Are Large Language Models Chronically Online Surfers? A Dataset for Chinese Internet Meme Explanation

Anonymous ACL submission

Abstract

Large language models (LLMs) are trained on vast amounts of text from the Internet, but do they truly understand the viral content that rapidly spreads online—commonly known as memes? In this paper, we introduce CHIME, a dataset for **CH**inese **I**nternet **M**eme **E**xplanation. The dataset comprises popular phrase-based memes from the Chinese Internet, annotated with detailed information on their meaning, origin, example sentences, types, etc. To evaluate whether LLMs understand these memes, we designed two tasks. In the first task, we assessed the models' ability to explain a given meme, identify its origin, and generate appropriate example sentences. The results show that while LLMs can explain the meanings of some memes, their performance declines significantly for culturally and linguistically nuanced meme types. Additionally, they consistently struggle to provide accurate origins for the memes. In the second task, we created a set of multiple-choice questions (MCQs) requiring LLMs to select the most appropriate meme to fill in a blank within a contextual sentence. While the evaluated models were able to provide correct answers, their performance remains noticeably below human levels. We include CHIME with the submission and hope it will facilitate future research on computational meme understanding.

Meme: treetree 的

Profanity: No **Offense:** No **Type:** Homophonic Pun

Meaning

"treetree 的" 是一个谐音梗，通常用来形容食物或物品的口感或外观上 "脆脆的" 感觉。

(A homophonic pun typically used to describe the texture or appearance of food or items that feel or look "crunchy.")

Origin

源于吃播，在直播中主播因为口音或习惯将 "脆脆" 发音为 "tree tree"，之后被网友在评论区中玩梗并传播开来，尤其在抖音等平台上常见。

(Originating from mukbang livestreams, this term came about when a streamer pronounced "crunchy" as "tree tree" due to their accent or speaking habits. It later became a popular meme among netizens in comment sections and spread widely, especially on platforms like Douyin (TikTok).)

Examples

1. 这款薯片好好吃，入口就是 treetree 的感觉。*(These chips are so delicious; they have that treetree texture as soon as you bite into them.)*
2. 每次吃这种饼干，我都觉得 treetree 的，让人忍不住想多吃几块。*(Every time I eat these cookies, they feel treetree, making it impossible to resist eating a few more.)*
3. 你试试这个油条，刚炸完，treetree 的。*(Try this fried dough stick—it's freshly made and super treetree!)*

Figure 1: A sample from our CHIME dataset.

1 Introduction

An Internet meme is a cultural item that conveys a specific idea, behavior, or style and spreads rapidly online, especially through social media and messaging platforms. While memes often gain popularity for their humorous and playful nature, they also reflect various facets of social, political, and cultural discourse (Szablewicz, 2014; Zhang and Kang, 2024). Internet memes take many forms, including phrases, images, and videos. In China, phrase-based memes have become a significant part of Internet culture, offering a distinctive blend of linguistic and cultural nuances. These phrases are typically short and straightforward. For example, some memes originate from slang (e.g., 熊孩子, "*brat*"), others are abbreviations (e.g., yyds/永远的神, "*the GOAT*" or "*the greatest of all time*"), and some are created using phonetic transformations (e.g., 因缺思厅, "*interesting*").

Despite their playful appearance, Internet memes pose intriguing challenges for natural language understanding systems. They often rely on subtle wordplay, intertextual references, and constantly evolving cultural contexts, making them difficult even for humans to interpret without sufficient background knowledge (Kostadinovska-Stojchevska and Shalevska, 2018). Specifically, Chinese Internet memes present unique challenges due to their use of puns, phonetic transformations, and extensive cultural references. Such memes frequently originate from online communities like Douyin (TikTok) and Weibo, where they can gain national attention in a matter of hours or days. Additionally, Chinese meme culture tends to blend homophones, dialect expressions, and creative abbreviations, resulting in content that is not only linguistically complex but also deeply rooted in shared social contexts. Recent advancements in large language models (LLMs) (OpenAI, 2024; Anthropic, 2024; Meta, 2024; Zhipu AI, 2024; Qwen Team, 2024; DeepSeek-AI, 2024) have shown promise in many natural language tasks, including conversational agents, information extraction, and machine translation. These models were pre-trained on vast amounts of text data from the Internet, which includes memes. However, whether these models can effectively capture the shifting and nuanced semantics of memes remains an open question.

To close this gap, we introduce the CHIME (**CH**inese **I**nternet **M**eme **E**xplanation) dataset—a collection of widely used Chinese phrase-based memes, each annotated with detailed metadata on its meaning, origin, example usage, etc. (see Figure 1 for a sample). Our goal is twofold. First, by assembling memes of varying linguistic complexity and cultural depth, CHIME serves as a resource to test whether LLMs can go beyond surface-level understanding. Second, by including annotations such as etymology and contextual usage, CHIME provides a more nuanced evaluation framework for computational meme comprehension. We posit that assessing how LLMs handle these memes offers fresh insights into the models' capabilities—and limitations—in reasoning about culturally rich, rapidly evolving content.

To this end, we propose two main tasks. The first task is an explanation-centric evaluation, where LLMs must describe a meme's meaning, provide its origin, and generate an appropriate example sentence. This setup probes both the breadth of the models' knowledge (e.g., recognizing the source and historical context of a meme) and the depth of their linguistic capabilities (e.g., producing example usage that aligns with social norms and cultural connotations). The second task is a multiple-choice question (MCQ) test, where the model must select the most fitting meme to fill in a blank within a contextual sentence. This requires not only semantic understanding but also the ability to discern subtle differences between multiple memes with overlapping or related meanings. Our findings suggest that while current LLMs can sometimes provide accurate meme explanations—especially for more straightforward or widely disseminated memes—their performance declines markedly for culturally and linguistically intricate cases. Furthermore, they struggle to pinpoint the correct origin of many memes, revealing gaps in their domain knowledge and context comprehension. By highlighting these challenges, we aim to spur further research in computational approaches for meme understanding, particularly those that incorporate cultural context into language models. We believe CHIME will pave the way for future investigations into how LLMs process and understand socially driven content on the Internet and contribute to the development of more humorous and human-like conversational agents.

2 Related Work

2.1 Meme Datasets

The concept of "meme" was first introduced by biologist Richard Dawkins in his book *The Selfish Gene* (Dawkins, 2016). The term "Internet meme" was formally defined by Castaño Díaz (2013) as a phrase, image, or video associated with real-life events that spreads widely online. Internet memes often employ humor as a means to convey and propagate their underlying message. Existing meme datasets mainly focus on image-based memes. Li et al. (2022) introduced a multimodal dataset for humor analysis using meme templates. Their study treats memes as image-text combinations, where a single image paired with different text can create varied humorous effects. The dataset includes 203 templates (images with text slots) and 5,184 annotated memes, each rated for humor levels. Xu et al. (2022) introduced MET-Meme, a multimodal meme dataset rich in metaphorical features. It contains 10,045 text-image pairs and has been used to demonstrate the importance of metaphor in sentiment analysis and semantic understanding. Additional multimodal meme datasets for identifying offensive content are available in (Hossain et al., 2022; Suryawanshi et al., 2020). In our research, we develop a novel meme explanation dataset that focuses exclusively on text, with the goal of accurately explaining phrase-based memes.

2.2 Humor Datasets

Humor is defined as the tendency of experiences to evoke laughter and provide amusement. Traditionally, humorous content has been represented as plain text. Zhang and Liu (2014) developed a humor recognition model to identify humorous tweets on Twitter, utilizing various linguistic features to achieve high accuracy. Yang et al. (2015) introduced humor datasets for classification, with positive examples from Pun of the Day[1] and the One-Liner dataset (Mihalcea and Strapparava, 2005), and negative examples from Yahoo Answers, The New York Times, AP News, and Proverbs. Additionally, Weller and Seppi (2019, 2020) presented a humor dataset extracted from Reddit. He et al. (2024) introduced *Chumor*, a Chinese humor dataset sourced from a Reddit-like platform, which contains jokes manually annotated with human explanations. Chen et al. (2024) proposed TalkFunny, a Chinese explainable humorous response dataset, which contains context-response pairs featuring chain-of-humor and humor mind map annotations.

Recent studies on computational humor have also focused on multimodal humor datasets. Hasan et al. (2019) constructed a multimodal humor dataset comprising TED videos and their English transcripts. Wu et al. (2021) proposed MUMOR, a multimodal humorous dialogue dataset sourced from TV-sitcoms, in both English and Chinese. Radev et al. (2016) analyzed a dataset of cartoons from *The New Yorker* paired with captions submitted by various users, evaluating the most humorous captions. Hessel et al. (2023) created humor benchmarks using The New Yorker Cartoon Caption Contest to assess three tasks: caption-cartoon matching, caption ranking, and humor explanation. Both multimodal and language-only models were tested, but results showed poor performance across all tasks, underscoring the challenges in computational humor understanding.

In our research, we focus on Chinese phrase-based memes, which are a unique form of humorous content and have been rarely explored in existing literature.

3 Dataset

The CHIME dataset was developed by collecting human-written meme explanations from online sources, followed by the automatic extraction of key information and subsequent manual verification. Each entry in the dataset is manually annotated with labels for meme type and the presence of profanity and offensive content. The following subsections provide a detailed explanation of these processes.

3.1 Raw Data Collection

We first collected human-written meme explanations from Geng Baike (梗百科, *Meme Encyclopedia*)[2], a website where users can contribute articles explaining specific phrase-based memes popular on the Chinese Internet. The explanations collected were created between August 17, 2020, and September 23, 2024. The data were then cleaned by correcting typographical errors and removing duplicates.

To filter out memes that are too niche, five annotators (three of the authors and two recruited individuals) reviewed all the collected meme explanations, indicating whether they were familiar with each one. The annotators, all frequent Internet users with adequate digital literacy, represent a range of birth years from the 1980s to the 2000s. We retained only those memes recognized by at least one of the five annotators. This process resulted in a final collection of 1,458 meme explanations.

3.2 Key Information Extraction

Since the crawled meme explanations were written by different individuals, they vary in format and style. To ensure consistency and extract relevant information, we utilized a large language model (LLM) to automatically identify and extract key elements from the explanations. Specifically, we focused on the following aspects:

- **Meaning**: A concise explanation of the meme, provided in a few sentences.
- **Origin**: The source of the meme, such as a famous movie, a celebrity quote, a TV show, or other cultural references. This information is included when available but is optional.
- **Examples**: For each meme, we extract up to three example sentences illustrating its usage. If the original explanation does not include examples, the LLM generates them.

[1] http://www.punoftheday.com/

[2] https://gengbaike.cn/

We asked GPT-4o (OpenAI, 2024) to extract the three components described above from each crawled meme explanation, using the prompt in Appendix B. However, the output of GPT-4o was not always fully accurate or reliable, as LLMs are known to generate erroneous or unfaithful content, commonly referred to as hallucinations (Huang et al., 2023). Additionally, some of the extracted examples were generated by GPT-4o rather than originating from human-written explanations. As a result, we manually reviewed all extracted information to ensure the accuracy of the meanings and origins, verify that no key details were omitted, and confirm that the examples appropriately demonstrated the usage of each meme.

3.3 Manual Annotation

To ensure the dataset meets safety and ethical standards, each meme was manually annotated with two labels: a **profanity** label, indicating the presence of sexually explicit content, and an **offense** label, marking content that may be offensive, such as racism or discrimination. One of the authors conducted the initial annotation, which was then verified by the other two authors.

Additionally, each meme was classified into one of the following types, based on a predefined taxonomy:

- **Experience** (现象): Memes derived from individuals summarizing their personal experiences or situations. These are often used to express limitations or unmet expectations, serving as a form of self-relief or self-deprecation.
- **Quotation** (引用): Memes originating from historical stories, public events, movie plots, TV shows, or celebrity quotes.
- **Stylistic device** (修辞): Memes crafted using rhetorical techniques such as metaphor, irony, or sarcasm, often to convey auxiliary ideas or emotions.
- **Homophonic pun** (谐音): Memes created by replacing original characters with those of similar or identical sounds to produce humorous or meaningful effects.
- **Slang** (俗语): Memes based on widely recognized and popular colloquial expressions specific to a particular time or place.
- **Abbreviation** (缩写): Memes formed by shortening proper nouns or general phrases. The abbreviation methods vary and include morpheme reductions, initialisms, and simplified spellings.

# Profanity	75 (5.1%)
# Offense	127 (8.7%)
# Experience	561 (38.5%)
# Quotation	438 (30.0%)
# Stylistic device	214 (14.7%)
# Homophonic pun	133 (9.1%)
# Slang	60 (4.1%)
# Abbreviation	52 (3.6%)
# Total	1,458

Table 1: Statistical overview of the CHIME dataset.

Table 1 presents the statistical overview of the CHIME dataset.

4 Can LLMs Explain Memes?

The CHIME dataset could serve as a benchmark to assess the ability of LLMs to interpret and generate explanations for memes without prior fine-tuning. To explore this capability, we conducted experiments where candidate language models are tasked with interpreting and generating explanations for memes from the CHIME dataset.

4.1 Experimental Setup

In this experiment, we employ a zero-shot setting, prompting the candidate language models to explain the meaning of a given Internet meme, provide its origin (if available), and construct an example sentence. The prompts used can be found in Appendix C. The evaluated language models include GPT-4o (OpenAI, 2024), Claude 3.5 Sonnet (Anthropic, 2024), GLM-4-9B, GLM-4-Plus (Zhipu AI, 2024), Qwen2.5-7B, Qwen2.5-72B (Yang et al., 2024; Qwen Team, 2024), and DeepSeek-V3 (DeepSeek-AI, 2024).

To assess and compare their performance across the six meme types, we randomly selected 40 memes from each type, resulting in a testing set of 240 memes. During the selection process, we deliberately excluded all memes that gained popularity after the training cut-off dates of the evaluated models. This same testing set was used for both automatic and human evaluation to facilitate direct comparison of the results.

Model	Cosine Similarity		BERTScore (F)		BARTScore (F)	
	Meaning	Origin	Meaning	Origin	Meaning	Origin
GPT-4o	0.815	0.647	0.800	0.675	−4.485	−4.717
Claude 3.5 Sonnet	0.788	0.625	0.789	0.696	−4.611	−4.695
GLM-4-9B	0.813	0.578	0.797	0.663	−4.453	−4.560
GLM-4-Plus	**0.844**	0.679	**0.822**	0.737	**−4.291**	−4.441
Qwen2.5-7B	0.792	0.605	0.782	0.661	−4.494	−4.779
Qwen2.5-72B	0.819	0.627	0.803	0.690	−4.366	−4.605
DeepSeek-V3	0.779	**0.709**	0.774	**0.751**	−4.331	**−4.344**

Table 2: Average cosine similarity, BERTScore, and BARTScore across all six meme types for each candidate model. The best-performing scores are highlighted in **bold**.

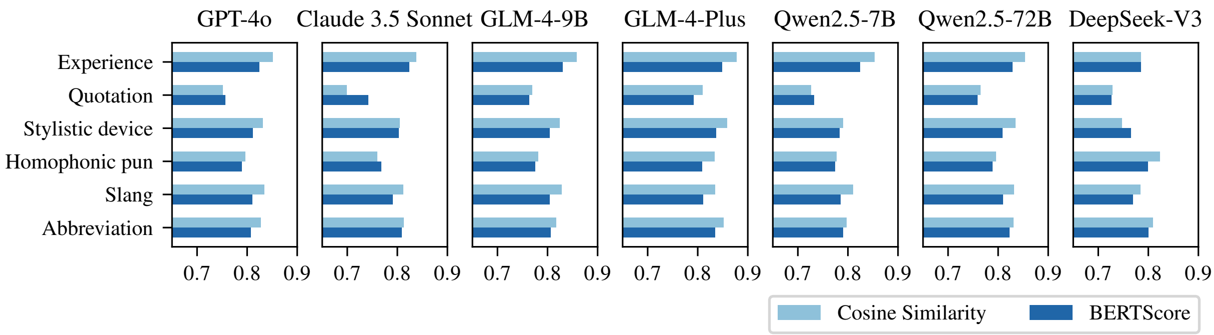

Figure 2: Average cosine similarity and BERTScore for the generated meanings of the candidate models, evaluated across each of the six meme types.

4.2 Automatic Evaluation

The purpose of automatic evaluation is to compare the LLM-generated meaning and origin of a meme with its ground truth meaning and origin. We adopted the following metrics:

- **Cosine similarity**. We used the BGE embedding model (*bge-large-zh-v1.5*) (Xiao et al., 2024) to generate sentence embeddings of the hypothesis and reference and calculated the cosine similarity between them.

- **BERTScore** (Zhang et al., 2020). BERTScore measures the similarity between the hypothesis and reference by summing the cosine similarities of their token embeddings. Here, we also employed the BGE embedding model to generate the token vector representations.

- **BARTScore** (Yuan et al., 2021). BARTScore utilizes an encoder-decoder language model to assess the likelihood that the hypothesis and reference are paraphrases. We used *bart-large-chinese* (Shao et al., 2024) for the underlying BART model.

Overall Results Table 2 presents the average cosine similarity, BERTScore, and BARTScore across all six meme types for each of the six candidate models. Since the BGE model was fine-tuned using contrastive learning, the absolute values of cosine similarity and BERTScore may not directly reflect performance quality; instead, the relative rankings are more informative. As shown in the table, GLM-4-Plus achieves the highest scores on the meaning task, while DeepSeek-V3 achieves the highest scores on the origin task. Additionally, all models perform better on the meaning task compared to the origin task, suggesting that identifying a meme's origin is more challenging than explaining its meaning. When comparing models of different sizes within the same series (e.g., GLM-4-9B versus GLM-4-Plus and Qwen 2.5-7B versus Qwen 2.5-72B), we observed that larger models consistently outperform their smaller counterparts.

Meme Type Specific Results Figure 2 provides a detailed breakdown of meaning scores (cosine similarity and BERTScore) for each of the six meme types. Among these types, *quotation* and *homophonic pun* emerge as the most challenging to ex-

Model	Meaning (%)			Origin (%)			Example (%)		
	A	N	D	A	N	D	A	N	D
GPT-4o	53.9	9.0	37.1	18.5	8.2	73.3	55.0	8.3	36.7
Claude 3.5 Sonnet	51.0	9.7	39.3	14.4	10.2	75.4	51.7	7.5	40.8
GLM-4-9B	40.4	9.0	50.6	7.7	10.3	82.0	41.1	6.0	52.9
GLM-4-Plus	68.5	8.9	22.6	**35.9**	8.7	55.4	70.7	5.6	23.7
Qwen2.5-7B	33.9	11.4	54.7	9.7	6.2	84.1	34.0	9.9	56.1
Qwen2.5-72B	45.7	10.0	44.3	14.4	10.2	75.4	46.8	6.8	46.4
DeepSeek-V3	**73.6**	10.3	**16.1**	35.4	12.3	**52.3**	**77.4**	6.2	**16.4**

Table 3: Average percentage of human ratings assigned as *Agree*, *Neutral*, and *Disagree* across all six meme types for each candidate model. A stands for *Agree*, N stands for *Neutral*, and D stands for *Disagree*. The best-performing scores are highlighted in **bold**.

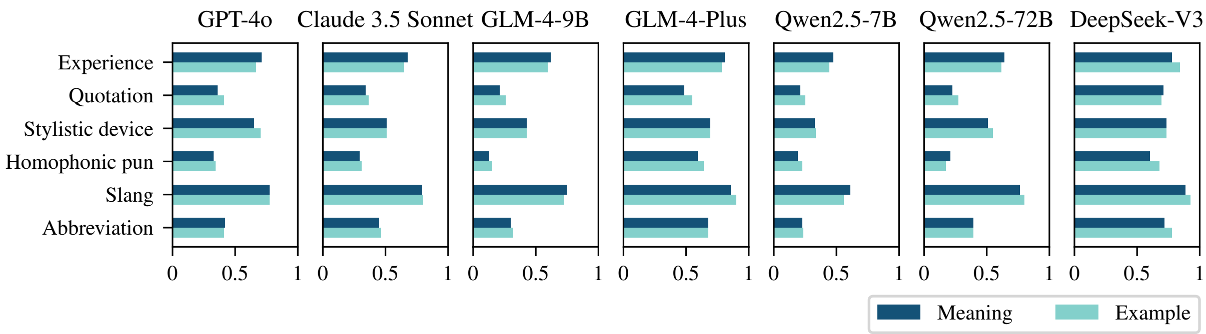

Figure 3: Average percentage of human ratings assigned as *Agree* for the generated meanings and example sentences of the candidate models, evaluated across each of the six meme types. The results of the origin task are omitted, as most memes with an identifiable origin belong to the *quotation* type.

plain. For exact meaning scores for each meme type, refer to Appendix D.

4.3 Human Evaluation

To provide a more comprehensive and accurate assessment of the candidate models' performance—particularly for the generated example sentences, which cannot be effectively evaluated through automated methods—we conducted a human evaluation. We recruited individuals to rate the content generated by the language models. For each testing meme, raters were first shown the true meaning, origin (if available), and three example sentences. Then, for each of the seven candidate models, raters were asked to evaluate the generated meaning, origin (if available), and example sentences using a 3-point Likert scale based on the following statements:

1. The explanation is completely accurate and aligns perfectly with the actual **meaning** of the meme. (*Disagree*, *Neutral*, *Agree*)

2. The provided **origin** perfectly matches the source of the meme without any discrepancies. (*Disagree*, *Neutral*, *Agree*)

3. The **example sentence** accurately reflects the actual usage of the meme, clearly and effectively demonstrating its meaning. (*Disagree*, *Neutral*, *Agree*)

The 240 testing memes were divided into 12 batches, each containing 20 memes for evaluation. For each batch, ratings were collected from three independent raters. More details of the human evaluation process are provided in Appendix E.

Overall Results For each group of meme evaluation tasks, we calculated the Fleiss' kappa score to assess inter-annotator agreement. The average Fleiss' kappa score across all 12 groups is 0.442, indicating moderate agreement among the raters. The results of the human evaluation are presented in Table 3, which shows the average percentage of ratings assigned as *Agree*, *Neutral*, and *Disagree* for each model, based on the aspects of meaning, origin, and example sentence. Different from the automatic evaluation results, DeepSeek-V3 demonstrates the best performance on the meaning and example tasks. All models perform significantly worse on the origin task compared to the meaning and example tasks, and larger models generally outperform their smaller counterparts.

Meme Type Specific Results Figure 3 provides a comparison of all models' performance across the six meme types, showing the percentage of *Agree* ratings for the meaning and example tasks. A strong correlation is observed between these two tasks, indicating that a model capable of accurately explaining the meaning of a meme is also likely to generate appropriate example sentences. Similar to the automatic evaluation results, *quotation*, *homophonic pun*, and *abbreviation* are identified as the most challenging meme types to explain. Additional details of the human evaluation are provided in Appendix F.

4.4 Discussion

Both automatic and human evaluations reveal significant variation in the performance of LLMs across different types of memes. While the models perform relatively well on *experience* and *slang* memes, their performance on *quotation*, *homophonic pun*, and *abbreviation* memes is considerably lower. This disparity likely stems from the nature of these meme types: *experience* memes often convey their meanings more directly, and *slang* memes are typically well-known expressions used in local dialects, making them more prevalent in training data. In contrast, understanding *quotation* memes often requires knowledge of their origin and contextual usage, while *homophonic pun* and *abbreviation* memes involve complex linguistic features that are harder to interpret at first glance. These findings suggest that comprehending memes with strong cultural and linguistic nuances remains a challenging task for LLMs, despite their advancements in overall language processing.

Though both evaluation methods indicate that GLM-4-Plus and DeepSeek-V3 are the two best-performing models, the rankings of the remaining models differ between automatic and human evaluations. Additionally, automatic metrics provide limited discriminatory power, as the scores among models are often quite close. While these metrics offer a quantitative measure of performance, they fail to capture subtleties such as contextual consistency and appropriateness in the generated content. The human evaluation results underscore the importance of incorporating qualitative assessments, particularly for tasks that demand nuanced understanding.

5 Can LLMs Use Memes?

To further assess the comprehensive capabilities of LLMs in understanding and applying Internet memes, we designed a second experiment. In this task, the LLMs are presented with a contextual sentence where the targeted meme is omitted, and they are required to select the most appropriate meme to fill in the blank.

5.1 Experimental Setup

In this experiment, we created a set of multiple-choice questions (MCQs) to evaluate the ability of candidate LLMs to select the most appropriate meme to complete a blank in a contextual sentence. Specifically, for each meme in the CHIME dataset, we randomly selected one of its example sentences and masked the targeted meme. We then identified four other memes with the highest cosine similarity, based on BGE embeddings, to serve as distractor options in the MCQ. For each meme type, we randomly selected 40 MCQs, resulting in a total of 240 MCQs for the testing set.

For each MCQ, the candidate models were prompted to choose the most appropriate meme from the given options while also generating an exemplar. The prompt used is provided in Appendix G. Each MCQ was presented to the models five times, with the final prediction determined by majority voting. To mitigate potential biases in LLMs toward specific answer positions (Zheng et al., 2024; Sabour et al., 2024), we further shuffled the order of the answer choices in four additional permutations, repeating the prediction process for each permutation. The average accuracy across these five runs was reported.

5.2 Results

Table 4 presents the accuracy of the candidate models on the MCQs, along with human performance. The results show that DeepSeek-V3 achieves the highest accuracy among the candidate models, outperforming the other models across all six meme types. The accuracy of the models varies significantly across different meme types, with *experience* and *slang* memes yielding higher accuracy compared to *stylistic device* and *homophonic pun*

Model	Experience	Quotation	Stylistic Device	Homophonic Pun	Slang	Abbreviation	Average
GPT-4o	0.795	0.740	0.700	0.590	0.850	0.760	0.739
Claude 3.5 Sonnet	0.785	0.735	0.710	0.625	0.825	0.770	0.742
GLM-4-9B	0.635	0.510	0.435	0.370	0.650	0.505	0.518
GLM-4-Plus	0.750	0.775	0.680	0.690	0.815	0.780	0.748
Qwen2.5-7B	0.690	0.400	0.475	0.300	0.600	0.490	0.493
Qwen2.5-72B	0.730	0.615	0.655	0.420	0.850	0.685	0.659
DeepSeek-V3	**0.820**	**0.855**	**0.785**	**0.705**	**0.870**	**0.795**	**0.805**
Human (Average)	0.933	0.825	0.833	0.883	0.950	0.892	0.886
Human (Best)	0.950	0.850	0.925	0.900	0.950	0.900	0.913

Table 4: Accuracy of the candidate models on the multiple-choice questions, along with human performance. The best-performing scores of the models are highlighted in **bold**.

Model	Accuracy
GPT-4o	0.898
Claude 3.5 Sonnet	0.872
GLM-4-9B	0.700
GLM-4-Plus	0.891
Qwen2.5-7B	0.778
Qwen2.5-72B	0.887
DeepSeek-V3	**0.918**

Table 5: Accuracy of the candidate models on the multiple-choice questions, **where the meaning of each meme option was provided to the LLMs**. The best-performing scores are highlighted in **bold**.

memes. As expected, larger models generally perform better than smaller models. The human performance, obtained from three recruited individuals, serves as a general upper bound, with the average accuracy of human raters surpassing that of the models. The best human performance is also provided for reference.

5.3 Discussion

The results of the MCQ experiment demonstrate that LLMs can effectively leverage their learned knowledge to select the most appropriate meme to complete a contextual sentence. However, the accuracy of the models varies across different meme types, with models performing much worse on linguistically more nuanced memes such as *stylistic device* and *homophonic pun*. This discrepancy is consistent with the findings from the meme explanation task, suggesting that the complexity of meme types significantly impacts the interpretive capabilities of LLMs.

We also conducted an experiment where the meaning of each meme option was provided to the LLMs, aiming to evaluate the impact of additional context on the models' performance. Table 5 presents the results in this setting. When the meaning of each meme option was provided to the models, the accuracy of all models increased, with the gap between the models narrowing. This finding suggests that LLMs can benefit from additional context to enhance their understanding and selection of memes, particularly for memes that involve complex linguistic features or cultural references.

6 Conclusion

This paper introduces CHIME, a novel dataset designed for the explanation of Chinese Internet memes. Each meme in the dataset is annotated with detailed information, including its meaning, origin, example sentences, and auxiliary labels, creating a robust benchmark for evaluating and enhancing the interpretive capabilities of LLMs. Through a comprehensive experimental framework, we evaluated the performance of seven prominent LLMs, uncovering significant variability in their ability to explain memes across different types. In addition, we designed a multiple-choice question (MCQ) experiment in which models select the most appropriate meme to complete a contextual sentence, further highlighting the challenges in computational meme understanding, particularly for culturally and linguistically nuanced content. Future work could explore expanding the dataset to include multimodal memes and developing models that deliver more engaging and human-like conversational experiences with the support of the CHIME dataset.

7 Limitations

While the CHIME dataset provides a comprehensive benchmark for evaluating the interpretive capabilities of LLMs, it has several limitations. First, the dataset is limited to Chinese Internet memes, which may not fully represent the diversity of memes across different cultures and languages. Second, the dataset focuses on textual content, excluding multimodal memes that incorporate images, videos, or other media. Third, the reliance on human annotations introduces potential subjectivity and bias, and the limited number of annotators may affect the consistency of labeling. Lastly, the dataset captures memes from a specific time period, so its relevance may diminish as meme culture rapidly evolves. Future work could address these limitations by expanding the dataset to include a broader range of meme types and modalities, increasing annotation diversity, and continually updating the dataset to reflect the dynamic nature of meme culture.

8 Ethical Considerations

The CHIME dataset was created with the utmost care to ensure that all content is safe and appropriate for research purposes. We conducted manual annotation to identify and label any potentially offensive or inappropriate content, including profanity and discriminatory language. We acknowledge that Internet memes can sometimes perpetuate harmful stereotypes or biases, and we have taken care to document these occurrences through our labeling system to enable responsible research. We also considered the privacy implications of including user-generated content and took steps to anonymize any personally identifiable information.

The broader impacts of this work are both positive and potentially concerning. On the positive side, this dataset can help advance our understanding of how cultural information spreads online and how language models process culturally-embedded content. It may also aid in developing more culturally aware AI systems. However, we acknowledge potential risks, such as the dataset being used to generate misleading content or manipulate online discourse. We encourage researchers using our dataset to consider these ethical implications and implement appropriate safeguards in their work.

References

Anthropic. 2024. Introducing Claude 3.5 Sonnet.

Carlos Mauricio Castaño Díaz. 2013. Defining and characterizing the concept of Internet meme. *Ces Psicología*, 6(2):82–104.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024. Talk Funny! A large-scale humor response dataset with chain-of-humor interpretation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17826–17834. AAAI Press.

Richard Dawkins. 2016. *The Selfish Gene*. Oxford University Press.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Md. Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2046–2056. Association for Computational Linguistics.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2024. Chumor 2.0: Towards benchmarking Chinese humor understanding. *CoRR*, abs/2412.17729.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Student Research Workshop, Online, November 20, 2022*, pages 32–39. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Bisera Kostadinovska-Stojchevska and Elena Shalevska. 2018. Internet memes and their socio-linguistic features. *European Journal of Literature, Language and Linguistics Studies*, 2(4).

Zefeng Li, Hongfei Lin, Liang Yang, Bo Xu, and Shaowu Zhang. 2022. Memeplate: A Chinese multimodal dataset for humor understanding in meme templates. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 527–538. Springer.

Meta. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 531–538. The Association for Computational Linguistics.

OpenAI. 2024. Hello GPT-4o.

Qwen Team. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Dragomir R. Radev, Amanda Stent, Joel R. Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in The New Yorker Cartoon Caption Contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5986–6004. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2024. CPT: A pre-trained unbalanced Transformer for both Chinese language understanding and generation. *Sci. China Inf. Sci.*, 67(5).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 32–41. European Language Resources Association (ELRA).

Marcella Szablewicz. 2014. The 'losers' of China's Internet: Memes as 'structures of feeling' for disillusioned young netizens. *China Information*, 28(2):259–275.

Orion Weller and Kevin D. Seppi. 2019. Humor detection: A Transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3619–3623. Association for Computational Linguistics.

Orion Weller and Kevin D. Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6136–6141. European Language Resources Association.

Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. 2021. MUMOR: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part I*, volume 13028 of *Lecture Notes in Computer Science*, pages 619–627. Springer.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 641–649. ACM.

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 *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2887–2899. ACM.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard H. Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21,* *2015*, pages 2367–2376. The Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on Twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 889–898. ACM.

Ruichen Zhang and Bo Kang. 2024. From propaganda to memes: Resignification of political discourse through memes on the Chinese Internet. *International Journal of Human–Computer Interaction*, 40(11):3030–3049.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhipu AI. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

A Computing Infrastructure

All the experiments were conducted by invoking the models through their official APIs, with default hyperparameters for generating responses. For GPT-4o, we used the version `gpt-4o-2024-08-06`, and for Claude 3.5 Sonnet, we used the version `claude-3-5-sonnet-20240620`. Total cost for the experiments (including the key information extraction when curating the dataset) was approximately $300, with the majority of the cost attributed to the usage of GPT-4o, Claude 3.5 Sonnet, and GLM-4-Plus.

B Key Information Extraction Prompt

We asked GPT-4o to extract the meaning, origin, and example sentences from the crawled meme explanation using the following prompt:

> 你需要根据提供的互联网流行梗的解释，提取它的含义、出处和3个例句。在提取时，保留所有关键信息，不要过度缩略。*(You need to extract the meaning, origin, and three examples of usage based on the explanation of the provided Internet meme. When extracting, retain all key information without excessive abbreviation.)*

C Explanation Task: Prompts

We gave the following prompts to the candidate models and let them explain the meaning of a given Internet meme, provide its origin (if available), and construct an example sentence:

> *For memes without a known origin:*
> 在中文互联网的语境下，解释以下网络流行梗的含义，并撰写1个例句。*(In the context of the Chinese Internet, explain the meaning of the following viral meme and create one example sentence.)*
>
> *For memes with a known origin:*
> 在中文互联网的语境下，解释以下网络流行梗的含义和出处，并撰写1个例句。*(In the context of the Chinese Internet, explain the meaning and origin of the following viral meme, and create one example sentence.)*

D Explanation Task: More Automatic Evaluation Results

Table 6 gives the exact meaning scores of the candidate models for each of the six meme types.

E Explanation Task: Human Evaluation Details

For our human evaluation process, we first divided the 240 testing memes into 12 batches of 20 memes each. For each batch, we created a questionnaire containing an instruction page followed by 20 evaluation pages (one per meme). The instruction page provided the following guidelines to raters (translated from Chinese):

> Internet memes, as a unique cultural phenomenon, not only reflect societal trends and public emotions but also hold signif-

Model	Experience			Quotation		
	Cos. Sim.	BERTS.	BARTS.	Cos. Sim.	BERTS.	BARTS.
GPT-4o	0.851	0.824	−4.293	0.751	0.757	−4.319
Claude 3.5 Sonnet	0.838	0.824	−4.410	0.699	0.742	−4.407
GLM-4-9B	0.858	0.830	−4.323	0.769	0.763	−4.217
GLM-4-Plus	**0.878**	**0.849**	**−4.086**	**0.810**	**0.792**	**−4.133**
Qwen2.5-7B	0.853	0.824	−4.237	0.726	0.733	−4.317
Qwen2.5-72B	0.854	0.829	−4.181	0.765	0.758	−4.256
DeepSeek-V3	0.785	0.785	−4.211	0.728	0.726	−4.204

Model	Stylistic Device			Homophonic Pun		
	Cos. Sim.	BERTS.	BARTS.	Cos. Sim.	BERTS.	BARTS.
GPT-4o	0.831	0.811	−4.386	0.796	0.789	−4.785
Claude 3.5 Sonnet	0.805	0.803	−4.507	0.760	0.768	−5.033
GLM-4-9B	0.824	0.804	−4.283	0.781	0.775	−4.813
GLM-4-Plus	**0.859**	**0.837**	**−4.198**	**0.834**	**0.809**	−4.588
Qwen2.5-7B	0.790	0.783	−4.424	0.777	0.774	−4.825
Qwen2.5-72B	0.835	0.809	−4.221	0.796	0.789	−4.651
DeepSeek-V3	0.747	0.765	−4.250	0.823	0.799	**−4.562**

Model	Slang			Abbreviation		
	Cos. Sim.	BERTS.	BARTS.	Cos. Sim.	BERTS.	BARTS.
GPT-4o	0.834	0.810	−4.424	0.827	0.807	−4.702
Claude 3.5 Sonnet	0.811	0.791	−4.483	0.813	0.809	−4.823
GLM-4-9B	0.829	0.804	−4.361	0.817	0.806	−4.720
GLM-4-Plus	**0.835**	**0.811**	−4.234	**0.851**	**0.835**	−4.505
Qwen2.5-7B	0.810	0.786	−4.389	0.797	0.790	−4.775
Qwen2.5-72B	0.832	0.810	**−4.227**	0.831	0.822	−4.657
DeepSeek-V3	0.784	0.770	−4.304	0.809	0.800	**−4.456**

Table 6: Average cosine similarity, BERTScore, and BARTScore for the generated meanings of the candidate models, for each of the six meme types. The best-performing scores are highlighted in bold.

icant social influence. To study the understanding of Chinese Internet memes by large language models, this project aims to systematically evaluate Internet memes within the context of the Chinese Internet through a questionnaire survey.

This questionnaire is divided into two parts: The first part will collect your name; the second part consists of 20 pages, each corresponding to one popular meme. You will be required to evaluate the explanations of each meme generated by six large language models across three dimensions: "meaning," "origin," and "example sentence."

You will answer approximately 120 questions, and the survey is expected to take about 40 minutes.

I. Instructions

1. Participation in this survey is entirely voluntary. You have the right to decide whether to participate. Your personal information will be kept strictly confidential and used solely for academic research purposes, with no disclosure to third parties.

2. To ensure the accuracy and reliability of the survey results, please provide

honest answers and avoid random responses or providing false information.

3. Please complete the questionnaire to the fullest extent possible and avoid skipping any questions. If you have any doubts, feel free to contact the project team for clarification.

4. Once you have completed the questionnaire, click the "Submit" button to confirm your submission. Please note that submissions cannot be modified, so review your responses carefully before submitting.

5. Be advised that the questionnaire may contain some vulgar, sexually suggestive, or offensive content. If you feel uncomfortable with such content, please consider whether to proceed.

II. Acknowledgments and Feedback

1. Thank you for taking the time to participate in this survey. Every response you provide will contribute valuable data to our research.

2. If you encounter any issues or have any suggestions while filling out the questionnaire, feel free to contact the project team at any time.

3. After the survey is complete, the project team will analyze the data and prepare a research report. If needed, we will share the results of the study with participants.

Thank you once again for your support and cooperation!

For each questionnaire, ratings were collected from three independent raters. We payed each rater around $14 per hour for their participation, which is much higher than the average hourly wage in China. We reruited a total number of 14 raters for the human evaluation task, and their birth years range from 1980s to 2000s. All raters were native Chinese speakers with a good understanding of Chinese Internet culture.

Batch	Meme Type	Fleiss' kappa
1	Slang	0.278
2	Slang	0.269
3	Stylistic device	0.318
4	Stylistic device	0.487
5	Quotation	0.421
6	Quotation	0.519
7	Experience	0.360
8	Experience	0.393
9	Abbreviation	0.736
10	Abbreviation	0.711
11	Homophonic pun	0.412
12	Homophonic pun	0.400

Table 7: Fleiss' kappa scores on each of the 12 evaluation batches in human evaluation.

F Explanation Task: More Human Evaluation Results

Table 7 gives the Fleiss' kappa scores on each of the 12 evaluation batches. Table 8 provides the detailed human evaluation results on the meaning task for each of the six meme types.

G MCQ Task: Prompts

For the multiple-choice questions (MCQs), we provided the following prompts to the candidate models (with English translation):

根据提供的句子，其中包含一个空白处，请从提供的5个选项中，根据上下文选择最合适的网络流行梗填入。只需给出选项的编号作为答案，不要做任何解释。
示例：
句子：这个方案真是_____，完全超出我的想象。
选项：
(1) 雪糕刺客
(2) yyds
(3) 狗带
(4) 实锤
(5) 偷感很重
答案：2

(English translation)
Based on the given sentence, which contains a blank, choose the most suitable Internet meme from the five provided options accord-

Model	Experience (%)			Quotation (%)			Stylistic Device (%)		
	A	N	D	A	N	D	A	N	D
GPT-4o	70.8	5.9	23.3	35.8	10.9	53.3	65.0	7.5	27.5
Claude 3.5 Sonnet	67.5	6.7	25.8	34.2	8.3	57.5	50.8	12.5	36.7
GLM-4-9B	61.6	1.7	36.7	20.8	15.9	63.3	42.5	8.3	49.2
GLM-4-Plus	**80.8**	3.4	15.9	48.3	15.8	35.8	69.1	9.2	**21.7**
Qwen2.5-7B	47.5	14.2	38.3	20.8	6.7	72.5	32.5	12.5	55.0
Qwen2.5-72B	64.2	3.3	32.5	22.5	15.8	61.7	50.8	12.5	36.7
DeepSeek-V3	77.5	15.0	**7.5**	**70.8**	11.7	**17.5**	**73.3**	3.4	23.3

Model	Homophonic Pun (%)			Slang (%)			Abbreviation (%)		
	A	N	D	A	N	D	A	N	D
GPT-4o	32.5	11.7	55.8	77.5	10.8	11.7	41.7	7.5	50.8
Claude 3.5 Sonnet	29.2	14.2	56.6	79.1	9.2	11.7	45.0	7.5	47.5
GLM-4-9B	12.5	12.5	75.0	75.0	10.0	15.0	30.0	5.8	64.2
GLM-4-Plus	59.2	13.3	**27.5**	85.8	8.4	5.8	67.5	3.3	29.2
Qwen2.5-7B	19.2	10.8	70.0	60.8	15.0	24.2	22.5	9.2	68.3
Qwen2.5-72B	20.8	15.9	63.3	76.6	11.7	11.7	39.2	0.8	60.0
DeepSeek-V3	**60.0**	10.8	29.2	**88.3**	9.2	**2.5**	**71.6**	11.7	**16.7**

Table 8: Average percentage of human ratings assigned as *Agree*, *Neutral*, and *Disagree* of the candidate models for each meme type, on the meaning task. A stands for *Agree*, N stands for *Neutral*, and D stands for *Disagree*. The best-performing scores are highlighted in **bold**.

ing to the context. Only provide the option number as the answer, without any explanation.
Example:
Sentence: This plan is truly ______, completely beyond my imagination.
Options:
(1) Ice Cream Assassin
(2) yyds (similar to GOAT in English)
(3) Go Die
(4) Solid Evidence
(5) Strong Sense of Stealing
Answer: 2

For MCQs where the meaning of each meme option was provided to the LLMs, the prompt was as follows (with English translation):

根据提供的句子，其中包含一个空白处，请从提供的5个选项中，根据上下文选择最合适的网络流行梗填入。只需给出选项的编号作为答案，不要做任何解释。
示例：
句子：这个方案真是______，完全超出我的想象。
选项：
(1) 雪糕刺客。含义："雪糕刺客"指的是那些看似普通但价格高昂的雪糕，购买时让人感到意外和"被刺"的疼痛感。这个表达反映了雪糕价格上涨和意外负担感。
(2) yyds。含义：yyds是"永远的神"的缩写，用来称赞某人或某事物非常优秀，值得敬仰和追随。
(3) 狗带。含义："狗带"是"go die"的谐音，意为去死或者死亡，通常用于幽默或夸张的表达方式。
(4) 实锤。含义："实锤"指的是能够证明某事件真实发生的可靠证据，通常具备较强的说服力。
(5) 偷感很重。含义：形容人在某些情境下感到拘谨、畏缩，显得偷偷摸摸或不自然。
答案：2

(English translation)
Based on the given sentence, which contains a blank, choose the most suitable Internet

meme from the five provided options according to the context. Only provide the option number as the answer, without any explanation.
Example:
Sentence: This plan is truly _____, completely beyond my imagination.
Options:
(1) Ice Cream Assassin. Meaning: “Ice Cream Assassin” refers to seemingly ordinary but unexpectedly expensive ice cream, making people feel “stabbed” by the price. This phrase reflects rising ice cream prices and the unexpected financial burden.
(2) yyds. Meaning: “yyds” is the abbreviation for “永远的神” (Eternal God), used to praise someone or something as excellent, admirable, and worthy of following.
(3) Go Die. Meaning: “Go Die” is a phonetic translation of “狗带” (gǒu dài), meaning “to die” or “go to hell,” often used humorously or exaggeratedly.
(4) Solid Evidence. Meaning: “Solid Evidence” refers to strong and reliable proof that confirms an event or claim, typically carrying strong credibility.
(5) Strong Sense of Stealing. Meaning: This phrase describes someone feeling awkward, timid, or unnatural in a certain situation, appearing sneaky or out of place.
Answer: 2