detecting different types of figurative meaning? Here, we analyze model performance under the original meme input setup. We start with the *binary* classification tasks (Table 3), and subsequently, move to *multi-label* classification (Table 4), which only applies to FigMemes.

As reported in Table 3, the eight MLLMs show considerable variation: Overall, according to their *average* score on eight types, **the largest model from each family consistently achieves the best results**, yet the differences in overall performance remain marginal. `Qwen-72B` performs the best among the eight models.

However, on individual label types, the performance gap between models can be more substantial.[5] For instance, on `metaphor`, the gap between `Qwen-7B` and `Qwen-32B` is as large as 55.34%. Models' performance on Memotion 2 and MET-Meme is more consistent across model sizes, with only minor variations among models from all three series.

On FigMemes, models generally perform best on `contrast`, with `Aya-8B` achieving the highest score of 77.95%. In comparison, performance on `anthropomorphism` is much lower, with the best model, `Qwen-32B`, reaching only 56.99%. These trends are consistent with the benchmarks reported by Liu et al. (2022a), in which discriminative models also score highest on `contrast` and lowest on `anthropomorphism`.

For multi-label classification, we only report the results of the largest model from each family due to space constraints. Table 4 shows that all three models surpass the baseline performance, which corresponds to random guessing of each label as positive or negative with a 50/50 probability. However, performance differs notably across the four metrics. This is largely due to the fact that, on average, each meme has only one positive label while the other five are negative. For instance, with input of original memes, `Gemma-27B` achieves the highest recall but the lowest precision among the three models, as the model tend to predict the presence of figurative meaning. This tendency is generally observed across all three models.

Q2: How much does text, image, or both contribute to the figurativeness of memes? We begin by examining the *binary classification* task. On FigMemes, we treat the best scores from Liu et al. (2022a) across modalities and models as a baseline. As shown in Table 3, based on the average score across types, performance generally drops when either modality is removed, with the exception of `Gemma-4B` and `Gemma-12B`, which score higher on `exaggeration`, `anthropomorphism` and

[5] To address potential concerns regarding the variance in model ranking, we conducted a series of McNemar's tests. Comparing the top two models for each task, we observed that the best-performing model significantly outperformed the second best one ($p < 0.001$) across every experimental seed.

Model	Size	Input	FigMemes Allusion	FigMemes Exag.	FigMemes Irony	FigMemes Anthrop.	FigMemes Metaphor	FigMemes Contrast	Memotion 2 Sarcasm	MET-Meme Metaphor	*Average*
Baseline			52.32	44.00	49.77	41.76	44.87	56.91	–	–	–
Aya	8B	Meme	70.04	58.63	56.06	51.94	50.62	**77.95**	63.48	66.66	61.92
		-Text	65.44	58.52	10.53	48.39	35.91	59.99	49.64	54.16	47.82
		-Image	34.78	30.16	57.07	8.47	40.09	23.46	63.01	57.28	39.29
	32B	Meme	66.87	61.01	61.31	54.67	54.97	70.85	63.37	66.77	62.48
		-Text	67.48	59.45	42.27	51.70	39.48	60.13	47.22	56.03	52.97
		-Image	48.99	37.66	59.09	9.30	37.85	63.53	59.82	50.11	45.79
Gemma	4B	Meme	**70.17**	53.94	62.70	51.37	53.22	56.71	63.47	66.91	59.81
		-Text	58.13	58.36	31.19	52.06	54.05	61.84	62.05	60.02	54.71
		-Image	33.29	43.46	58.33	3.86	47.19	61.20	63.08	61.42	46.48
	12B	Meme	60.27	56.09	63.01	53.00	42.87	57.48	**63.54**	67.81	58.01
		-Text	64.06	56.96	45.54	53.94	31.87	59.48	58.45	51.84	52.77
		-Image	53.45	49.79	57.97	11.34	35.95	66.58	61.78	54.80	48.96
	27B	Meme	61.32	62.32	62.51	52.99	63.64	62.82	63.38	68.21	62.15
		-Text	58.66	60.87	42.68	53.81	52.77	61.65	60.46	53.99	55.61
		-Image	54.66	41.53	60.68	13.78	47.18	66.80	62.77	57.65	50.63
Qwen	7B	Meme	59.92	52.96	52.58	48.52	9.10	75.62	63.22	**70.90**	54.10
		-Text	56.77	53.26	9.94	48.49	3.27	71.36	36.60	36.22	39.49
		-Image	23.48	30.33	26.20	1.77	6.39	45.95	56.85	39.29	28.78
	32B	Meme	62.48	60.06	**69.81**	**56.99**	**64.34**	68.67	62.58	69.48	64.30
		-Text	64.78	**63.66**	13.56	53.51	31.80	68.68	27.00	37.65	45.08
		-Image	56.69	38.21	53.36	17.78	32.63	65.56	56.85	57.85	47.37
	72B	Meme	65.31	61.61	65.35	54.09	61.76	76.76	62.81	70.57	**64.78**
		-Text	66.18	60.93	22.95	55.24	41.63	74.90	42.90	50.53	51.91
		-Image	53.80	37.96	61.81	16.92	30.53	68.46	60.30	55.62	48.18

Table 3: F1 scores (in %) of eight models on the *binary* classification of figurative meaning types in memes from three datasets. *-Text* and *-Image* denote the ablation of the text or image modality, respectively. The **best** and second-best results are shown in bold and underlined, respectively. Following Liu et al. (2022a), in FigMemes we consider memes with 0 labels as negative cases, and take the best-performing benchmark score as baseline. Higher scores are highlighted in orange, lower scores in blue. Results are averaged over five runs.

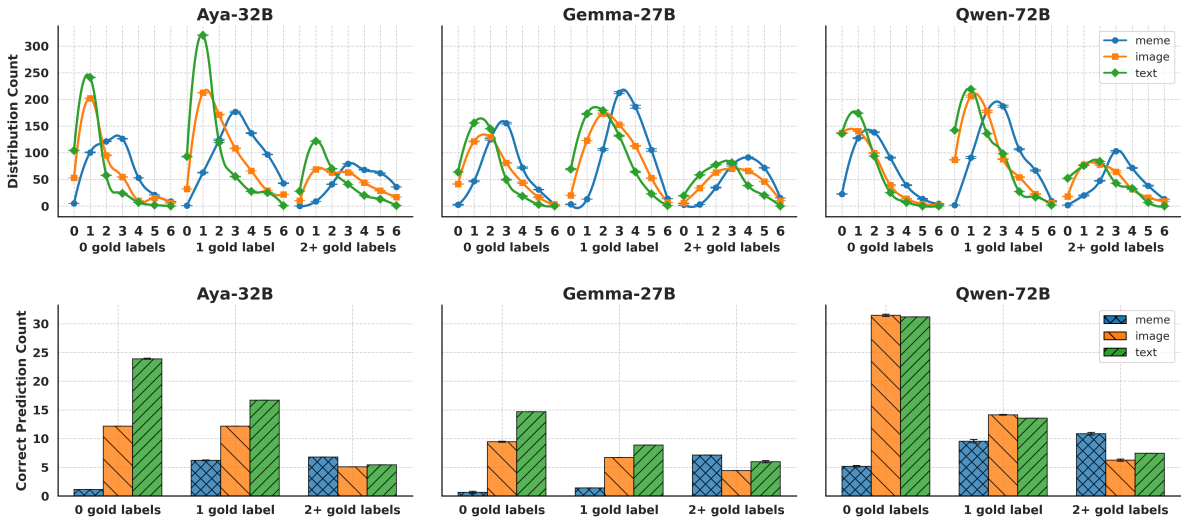

Figure 2: Performance of `Aya-32B`, `Gemma-27B`, and `Qwen-72B` on the multi-label classification task (only in FigMemes) across modalities and meme groups. Error bars indicate standard deviation over 5 runs. **Top row:** Count of memes with 0, 1, or 2+ figurative types (0, 1, 2+ gold labels) that are predicted to contain from 0 to 6 figurative types. **Bottom row:** Count of memes containing 0, 1, or 2+ figurative types that are fully correctly predicted, where all six predicted labels match the gold labels.

`contrast` in FigMemes when one modality is removed.

Interestingly, for all evaluated models, the negative impact of removing the image (*-image*) is more pronounced than removing the text (*-text*) for `irony` in FigMemes, `sarcasm` in Memotion 2, and `metaphor` in MET-Meme; whereas for `allusion`, `exaggeration`, and `anthropomorphism` in FigMemes the opposite pattern occurs. This indicates that **some figurative types are visually grounded while others are primarily conveyed through text**, and MLLMs rely on different modalities to in-

Model	Input	Acc.	Prec.	Rec.	F1
random	–	50.04	15.57	49.88	23.73
`Aya-32B`	Meme	57.60	24.63	83.54	38.05
	-Text	66.80	25.30	57.86	35.20
	-Image	74.06	28.36	43.51	34.34
`Gemma-27B`	Meme	55.97	24.44	**87.50**	38.21
	-Text	62.08	23.79	65.04	34.84
	-Image	68.56	25.86	54.83	35.15
`Qwen-72B`	Meme	64.69	28.24	82.12	**42.03**
	-Text	71.91	28.76	54.31	37.60
	-Image	**74.27**	**29.07**	45.21	35.39

Table 4: Accuracy, precision, recall and Micro-F1 (in %) of models on *multi-label* classification of figurative meanings in FIGMEMES. The highest score for each metric is highlighted in **bold**. Higher scores are highlighted in orange, lower scores in blue.

terpret different types of figurative meaning. Moreover, for `metaphor` in FIGMEMES and MET-MEME, all models (except for `Gemma-4B`) exhibit substantial performance drops under both *-Image* and *-Text* conditions, highlighting that detecting metaphors depends not on a single modality alone but on the interplay between text and image.

For the *multi-label classification* task (Table 4), micro-recall across all models consistently follows the trend: original meme > *-text* > *-image*, indicating that **models tend to predict the presence of figurative meaning more often under the full meme input setup**.

To analyze whether models predict more labels due to richer figurative content in a meme (without any ablation) or merely because of a bias toward the meme format, we categorize memes from FIGMEMES by the number of gold labels and count the predicted labels under different modality inputs. As shown at the top row of Figure 2, independent of the number of gold figurative labels a meme carries, the models tend to assign more figurative labels when they see the original meme compared to the ablated versions. Notably, among 495 memes without any figurative labels, all three models assign 0 labels to very few cases, while most cases receive 1-3 labels. Due to this bias, all three models fail to correctly classify memes that do not contain any figurative labels, as reported in the bottom row of Figure 2. For such cases, models even perform better with text-only or image-only input, especially for `Qwen-72B`.

In contrast, for memes containing more than two types of figurative meaning, using the full meme input yields better performance. This tendency is also observed in other smaller models. We present the performance of the eight models in Appendix B.4.

4.2. Human Evaluation of Explanation

Q3: How effective are the model-provided explanations for its predicted labels and memes? We manually evaluate the explanations generated by the eight models following criteria introduced in Section 3.3. As shown in Figure 3, our comparison reveals an interesting discrepancy: while `Qwen-72B` was previously found best on automatic metrics, it falls short in human assessment. While `Gemma` and `Qwen` models are judged with higher ratings than `Aya` models of comparable size, `Qwen-72B` surpasses `Qwen-32B` only on `faithfulness`, indicating fewer hallucinations, while on other criteria, its performance is roughly equal to or slightly worse than `Qwen-32B`. Furthermore, for `Aya` and `Gemma`, larger models outperform smaller ones.

Focusing on the internal coherence between predicted labels and explanations, we find that **all models perform better on `consistency` than `relevance`**. Model-generated explanations often lack substance: despite aligning with the predicted labels, they frequently fail to provide concrete evidence for the specific figurative meaning, such as identifying the target of sarcasm or the underlying metaphor.

Human annotators rated the models highest on `textual completeness`, indicating a strong consensus that the generated explanations successfully capture the key textual elements essential to the memes' figurative meanings.

By comparison, the models perform worse in capturing visual components. Most models perform worst on `correctness`: `Aya-8B`, `Aya-32B`, and `Gemma-4B` all score below 2, suggesting that human annotators generally disagree with the models' interpretations of the memes intent.

Q4: What are the common types of errors in model explanations? As we found in Section 4.1, MLLMs' strong bias toward predicting the presence of figurative meaning in meme formats results in two key limitations in their explanations. One is **label-explanation inconsistency**: for example, Figure 4a is a motivating meme without sarcastic meaning. `Qwen-72B` labels it as *sarcastic*, but in its explanation, it contradicts the label and asserts that there is no sarcasm. The other is **over-interpretation**, where the model attempts to identify an uncommon or exaggerated aspect to justify a positive label. For instance, `Gemma-12B` labels Figure 4a as *sarcastic* and explains that it "utilizes the `Success Kid` meme paired with the phrase `Planning Like A Boss`, creating a strong sense of sarcasm because the baby's determined expression contrasts humorously with the often chaotic reality of planning."

Mismatches between model-generated explanations and the memes are largely due to failure in

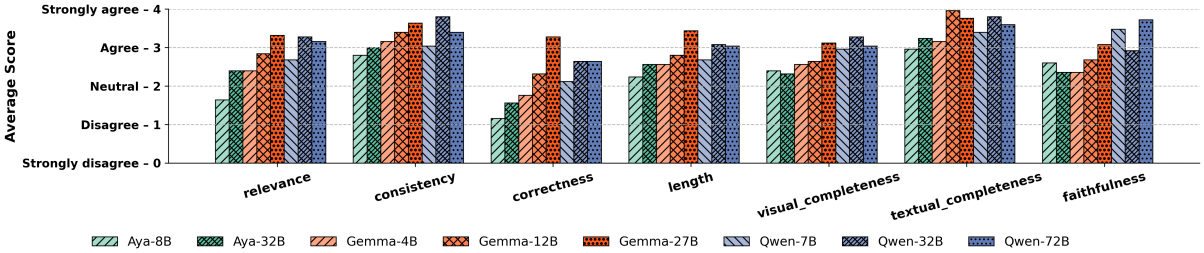

Figure 3: Human evaluation results on model-generated explanations.

(a) (b) (c) (d)

Figure 4: Examples of memes for which MLLMs struggle to generate high-quality explanations.

visual information extraction rather than in textual processing. These issues typically arise from either an **incorrect depiction** of the image or a **neglect of visual cue**. For the former, `Aya-32B` misinterprets Figure 4b as "[...] The finger over the cat's mouth is a visual metaphor for silence, implying that the cat should keep quiet or be hushed, thereby attributing human-like behavior to the animal." This is inaccurate, as people usually use a vertical finger in front of the mouth to signal silence, whereas the image shows a finger lifting the cat's nose, causing its mouth to open and make an "A". For the latter, although all models predict the label *metaphoric* correctly, but fail to interpret Figure 4c, because they overlook the crucial visual cue, the water spilling away before it reaches the glass, and thus misread the meme as "[...] symbolizing how information is absorbed and processed", rather than the intended meaning of failing to do so.

We also find that models struggle with memes depicting everyday human behavior. Figure 4d satirizes people who are willing to spend $1,000 on a phone but feel annoyed when an app is not free, even if it costs only 99 cents. This demonstrates the common psychological pattern where consumers rationalize big purchases but resist minimal charges, yet half of the models fail to interpret it. This type of error can be described as a **behavioral reasoning error**, where the model misses the social or psychological irony underlying the meme due to lack of social experience, which is also pointed out by Chen et al. (2024) and Mathur et al. (2024).

(a) (b)

Figure 5: Two memes from MEMOTION 2 with gold label "not sarcastic".

Q5: Can explanations be valid for "wrong" labels? When qualitatively evaluating model-generated labels and explanations, we observed that models can assign labels that are different from the gold label while still providing reasonable and evidence-based explanations. For example, both memes in Figure 5 are annotated with *not sarcastic*, yet `Gemma-27B` classifies the first as *sarcastic*, describing it as "creating a sarcastic tone about the unchanging nature of this specific behavior and subtly mocking the continued enthusiasm." and `Qwen-72B` labels the second as *sarcastic*, explaining it as "mocking the concept of age-related restrictions on simple pleasures". Both annotators (strongly) agree that these explanations fit well with the seven human evaluation criteria. These cases highlight the inherent subjectivity in interpretation of figurative meaning and indicate that minority-vote labels are not necessarily errors.

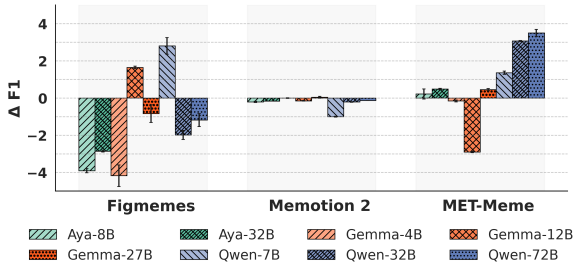

Figure 6: Difference in macro-F1 (in %) between prompting score and prompting label setting for eight models, averaged over five seeds. Bars with positive values indicate that prompting score improved performance; bars with negative values indicate a decrease. Results cover binary classification tasks across three datasets.

4.3. Prompt Sensitivity

Q6: How robust are models to different prompts? Since MLLMs are highly sensitive to prompt design, we validate our results with two prompting strategies (see Section 3.2). The full prompt template and the corresponding experimental results are provided in Appendix B.5.

Although F1 scores under the same prompt setting show little variation (less than 1% standard deviation), model performance under different prompt settings varies moderately. Figure 6 quantifies the gap between results of the binary classification task under two prompting setups. Overall, the gap of performance of all models is within $\pm 4\%$. Models demonstrate higher robustness on MEMOTION 2 than on FIGMEMES and MET-MEME. `Gemma-27B` is the most consistent model, with performance variations across the three datasets less than 1%. There is no clear evidence that any particular prompt setting consistently benefits a specific model in all setups; performance still largely depends on the dataset.

5. Conclusion

We explore the ability of current generative multimodal models to *detect and explain* figurative meaning in internet memes. We evaluate in total eight models from three families, `Aya`, `Gemma`, and `Qwen` with different sizes, and analyze the influence of input modalities (text, image, both) quantitatively and qualitatively.

Larger models generally achieve better results on the binary classification task. Textual components contribute more to detecting `irony` and `metaphor`, whereas visual components are more critical for `allusion`, `exaggeration`, and `anthropomorphism`. All evaluated models exhibit a bias toward predicting the presence of figurative meaning in meme format input, which is typically composed of images with embedded text, suggesting a higher reliance on surface characteristics.

Importantly, our human evaluation reveals several deficiencies of model-generated *explanations*. Across all models, identifying the core visual components that convey figurative meaning is harder than identifying the textual ones.

When model-generated explanations fail to capture the figurative meaning of a meme, it is largely due to failures in visual information extraction rather than in textual processing. Furthermore, models sometimes assign labels that differ from the gold annotation while still generating valid, evidence-based explanations, suggesting that minority-vote labels are not necessarily errors but may reflect the inherent subjectivity of figurative meaning interpretation.

Ultimately, our findings highlight the need for future research to move beyond simple label prediction toward developing more sophisticated multimodal architectures that can ensure reasoning faithfulness and a deeper sensitivity to the complex socio-psychological contexts inherent in meme communication.

Limitations

Data and Benchmark Limitations The three-year release history of the evaluation datasets poses a risk of data contamination, potentially inflating performance if models encountered these samples during pre-training. Furthermore, the limited variety of meme formats, imbalanced label distribution (e.g., sparse figurative types in FIGMEMES), and inconsistent annotation schemes across datasets collectively constrain the representativeness of our findings. These factors complicate fair comparisons and may bias the assessment of models' generalization capabilities across diverse communicative contexts.

Robustness to Prompt Specification Despite evaluating two prompt templates across five experimental runs, potential sensitivity to prompt specification remains. For example, it is unclear how consistency might be affected by variations in output format (e.g., XML vs. JSON) or subtle shifts in instructional wording. Furthermore, while our continuous scoring-type prompt (0–1) provides fine-grained scalar feedback, a discrete Likert-like scale might introduce different inductive biases.

Ethical Considerations

Hateful content in FigMemes To assess potential ethical concerns, we randomly sampled 30 memes from FigMemes, finding that 17 could be considered potentially offensive, targeting groups such as religious or social communities. While the dataset contains sensitive content, all annotations in our study were conducted with care, and no annotator reported any psychological distress or harm during the annotation process. We emphasize that our study focuses on understanding figurative language in these memes for research purposes and does not promote or disseminate harmful content. Access to the dataset is controlled, and all participants are expected to follow ethical research practices when using it.

Use of AI Assistants. The authors acknowledge the use of ChatGPT solely for correcting grammatical errors, enhancing the coherence of the final manuscripts, and providing assistance with coding.

Acknowledgements

We thank MaiNLP lab members Felicia Körner, Philipp Mondorf and Andreas Säuberli for giving feedback on earlier drafts of this paper, as well as to the anonymous reviewers for their feedback. This work is supported by the KLIMA-MEMES project funded by the Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities. The authors are responsible for the content of this publication.

Bibliographical References

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. MetaCLUE: Towards Comprehensive Visual Metaphors Research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report.

Luke Bates, Peter Ebert Christensen, Preslav Nakov, and Iryna Gurevych. 2025. A template is all you meme. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10443–10475, Albuquerque, New Mexico. Association for Computational Linguistics.

Christian Burgers, Elly A. Konijn, and Gerard J. Steen. 2016. Figurative Framing: Shaping Public Discourse Through Metaphor, Hyperbole, and Irony. *Communication Theory*, 26(4):410–430.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

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

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:127428.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR 3.0 Technical Report.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks

from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, and Vinutha B. NarayanaMurthy. 2024. CM-Off-Meme: Code-Mixed Hindi-English Offensive Meme Detection with Multi-Task Learning by Leveraging Contextual Knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3380–3393, Torino, Italia. ELRA and ICCL.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022a. FigMemes: A Dataset for Figurative Language Identification in Politically-Opinionated Memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria. Association for Computational Linguistics.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing Social Intelligence in AI Agents: Technical Challenges and Open Questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, Miami, Florida, USA. Association for Computational Linguistics.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. ***YesBut***: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895, Miami, Florida, USA. Association for Computational Linguistics.

Khoi P. N. Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, Affan Kashfi, Aaron Xu, Suyesh Shrestha, Megan Vu, Jerry Wang, and Vincent Ng. 2025. MemeQA: Holistic Evaluation for Meme Understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18926–18946, Vienna, Austria. Association for Computational Linguistics.

Asaf Nissenbaum and Limor Shifman. 2017. Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society*, 19(4):483–501.

Shraman Pramanick, Aniket Roy, and Vishal M. Patel Johns. 2022. Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 546–556.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *De-Factify @AAAI*, volume 3199 of *CEUR Workshop Proceedings*.

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

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Limor Shifman. 2013. Memes in a Digital World: Reconciling with a Conceptual Troublemaker. *Journal of Computer-Mediated Communication*, 18(3):362–377.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. 2025. Gemma 3 technical report.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Bradley E Wiggins. 2019. *The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality*. Routledge.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. MET-Meme: A Multimodal Meme Dataset Rich in Metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '22, pages 2887–2899, New York, NY, USA. Association for Computing Machinery.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring Chain-of-Thought for Multi-modal Metaphor Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. 2024. GOME: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18500–18510, Miami, Florida, USA. Association for Computational Linguistics.

Naitian Zhou, David Jurgens, and David Bamman. 2024. Social meme-ing: Measuring linguistic variation in memes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3005–3024, Mexico City, Mexico. Association for Computational Linguistics.

Shijia Zhou, Siyao Peng, Simon M. Luebke, Jörg Haßler, Mario Haim, Saif M. Mohammad, and Barbara Plank. 2025. What Media Frames Reveal About Stance: A Dataset and Study about Memes in Climate Change Discourse.

Appendices

A. Image Preprocessing

To ensure the complete removal of textual information from the input—rather than merely relying on prompt-based attention control—we preprocessed the original images. Specifically, we employed PaddleOCR (Cui et al., 2025) to detect text and mask it, and then use LaMa (Suvorov et al., 2021) for inpainting. In most cases, the text was successfully eliminated, as illustrated in Figure 7. For a small subset of low-resolution images where automated processing was suboptimal, we manually refined the results to ensure data quality as shown in Figure 8.

Figure 7: Examples of automatic preprocessing by inpainting text. The left column shows the original meme images, while the right column displays the results after text removal using PaddleOCR and LaMa.

B. Implements Detail and Additional Results

B.1. Macro-Averaged Results for Multi-label Classification

For completeness, we also report macro-averaged accuracy, precision, recall and F1 for multi-label classification on FigMemes in Table 5. Unlike micro-averaging, macro-averaging computes scores independently for each label and takes their un-

Figure 8: Examples of automatic and manual text removal. For cases where low image resolution hindered automated OCR detection, the original memes were manually edited to ensure a clean, text-free input for the model.

weighted mean, which may be less stable in the presence of label sparsity. These scores are provided as supplementary reference alongside the micro-averaged results reported in the main paper.

Model	Input	Acc.	Prec.	Rec.	F1
Aya-32B	Meme	57.60	24.41	82.72	37.27
	-Text	66.80	25.73	61.90	34.71
	-Image	74.06	26.90	39.98	30.23
Gemma-27B	Meme	87.50	37.88	55.97	24.57
	-Text	69.09	34.82	62.08	24.12
	-Image	51.24	31.44	68.56	23.70
Qwen-72B	Meme	81.19	41.73	64.69	28.69
	-Text	60.06	35.90	71.91	28.71
	-Image	41.90	31.92	74.27	28.57

Table 5: Macro-averaged accuracy, precision, recall and F1 (in %) of models on *multi-label* classification of figurative meanings in FIGMEMES.

B.2. Main Results with Standard Deviation

To account for potential performance variations, we conducted five independent trials for the experiments. The resulting average scores, along with the standard deviations which indicate the robustness of our approach, are reported in Table 7.

B.3. Model Setup

We evaluate the following vision-language models: Aya-Vision [6] (Üstün et al., 2024), Gemma 3[7] (Team et al., 2025), and Qwen2.5-VL[8] (Bai et al., 2025).

For `Gemma` and `Qwen`, we test their instruction-tuned versions.

All models ran locally on NVIDIA A100 and H200 GPUs with the vLLM[9] framework (Kwon et al., 2023) for efficient and consistent inference. Table 6 summarizes the decoding hyperparameters used during inference across all models. Each experiment is repeated 5 times using fixed random seeds: 42, 52, 62, 72, and 82.

Parameter	Value
Temperature	0.7
Top-*p*	0.1
Repetition Penalty	1.05
Max Tokens	512

Table 6: Sampling parameters used during inference.

B.4. Performance by Label Count

Due to space constraints, Figure 2 in the main text only presents the performance of the largest model from each series. Here, we provide the complete results for all tested models, as shown in Figure 9 and Figure 10.

B.5. Second Prompt Setting Results

In the second prompt setting, we prompt the model to provide a continuous, probability-like value between 0 and 1 to indicate the degree to which the input expresses a specific type of figurative meaning.

[6] https://huggingface.co/collections/CohereLabs/cohere-labs-aya-vision-67c4ccd395ca064308ee1484

[7] https://huggingface.co/collections/google/gemma-3-release-67c6c6f89c4f76621268bb6d

[8] https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5

[9] https://github.com/vllm-project/vllm

```
System:
You are an expert linguistic annotator.

User:
You are asked to annotate whether the input
   contains a {FIGURATIVE TYPE} expression
   or not.

Definition:
   {FIGURATIVE TYPE}: {DEFINITION}

{INPUT}

Please explain why you are assigning this
   label. Your explanation should clearly
   justify your choice and reference the
   relevant visual and/or textual compon-
   ents in the input.

Return your answer ONLY as a JSON object
   in this exact format:

{
  "score":
    {FIGURATIVE TYPE}:
       probability_between_0_and_1,
  "explanation": {generated_explanation}
}

ONLY return a valid JSON object in the
   exact format above.
```

The averaged score over 5 runs with standard deviations of experiments with second prompt template setup in Table 8.

Model	Size	Input	FigMemes						Memotion 2	MET-Meme
			Allusion	Exag.	Irony	Anthrop.	Metaphor	Contrast	Sarcasm	Metaphor
Baseline			52.32	44.00	49.77	41.76	44.87	56.91	–	–
Aya	8B	Meme	70.04±0.00	58.63±0.00	56.06±0.06	51.94±0.00	50.62±0.15	77.95±0.00	63.48±0.00	66.66±0.26
		-Text	65.44±0.00	58.52±0.12	10.53±0.00	48.39±0.19	35.91±0.22	59.99±0.38	49.64±0.00	54.16±0.22
		-Image	34.78±0.00	30.16±0.00	57.07±0.00	8.47±0.00	40.09±0.00	23.46±0.00	63.01±0.01	57.28±0.13
	32B	Meme	66.87±0.00	61.01±0.00	61.31±0.03	54.67±0.00	54.97±0.00	70.85±0.07	63.37±0.00	66.77±0.04
		-Text	67.48±0.00	59.45±0.00	42.27±0.17	51.70±0.00	39.48±0.22	60.13±0.07	47.22±0.09	56.03±1.31
		-Image	48.99±0.00	37.66±0.20	59.09±0.04	9.30±0.00	37.85±0.05	63.53±0.00	59.82±0.03	50.11±1.62
Gemma	4B	Meme	70.17±0.49	53.94±0.14	62.70±0.06	51.37±0.36	53.22±0.30	56.71±0.11	63.47±0.01	66.91±0.05
		-Text	58.13±0.33	58.36±0.09	31.19±0.51	52.06±0.19	54.05±0.13	61.84±0.06	62.05±0.05	60.02±0.35
		-Image	33.29±0.06	43.46±0.10	58.33±0.07	3.86±0.73	47.19±0.09	61.20±0.13	63.08±0.00	61.42±0.20
	12B	Meme	60.27±0.05	56.09±0.05	63.01±0.04	53.00±0.00	42.87±0.19	57.48±0.04	63.54±0.02	67.81±0.01
		-Text	64.06±0.11	56.96±0.04	45.54±0.00	53.94±0.23	31.87±0.00	59.48±0.00	58.45±0.04	51.84±0.04
		-Image	53.45±0.19	49.79±0.20	57.97±0.15	11.34±0.05	35.95±0.56	66.58±0.28	61.78±0.01	54.80±0.08
	27B	Meme	61.32±0.06	62.32±0.05	62.51±0.09	52.99±0.24	63.64±0.16	62.82±0.15	63.38±0.01	68.21±0.01
		-Text	58.66±0.04	60.87±0.00	42.68±0.26	53.81±0.08	52.77±0.12	61.65±0.11	60.46±0.04	53.99±0.02
		-Image	54.66±0.06	41.53±0.09	60.68±0.22	13.78±0.06	47.18±0.16	66.80±0.27	62.77±0.02	57.65±0.07
Qwen	7B	Meme	59.92±0.08	52.96±0.18	52.58±0.22	48.52±1.27	9.10±0.41	75.62±0.37	63.22±0.00	70.90±0.05
		-Text	56.77±0.08	53.26±0.18	9.94±0.26	48.49±0.44	3.27±0.01	71.36±0.20	36.60±0.03	36.22±0.00
		-Image	23.48±0.00	30.33±0.00	26.20±0.24	1.77±0.00	6.39±0.01	45.95±0.36	56.85±0.00	39.29±0.00
	32B	Meme	62.48±0.15	60.06±0.36	69.81±0.14	56.99±0.16	64.34±0.25	68.67±0.07	62.58±0.02	69.48±0.04
		-Text	64.78±0.07	63.66±0.10	13.56±0.24	53.51±0.00	31.80±0.27	68.68±0.13	27.00±0.18	37.65±0.02
		-Image	56.69±0.00	38.21±0.00	53.36±0.00	17.78±0.00	32.63±0.00	65.56±0.00	56.85±0.09	57.85±0.05
	72B	Meme	65.31±0.17	61.61±0.06	65.35±0.04	54.09±0.17	61.76±0.19	76.76±0.38	62.81±0.00	70.57±0.18
		-Text	66.18±0.08	60.93±0.09	22.95±0.26	55.24±0.33	41.63±0.28	74.90±0.19	42.90±0.02	50.53±0.40
		-Image	53.80±0.00	37.96±0.00	61.81±0.00	16.92±0.00	30.53±0.00	68.46±0.00	60.30±0.01	55.62±0.05

Table 7: F1 scores (in %) of eight models on the *binary* classification of figurative meaning types in memes from three datasets. *-Text* and *-Image* denote the ablation of the text or image modality, respectively. Following Liu et al. (2022a), in FigMemes we consider memes with 0 labels as negative cases, and take the best-performing benchmark score as baseline. Results are averaged over five runs.

Model	Size	Input	FigMemes						Memotion 2	MET-Meme
			Allusion	Exag.	Irony	Anthrop.	Metaphor	Contrast	Sarcasm	Metaphor
Baseline			52.32	44.00	49.77	41.76	44.87	56.91	–	–
Aya	8B	Meme	60.97±0.05	58.72±0.11	59.99±0.03	49.13±0.07	53.19±0.13	59.80±0.15	63.27±0.03	66.88±0.01
		-Text	61.41±0.05	59.42±0.08	46.37±0.00	49.39±0.05	53.68±0.06	57.01±0.15	59.17±0.02	63.51±0.04
		-Image	51.66±0.15	30.95±0.08	58.17±0.06	6.43±0.04	52.34±0.13	45.69±0.00	56.03±0.04	53.62±0.05
	32B	Meme	69.11±0.00	57.43±0.04	60.55±0.00	54.56±0.10	57.41±0.05	53.47±0.04	63.21±0.00	67.25±0.00
		-Text	66.05±0.00	57.29±0.00	43.72±0.04	54.49±0.00	55.69±0.28	48.31±0.08	52.69±0.05	63.12±0.02
		-Image	49.40±0.00	41.75±0.05	61.93±0.04	12.12±0.00	46.79±0.07	51.58±0.16	59.24±0.04	49.47±0.11
Gemma	4B	Meme	68.25±0.57	52.29±0.36	60.95±0.47	45.66±0.94	52.14±0.16	43.75±0.43	63.47±0.02	66.75±0.02
		-Text	60.52±0.45	55.78±0.26	45.17±0.40	49.62±0.36	53.34±0.30	50.74±0.33	61.30±0.02	61.75±0.01
		-Image	35.48±0.00	46.85±0.12	58.46±0.04	6.52±0.03	48.48±0.04	53.52±0.00	63.37±0.00	60.31±0.00
	12B	Meme	61.29±0.00	58.72±0.04	61.86±0.00	52.51±0.00	52.70±0.05	55.54±0.04	63.39±0.00	64.91±0.03
		-Text	63.42±0.00	57.75±0.00	46.98±0.00	54.16±0.09	41.09±0.00	57.35±0.00	54.51±0.08	47.50±0.12
		-Image	53.62±0.00	45.45±0.11	57.93±0.05	6.78±0.00	41.62±0.34	69.55±0.28	61.07±0.00	54.41±0.11
	27B	Meme	59.13±0.52	60.75±0.39	61.80±0.17	52.04±0.44	62.73±0.44	64.16±0.58	63.42±0.05	68.67±0.05
		-Text	60.02±0.04	61.73±0.09	52.79±0.27	51.90±0.00	54.37±0.28	59.43±0.06	57.34±0.05	57.90±0.05
		-Image	54.82±0.03	41.91±0.00	59.95±0.17	14.69±0.05	49.44±0.14	64.35±0.14	60.38±0.03	60.17±0.03
Qwen	7B	Meme	62.60±0.00	51.23±0.17	58.77±0.15	51.18±0.00	30.53±0.26	61.20±0.22	62.21±0.03	72.26±0.06
		-Text	61.85±0.08	53.27±0.05	20.96±0.27	47.93±0.32	31.31±0.17	57.71±0.15	31.80±0.17	27.43±0.00
		-Image	24.63±0.00	29.65±0.00	40.82±0.00	6.72±0.00	30.05±0.00	53.73±0.06	44.17±0.00	32.62±0.00
	32B	Meme	61.76±0.04	59.10±0.07	69.79±0.13	56.07±0.09	63.50±0.32	60.21±0.29	62.37±0.02	72.54±0.06
		-Text	65.67±0.10	60.07±0.16	24.03±0.61	52.72±0.07	52.39±0.64	62.74±0.39	28.65±0.08	28.12±0.01
		-Image	54.60±0.04	42.71±0.00	54.97±0.08	18.92±0.00	46.50±0.10	61.36±0.08	56.31±0.07	49.68±0.06
	72B	Meme	67.61±0.24	57.83±0.15	63.95±0.22	53.98±0.28	63.99±0.51	70.42±0.53	62.68±0.00	74.06±0.01
		-Text	64.63±0.07	60.07±0.09	37.83±0.34	55.14±0.27	50.73±0.36	67.59±0.20	49.17±0.02	47.74±0.00
		-Image	53.61±0.00	38.81±0.00	60.76±0.05	13.85±0.00	40.52±0.11	69.16±0.00	58.86±0.06	51.08±0.05

Table 8: F1 scores (in %) of eight models on the *binary* classification of figurative meaning types in memes from three datasets with the second prompt template. *-Text* and *-Image* denote the ablation of the text or image modality, respectively. Following Liu et al. (2022a), in FigMemes we consider memes with 0 labels as negative cases, and take the best-performing benchmark score as baseline. Results are averaged over five runs.

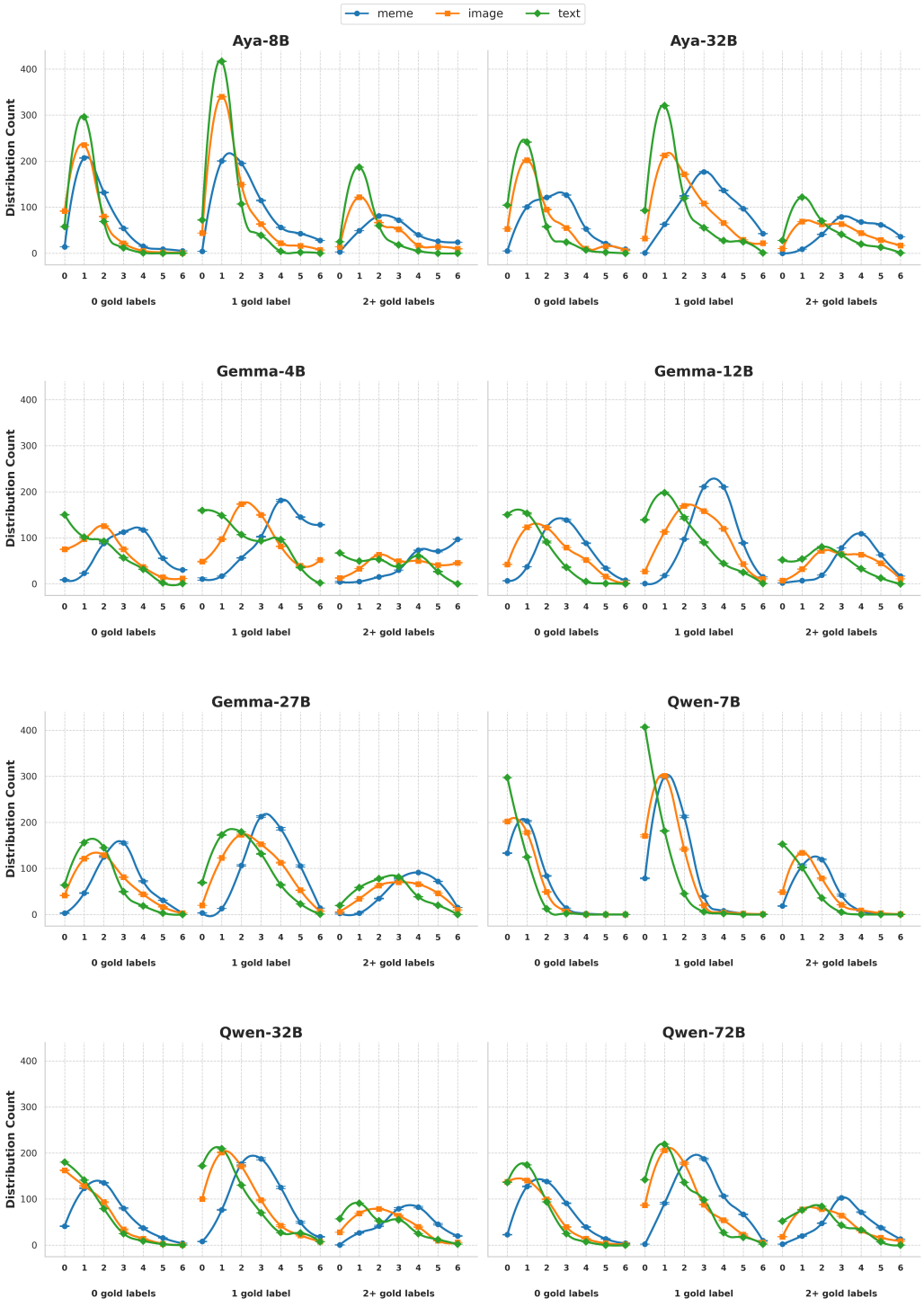

Figure 9: Distribution of predicted label counts across eight models. Each subplot shows the frequency of models predicting 0 to 6 labels, grouped by the actual number of gold labels (0, 1, and 2+).

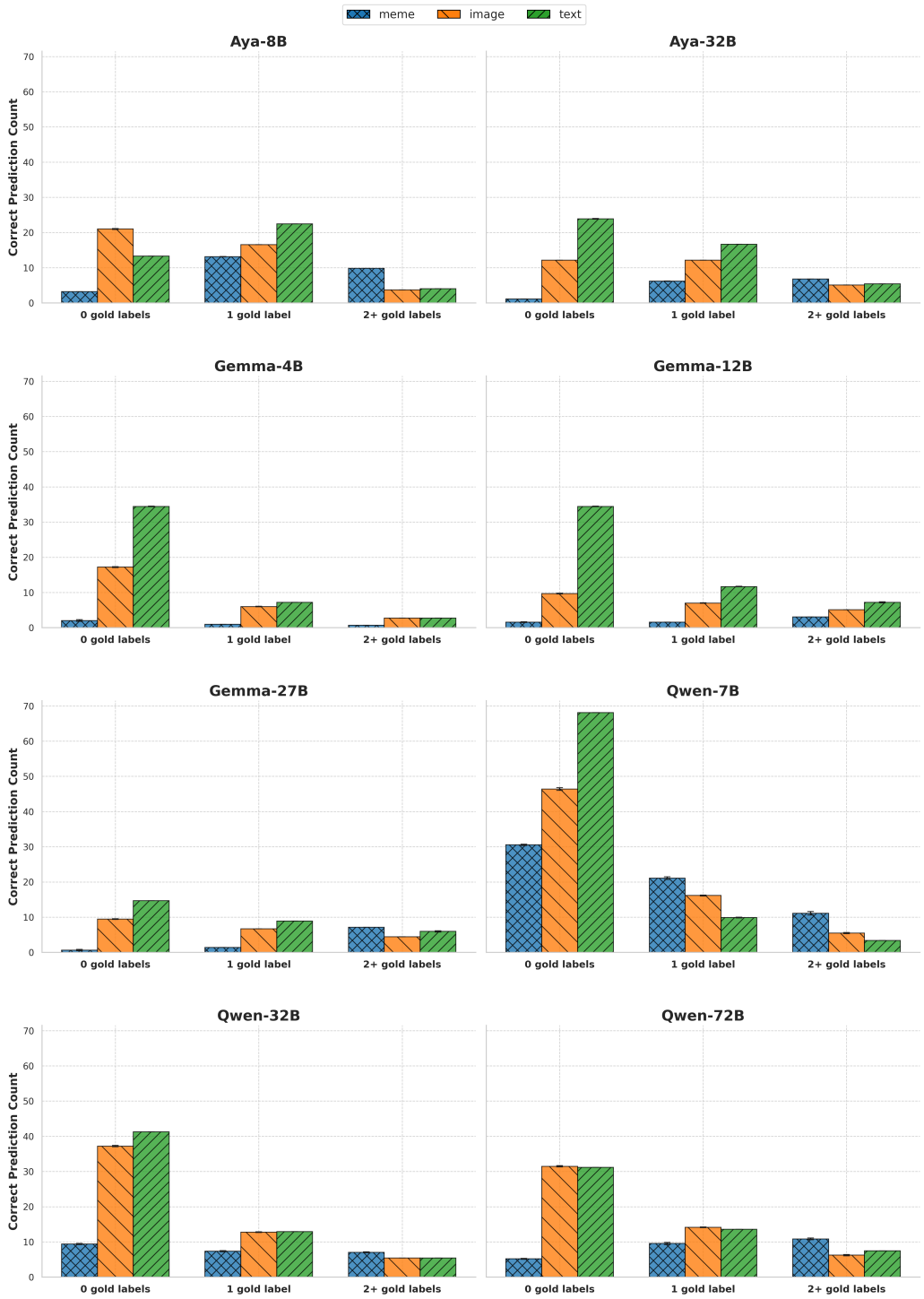

Figure 10: Correct prediction counts for eight models across different label complexities. The bar charts represent the Mean and Standard Deviation (error bars) over five experimental seeds.