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

Yubo Xie[1]*, Chenkai Wang[2], Zongyang Ma[3], Fahui Miao[1]
[1]Shanghai Maritime University, Shanghai, China
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3]Xi'an Jiaotong Liverpool University, Suzhou, China
yuboxie@hotmail.com, wangchenkaicn@foxmail.com,
Zongyang.Ma@xjtlu.edu.cn, miaofahui@126.com

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 have made CHIME public[1] and hope it will facilitate future research on computational meme understanding.

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*").

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: An example from our CHIME dataset.

Despite their playful appearance, Internet memes pose intriguing challenges for natural language understanding systems. They often rely on subtle wordplay, intertextual references, and con-

*Corresponding author.
[1]https://github.com/yuboxie/chime

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17062–17083
November 4-9, 2025 ©2025 Association for Computational Linguistics

stantly 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 simplified Chinese phrase-based memes, each annotated with detailed metadata on its meaning, origin, example usage, etc. (see Figure 1 for an example). 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, 1976). 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. Existing meme datasets mainly focus on image-based memes. Li et al. (2022) introduced a multimodal dataset for humor analysis using meme templates. Xu et al. (2022) introduced MET-Meme, a multimodal meme dataset rich in metaphorical features. Hossain et al. (2022); Suryawanshi et al. (2020) introduced multimodal meme datasets for identifying hateful and offensive content, while Lu et al. (2024); Gu et al. (2024) built multimodal Chinese harmful meme datasets. 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 Non-Literal Language

Non-literal language encompasses various forms of expression, including slangs, idioms, and figurative language. Several existing works have focused on the challenges of understanding non-literal language. Zheng et al. (2019); De Luca Fornaciari et al. (2024) focus on idioms and their assessment in LLMs. Liu et al. (2022) assessed language models' ability to interpret figurative language by collecting creative metaphors from crowdsourcing workers. Mei et al. (2024) developed an English slang dataset from Urban Dictionary that reflects Internet language trends. Our dataset differs not only in language but also in how memes and slangs are created: Chinese Internet memes frequently utilize phonetic wordplay and visual puns, whereas English slangs typically rely on Latin letters and tend to favor abbreviations and acronyms. Sun et al. (2024) also constructed an English slang dataset, but primarily from movie subtitles, which may not capture the most recent Internet language trends.

Some other works have focused on toxic and offensive language detection, which may contain Internet slangs. Lu et al. (2023) constructed a fine-grained dataset and insult lexicon to detect Chinese toxic language. Xiao et al. (2024b) evaluated the robustness of language models in detecting disguised Chinese offensive content.

2.3 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. Yang et al. (2015); Weller and Seppi (2019, 2020) introduced various English humor datasets. He et al. (2024) introduced Chumor, a Chinese humor dataset sourced from Ruo Zhi Ba. Chen et al. (2024) proposed TalkFunny, a Chinese explainable humorous response dataset. Recent studies have also focused on multimodal humor datasets. Hasan et al. (2019); Wu et al. (2021); Radev et al. (2016); Hessel et al. (2023) constructed and analyzed humor datasets from various sources like TED videos, TV sitcoms, and The New Yorker cartoons. Our research focuses 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.

We asked GPT-4o (OpenAI, 2024) to extract the three components described above from each crawled meme explanation, using the prompt in

[2] https://gengbaike.cn/

Appendix B.1. 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.

More details on the manual annotation process can be found in Appendix B.2.

Table 1 presents the statistical overview of the CHIME dataset. Appendix B.3 provides additional statistics on the origins of the memes. We also provide a few representative examples for all six meme types in Appendix B.4.

4 Can LLMs Explain Memes?

The CHIME dataset functions as a benchmark for evaluating LLMs' capacity to interpret and explain memes without fine-tuning. To investigate this capability, we tasked candidate models with generating explanations for memes from this dataset.

4.1 Experimental Setup

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.1. We also experimented with one-shot prompting, but the results were mostly inferior to zero-shot prompting (see Appendix C.2). 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).

4.2 Automatic Evaluation

Automatic evaluation was conducted on the entire dataset (1,458 memes), wherein LLM-generated interpretations of meme meaning and origin were systematically compared against the ground truth. We adopted the following metrics: cosine similarity, BERTScore (Zhang et al., 2020), and BARTScore (Yuan et al., 2021). For cosine similar-

Model	Cosine Similarity		BERTScore (F)		BARTScore (F)	
	Meaning	Origin	Meaning	Origin	Meaning	Origin
GPT-4o	0.805	0.628	0.790	0.680	−4.367	−4.695
Claude 3.5 Sonnet	0.773	0.614	0.776	0.679	−4.559	−4.877
GLM-4-9B	0.792	0.640	0.785	0.696	−4.321	−4.493
GLM-4-Plus	**0.832**	0.689	**0.809**	**0.744**	**−4.238**	**−4.423**
Qwen2.5-7B	0.778	0.579	0.765	0.632	−4.448	−4.855
Qwen2.5-72B	0.805	0.622	0.789	0.676	−4.321	−4.602
DeepSeek-V3	0.787	**0.694**	0.782	0.740	−4.289	−4.463

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**.

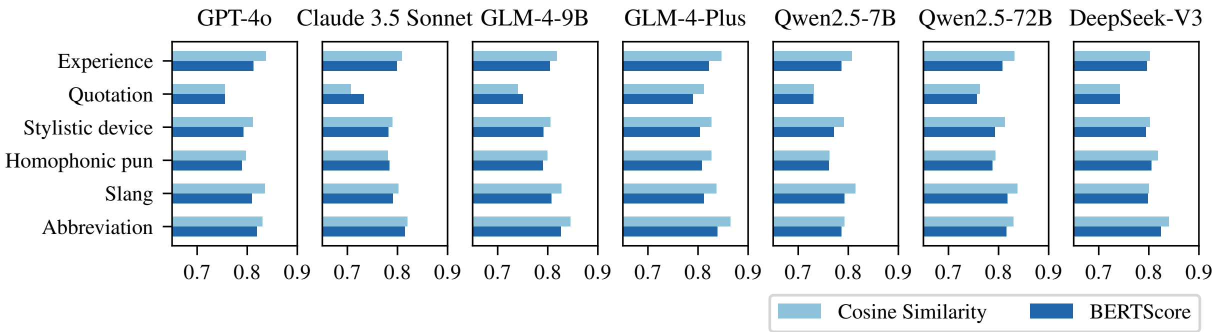

Figure 2: Average cosine similarity and BERTScore for the generated meanings of the candidate models, evaluated across each of the six meme types.

ity and BERTScore, we used the BGE embedding model (*bge-large-zh-v1.5*) (Xiao et al., 2024a) to generate embeddings. For BARTScore, we used *bart-large-chinese* (Shao et al., 2024).

Overall Results Table 2 presents the average cosine similarity, BERTScore, and BARTScore across all six meme types for each of the six candidate models.[3] As shown in the table, GLM-4-Plus achieves the highest scores on most metrics, while DeepSeek-V3 achieves the highest score on the origin task with cosine similarity. 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.

[3]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.

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 explain. For exact meaning scores for each meme type, refer to Appendix C.3.

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

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**.

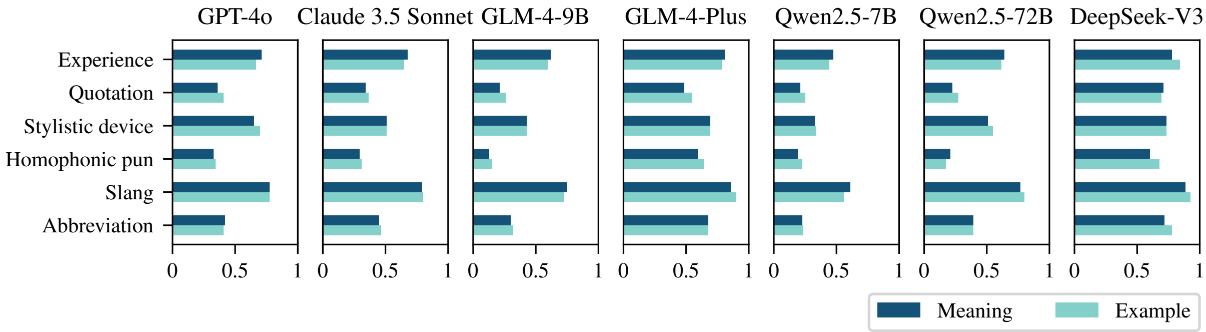

Figure 3: Average percentage of human ratings assigned as *Agree* for the generated meanings and example sentences of the candidate models, evaluated across each of the six meme types. The results of the origin task are omitted, as most memes with an identifiable origin belong to the *quotation* type.

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*)

We randomly selected 240 testing memes (40 per category) and then divided them into 12 batches, each containing 20 memes for evaluation. For each batch, ratings were collected from three independent raters. More details on the human evaluation process are provided in Appendix C.4.

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 on the human evaluation results are provided in Appendix C.5.

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.

Error Analysis To further investigate the performance of LLMs, we conducted an error analysis on the generated meanings and origins. We have identified several consistent patterns: (1) **Origin confusion**: Models frequently attributed memes to incorrect sources, particularly with *quotation* memes. In many instances, LLMs provided vague attributions (e.g., "originating from social media") rather than specific origins. (2) **Semantic shift**: For most misinterpreted *homophonic pun* memes, models explained related concepts with similar phonetics rather than capturing the actual meme meaning. In other cases, models failed to recognize the phonetic wordplay entirely and simply explained the literal meaning. (3) **Cross-type confusion**: *Abbreviation* memes were occasionally misinterpreted as homophonic puns, indicating difficulty in distinguishing between these distinct linguistic mechanisms. We provide a more comprehensive error analysis with illustrative case studies in Appendix C.6.

5 Can LLMs Use Memes?

To evaluate LLMs' comprehensive meme literacy, we designed a second experiment where models must select the most appropriate meme to complete a contextual sentence with an intentional omission.

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. As a result, the final testing set contains 1,268 MCQs.[4]

For each MCQ, the candidate models were prompted to choose the most appropriate meme from the given options. The prompt used is provided in Appendix D.1. 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 except *slang*. The accuracy of the models varies significantly across different meme types, with *experience* and *slang* memes yielding higher

[4]The number of MCQs is less than the total number of memes because we used strict matching for masking targeted memes, but certain example sentences employ memes in a contextually flexible way.

Model	Experience	Quotation	Stylistic Device	Homophonic Pun	Slang	Abbreviation	Average
GPT-4o	0.779	0.708	0.761	0.549	0.858	0.750	0.734
Claude 3.5 Sonnet	0.758	0.644	0.778	0.597	0.800	0.729	0.718
GLM-4-9B	0.574	0.527	0.536	0.360	0.654	0.504	0.526
GLM-4-Plus	0.784	0.748	0.817	0.640	0.804	0.792	0.764
Qwen2.5-7B	0.602	0.520	0.524	0.294	0.642	0.512	0.516
Qwen2.5-72B	0.733	0.691	0.691	0.486	**0.869**	0.671	0.690
DeepSeek-V3	**0.831**	**0.791**	**0.828**	**0.713**	0.858	**0.833**	**0.809**
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	Δ
GPT-4o	0.896	+0.162
Claude 3.5 Sonnet	0.881	+0.163
GLM-4-9B	0.692	+0.166
GLM-4-Plus	0.887	+0.123
Qwen2.5-7B	0.786	+0.270
Qwen2.5-72B	0.881	+0.191
DeepSeek-V3	**0.897**	+0.088

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 score is highlighted in **bold**. The column Δ indicates the improvement in accuracy compared to the setting without meme meanings (Table 4).

accuracy compared to *stylistic device* and *homophonic pun* memes. As expected, larger models generally perform better than smaller models. The human performance, obtained from three recruited individuals on 240 randomly selected MCQs (balanced across meme types), 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 *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 (prompt provided in Appendix D.2). 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.

To further understand the relationship between explanation and usage capabilities, we conducted cross-task analysis and found interesting patterns: (1) Models that correctly explain meme meanings achieve around 83% accuracy in MCQ selection; (2) Conversely, models that select correct memes in context only achieve around 73% accuracy in explanation. This asymmetry suggests that receptive understanding (recognizing appropriate usage) is easier than productive understanding (generating explanations), highlighting the distinct cognitive demands of these two tasks.

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. Particularly, our dataset focuses on Simplified Chinese, because the source platform (Geng Baike) primarily hosts Simplified Chinese content and most Chinese Internet memes originate from mainland China platforms (Douyin, Weibo) where Simplified Chinese dominates. Future work could explore Traditional Chinese memes from Taiwan/Hong Kong platforms. 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. 1976. *The Selfish Gene*. Oxford University Press.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

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

Tianlong Gu, Mingfeng Feng, Xuan Feng, and Xuemin Wang. 2024. SCARE: A novel framework to enhance Chinese harmful memes detection. *IEEE Transactions on Affective Computing*, pages 1–14.

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.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4437–4452. Association for Computational Linguistics.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. 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 16235–16250. Association for Computational Linguistics.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of Chinese harmful memes. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12558–12575. Association for Computational Linguistics.

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

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).

Zhewei Sun, Qian Hu, Rahul Gupta, Richard S. Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1683–1701. Association for Computational Linguistics.

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. 2024a. 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.

Yunze Xiao, Yujia Hu, Kenny T. W. Choo, and Roy Ka-Wei Lee. 2024b. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6012–6025. Association for Computational Linguistics.

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, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese idiom dataset for cloze test. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 778–787. Association for Computational Linguistics.

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, except for GLM-4-9B, which was run on a machine with one Intel Xeon Platinum 8352V 2.10 GHz CPU and two NVIDIA GeForce RTX 4090 GPUs. 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 $1500, with the majority of the cost attributed to the usage of GPT-4o, Claude 3.5 Sonnet, and GLM-4-Plus.

B Dataset Construction

B.1 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.)*

B.2 Manual Annotation

For the annotation of profanity, offense, and meme type, one of the authors conducted the initial annotation, which was then verified by the other two authors. When disagreements arose, the three authors held discussions to resolve any discrepancies, with final determinations made via majority voting. This process yielded perfect agreement (100%) for profanity and offense classifications, and near-perfect agreement (92.9%) for meme type. The final labels were agreed upon by all three authors, ensuring a consistent and accurate representation of the dataset.

During the pilot annotation phase of meme types, we initially followed a taxonomy derived from Chinese literature[5] on Internet memes, which classified memes into four categories based on their creation methodology: *phonetic*, *experiential*, *story-based*, and *sarcastic*. However, when conducting comprehensive annotation of all memes, we discovered that this existing taxonomy could not adequately capture the full range of types we encountered. Consequently, we refined this classification system and expanded it to include two additional categories, resulting in a more comprehensive taxonomy that better represents the diversity of Chinese Internet memes.

[5] https://www.cpd.com.cn/n15737398/n26490099/523/t_1086228.html

B.3 Statistics on Memes' Origins

Most memes with clear origins fall into the *quotation* category (as expected). Of the total 1,458 memes, 525 have clear origins. We have compiled detailed statistics in Table 6.

Meme Type	**# Memes with Origin**
Quotation	411 (78.3%)
Homophonic pun	41 (7.8%)
Experience	36 (6.9%)
Stylistic device	27 (5.1%)
Abbreviation	8 (1.5%)
Slang	2 (0.4%)

Table 6: Statistics on the origins of memes in the CHIME dataset.

B.4 Examples of the CHIME Dataset

Table 7 and Table 8 provide a few representative examples of memes illustrating each meme type in the CHIME dataset.

C Explanation Task

C.1 Zero-Shot Prompts

We gave the following zero-shot 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

Experience

Meme: 没苦硬吃 (Eating Bitterness Unnecessarily)

Meaning: “没苦硬吃” 是形容老一辈人为了节约，在并不需要吃苦的情况下，主动去承受不必要的艰难，从而导致适得其反的行为。(Describes older generations who voluntarily endure unnecessary hardships to save money, often with counterproductive results.)

Examples:

1. 他家明明有空调，但为了省电费，硬是夏天不开，真是典型的 “没苦硬吃”。(They clearly have an air conditioner at home, but to save on electricity bills, they insist on not using it during summer. This is a typical case of “eating bitterness unnecessarily.”)
2. 奶奶总是捡路边的瓶子和纸壳子，看来是又把新家搞成 “没苦硬吃” 的现场了。(Grandma always picks up bottles and cardboard boxes from the roadside. It seems she’s turned her new home into another scene of “eating bitterness unnecessarily.”)
3. 为了省钱总吃剩菜，结果还吃进了医院，这简直就是 “没苦硬吃” 的现实案例。(To save money, they always eat leftover food, which eventually landed them in the hospital. This is a perfect real-life example of “eating bitterness unnecessarily.”)

Quotation

Meme: 臣妾做不到啊 (Your Humble Servant Cannot Do This)

Meaning: “臣妾做不到啊” 用来表达对某些要求或任务的无奈或难以完成，通常用夸张和幽默的方式来表达内心的抗拒和无力感。(Used to express helplessness or inability to meet certain demands or complete tasks. It’s typically used in an exaggerated and humorous way to convey inner resistance and a sense of powerlessness.)

Origin: 该梗源自电视剧《甄嬛传》中的一幕，皇帝要求皇后憎恨他，皇后就哭诉着喊出这句台词。这一台词因其夸张和情感冲击力而被网友广泛引用并恶搞。(Originated from a scene in the TV drama “Empresses in the Palace,” where the emperor demands that the empress hate him, to which the empress tearfully cries out this line. This dialogue became widely quoted and parodied by Internet users due to its exaggerated delivery and emotional impact.)

Examples:

1. 当朋友打算放弃晚餐的甜点时，他诧异地说：“饭后不吃甜点? 臣妾做不到啊！” (When a friend was about to skip dessert after dinner, he exclaimed in surprise: “No dessert after a meal? Your humble servant cannot do this!”)
2. 明知要减肥，却难抵深夜美食诱惑时，她无奈地说道：“臣妾做不到啊！” (Knowing she should be on a diet, yet unable to resist the temptation of late-night snacks, she helplessly said: “Your humble servant cannot do this!”)
3. 他看到明天就要交的论文，心如死灰地在社交媒体上张贴：“明天就要交论文，臣妾做不到，做不到啊！” (Looking at the paper due tomorrow, he hopelessly posted on social media: “The paper is due tomorrow, your humble servant cannot do this, simply cannot do this!”)

Stylistic Device

Meme: 人体描边大师 (Human Outline Master)

Meaning: “人体描边大师” 用来调侃射击类游戏中玩家的糟糕枪法，通常指射击时弹道完美避开目标，就像在目标周围画了个轮廓线。(Used to mock poor shooting skills in shooting games, typically referring to when bullets perfectly avoid the target, as if drawing an outline around the target’s body.)

Examples:

1. 昨天晚上玩《绝地求生》，我跟朋友都成了人体描边大师，全程没打中一个人。(Last night while playing PUBG, my friend and I became Human Outline Masters, not hitting a single person the entire time.)
2. 他在《彩虹六号》里开火直接绕着敌人打了一圈，果然是人体描边大师。(When he fired in Rainbow Six Siege, his bullets went completely around the enemy, truly proving himself to be a Human Outline Master.)
3. 看完这部电影，我觉得反派的射击水平只有人体描边大师能比得上。(After watching this movie, I think the villains’ shooting skills could only be matched by Human Outline Masters.)

Table 7: Representative examples from the CHIME dataset, illustrating the *experience*, *quotation*, and *stylistic device* meme types.

Homophonic Pun
Meme: 虾仁猪心 (Xia Ren Zhu Xin, literally “shrimp meat, pig heart”) **Meaning**: “虾仁猪心” 是 “杀人诛心” 的谐音梗，用来调侃和形容揭露、指责人的思想和动机。本意是指消灭肉体不如谴责动机，但在网络用语中，多用于游戏场景中，引申为攻击他人时意在击中对方的心理弱点。(A homophonic pun for “Sha Ren Zhu Xin” (killing the person and condemning their heart/intention). It’s used to describe exposing or criticizing someone’s thoughts and motives. The original meaning suggests that condemning one’s motives is more devastating than physical harm. In Internet slang, particularly in gaming contexts, it refers to attacking someone’s psychological vulnerabilities rather than just defeating them.) **Examples**: 1. 玩游戏的时候，他总是虾仁猪心，不仅打败我，还要说些让我气愤的话。(When playing games, he always goes for the psychological kill, not just defeating me but also saying things that make me angry.) 2. 她说话总是虾仁猪心，看来她不仅想赢得比赛，还想让我怀疑自己的能力。(She always speaks in a way that cuts to the core, seemingly not just wanting to win the competition but also making me doubt my abilities.) 3. 你这样做简直虾仁猪心，揭穿我的意图让我无地自容。(What you did was absolutely brutal to my psyche, exposing my intentions and leaving me utterly embarrassed.)
Slang
Meme: 小老弟 (Little Brother) **Meaning**: “小老弟” 是一种对比自己年轻或经验浅的男性的亲切称呼，现指比自己实力弱、缺乏经验的人。(An affectionate term for males who are younger or less experienced than oneself. It now refers to someone who has less ability or experience than you.) **Examples**: 1. 看到他在那里手忙脚乱，我忍不住说：“怎么回事小老弟？” (Seeing him all flustered there, I couldn’t help but say: “What’s going on, little brother?”) 2. 当他第一次玩这个游戏的时候，老玩家们都调侃他：“继续加油，小老弟。” (When he played this game for the first time, the veteran players teased him: “Keep it up, little brother.”) 3. 这项目果然难度很大，几个小老弟还得多练习一下。(This project is indeed quite challenging; these little brothers still need more practice.)
Abbreviation
Meme: 请允悲 (Please Allow Sadness) **Meaning**: “请允悲” 是 “请允许我做一个悲伤的表情” 的缩写，常用于表达对他人不幸遭遇的调侃。在对方倾诉糗事或不幸经历时，内心觉得搞笑但表面上假装同情，因此说 “请允悲” 以表达这种内心的矛盾与调侃。(An abbreviation for “please allow me to make a sad expression,” commonly used to express mock sympathy for someone else’s misfortune. When someone shares an embarrassing situation or unfortunate experience, you might find it amusing internally while pretending to be sympathetic externally. Saying “please allow sadness” expresses this inner contradiction and gentle mockery.) **Examples**: 1. 朋友跟我说他早上出门急得穿错鞋，我忍不住想笑，但还是回复了一句：“请允悲。” (My friend told me he was in such a rush this morning that he wore mismatched shoes. I could barely hold back my laughter but still replied: “Please allow sadness.”) 2. 他居然因为贪吃掉进了水坑，看他湿透的样子，我只好假装同情地说：“请允悲。” (He actually fell into a puddle because he was distracted by food. Seeing him all soaked, I had to feign sympathy and say: “Please allow sadness.”) 3. 刚听完她怎么被狗追得掉水沟的故事，我差点笑出声，只能一本正经地说：“请允悲。” (After hearing her story about being chased by a dog and falling into a ditch, I almost burst out laughing, but managed to say with a straight face: “Please allow sadness.”)

Table 8: Representative examples from the CHIME dataset, illustrating the *homophonic pun*, *slang*, and *abbreviation* meme types.

句。*(In the context of the Chinese Internet, explain the meaning and origin of the following viral meme, and create one example sentence.)*

C.2 One-Shot Prompts

We also experimented with one-shot prompts, where we provided the model with an example of a meme and its explanation, (possibly) origin, and 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。
示例：
技术宅拯救世界
含义：指技术宅能够通过自身强大的动手和创造能力，解决各种实际问题，甚至能承担起拯救世界的重任。
例句：在电影里，当病毒席卷全球，终究是技术宅拯救世界，用代码解开谜团。
(English translation)
In the context of the Chinese Internet, explain the meaning of the following viral meme and create one example sentence.
Example:
Tech Geeks Save the World
Meaning: Refers to how tech enthusiasts can solve various practical problems through their strong hands-on and creative abilities, and even take on the responsibility of saving the world.
Example sentence: In the movie, when the virus swept across the globe, it was ultimately the tech geeks who saved the world by decoding the mystery with their programming skills.

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。
示例：
水灵灵
含义：形容一种年轻、有活力的状态。
出处：源自一位韩国女子组合成员在采访中的发言，她在展示合照封面时说自己"水灵灵地在中间"。
例句：水灵灵地挤个地铁，每天都充满活力。
(English translation)
In the context of the Chinese Internet, explain the meaning and origin of the following viral meme, and create one example sentence.
Example:
Fresh and Dewy
Meaning: Describes a youthful, energetic state or condition.
Origin: Originated from a statement made by a Korean girl group member during an interview, where she described herself as "fresh and dewy in the middle" when showing a group photo cover.
Example sentence: Taking the subway with a fresh and dewy attitude, filled with vitality every day.

Our analysis revealed that one-shot prompts did not significantly improve model performance, but greatly diminished it on the meaning explanation task, compared to zero-shot prompts, as demonstrated in Table 9. We hypothesize that this performance degradation stems from the inherent nature of meme interpretation, which demands flexible analysis rather than rigid pattern matching or format adherence. Consequently, we focused exclusively on zero-shot prompting results in the main text.

C.3 More Automatic Evaluation Results

Table 10 gives the exact meaning scores of the candidate models for each of the six meme types.

C.4 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 significant social influence. To study the understanding of Chinese Internet memes by large language models, this project aims

Model	Cosine Similarity		BERTScore (F)		BARTScore (F)	
	Meaning	**Origin**	**Meaning**	**Origin**	**Meaning**	**Origin**
GPT-4o						
Zero-Shot	0.815	0.647	0.800	0.675	−4.485	−4.717
One-Shot	0.825 ↑	0.652 ↑	0.805 ↑	0.717 ↑	−4.426 ↑	−4.565 ↑
Claude 3.5 Sonnet						
Zero-Shot	0.788	0.625	0.789	0.696	−4.611	−4.695
One-Shot	0.736 ↓	0.660 ↑	0.761 ↓	0.719 ↑	−4.630 ↓	−4.750 ↓
GLM-4-9B						
Zero-Shot	0.813	0.578	0.797	0.663	−4.453	−4.560
One-Shot	0.750 ↓	0.549 ↓	0.746 ↓	0.607 ↓	−4.470 ↓	−4.656 ↓
GLM-4-Plus						
Zero-Shot	0.844	0.679	0.822	0.737	−4.291	−4.441
One-Shot	0.797 ↓	0.689 ↑	0.796 ↓	0.743 ↑	−4.283 ↑	−4.468 ↓
Qwen2.5-7B						
Zero-Shot	0.792	0.605	0.782	0.661	−4.494	−4.779
One-Shot	0.731 ↓	0.639 ↑	0.731 ↓	0.693 ↑	−4.573 ↓	−4.677 ↑
Qwen2.5-72B						
Zero-Shot	0.819	0.627	0.803	0.690	−4.366	−4.605
One-Shot	0.799 ↓	0.626 ↓	0.789 ↓	0.697 ↑	−4.370 ↓	−4.498 ↑
DeepSeek-V3						
Zero-Shot	0.779	0.709	0.774	0.751	−4.331	−4.344
One-Shot	0.746 ↓	0.689 ↓	0.754 ↓	0.722 ↓	−4.380 ↓	−4.539 ↓

Table 9: Comparative analysis of average cosine similarity, BERTScore, and BARTScore across six meme types for all candidate models, contrasting zero-shot and one-shot prompting approaches. ↑ indicates superior performance, and ↓ denotes inferior performance. Results were derived from a balanced sample of 240 memes, comprising 40 from each meme type.

Model	Experience			Quotation		
	Cos. Sim.	BERTS.	BARTS.	Cos. Sim.	BERTS.	BARTS.
GPT-4o	0.837	0.812	−4.261	0.756	0.755	−4.354
Claude 3.5 Sonnet	0.809	0.799	−4.414	0.707	0.733	−4.594
GLM-4-9B	0.818	0.804	−4.236	0.740	0.750	−4.285
GLM-4-Plus	**0.846**	**0.822**	**−4.150**	**0.812**	**0.790**	**−4.199**
Qwen2.5-7B	0.807	0.786	−4.337	0.731	0.730	−4.447
Qwen2.5-72B	0.832	0.807	−4.220	0.763	0.757	−4.294
DeepSeek-V3	0.802	0.796	−4.203	0.742	0.742	−4.306
Model	**Stylistic Device**			**Homophonic Pun**		
	Cos. Sim.	**BERTS.**	**BARTS.**	**Cos. Sim.**	**BERTS.**	**BARTS.**
GPT-4o	0.811	0.792	−4.365	0.797	0.789	−4.751
Claude 3.5 Sonnet	0.790	0.782	−4.499	0.781	0.784	−5.101
GLM-4-9B	0.805	0.791	−4.303	0.799	0.790	−4.741
GLM-4-Plus	**0.827**	**0.804**	−4.248	**0.826**	**0.808**	**−4.589**
Qwen2.5-7B	0.791	0.771	−4.387	0.762	0.761	−4.875
Qwen2.5-72B	0.813	0.793	−4.312	0.794	0.788	−4.746
DeepSeek-V3	0.802	0.794	**−4.245**	0.818	0.805	−4.610
Model	**Slang**			**Abbreviation**		
	Cos. Sim.	**BERTS.**	**BARTS.**	**Cos. Sim.**	**BERTS.**	**BARTS.**
GPT-4o	0.835	0.809	−4.388	0.830	0.819	−4.612
Claude 3.5 Sonnet	0.802	0.791	−4.531	0.820	0.815	−4.736
GLM-4-9B	0.827	0.807	**−4.253**	0.845	0.826	−4.607
GLM-4-Plus	0.837	0.812	−4.332	**0.865**	**0.839**	−4.479
Qwen2.5-7B	0.814	0.792	−4.415	0.792	0.786	−4.863
Qwen2.5-72B	**0.837**	**0.818**	−4.290	0.830	0.816	−4.636
DeepSeek-V3	0.800	0.798	−4.272	0.840	0.824	**−4.467**

Table 10: 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**.

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. Of the 14 raters, 9 annotated three batches, 4 annotated two batches, and 1 annotated a single batch. The average number of batches per rater was 2.57, with a median of 3.

C.5 More Human Evaluation Results

Table 11 gives the Fleiss' kappa scores on each of the 12 evaluation batches. Based on the kappa scores, we observe that *abbreviation* memes show highest agreement ($\kappa \approx 0.71$ to 0.74) due to their straightforward nature; *slang* memes show lowest agreement ($\kappa \approx 0.27$ to 0.28) because cultural familiarity varies among annotators; *homophonic puns* show moderate disagreement ($\kappa \approx 0.40$ to 0.41) due to subjective interpretation of wordplay effectiveness. We conjecture that cultural context dependency is the primary driver of annotation disagreement—memes requiring deeper cultural knowledge (slang, stylistic devices) are harder to evaluate consistently than structurally-defined ones (abbreviations). Table 12 provides the detailed human evaluation results on the meaning task for each of the six meme types.

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

C.6 Error Analysis with Illustrative Case Studies

To further investigate the performance of LLMs, we conducted a qualitative error analysis on the generated meanings and origins. Specifically, we have identified three common types of errors in the generated meanings and origins, which are as follows:

1. **Origin confusion**: Models frequently attributed memes to incorrect sources, particularly with quotation memes. In many instances, LLMs provided vague attributions (e.g., "originating from social media") rather than specific origins. For example, 再爱就不礼貌了 (Any More Love Would Be Impolite) originated when a Japanese short video blogger made it into a Japanese language teaching video, which Internet users then parodied into a "fake Japanese version," creating a comedic atmosphere. However, all models except Qwen2.5-7B and DeepSeek-V3 have provided vague origins, such as "the exact origin of this meme is unclear, but it gradually gained popularity in social media and everyday online communication" (by GLM-4-Plus), while Qwen2.5-7B and DeepSeek-V3 provided completely incorrect origins.

2. **Semantic shift**: For most misinterpreted homophonic pun memes, models explained related concepts with similar phonetics rather than capturing the actual meme meaning. In other cases, models failed to recognize the phonetic wordplay entirely and simply explained the literal meaning. For example, 肾炎 (Shen Yan, literally "nephritis") is a homophonic pun on "divine face/godly appearance" in Chinese, used to mock fans who exaggeratedly describe their idols as having "godly looks." However, Claude 3.5 Sonnet misinterpreted it as another homophonic word 神言, which means "divine words" (words of god/-godlike statement), while GLM-4-Plus simply explained its literal meaning, interpreting it as a kidney disease.

3. **Cross-type confusion**: Abbreviation memes were occasionally misinterpreted as homophonic puns, indicating difficulty in distinguishing between these distinct linguistic mechanisms. For example, 人干事 (Human Doing Things) is an abbreviation of 这是人干的事吗 (Is This Something A Human Would Do?), mainly used to criticize unreasonable or unacceptable things. However, Qwen2.5-7B recognized it as a homophonic pun, 人设 (Persona), and explained "used to mock people whose online image is inconsistent with their actual behavior."

Table 13 gives the complete model outputs for the above three error types.

D MCQ Task

D.1 MCQ Prompt without Meaning

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

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 12: 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**.

(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
(2) yyds (similar to GOAT in English)
(3) Go Die
(4) Solid Evidence
(5) Strong Sense of Stealing
Answer: 2

D.2 MCQ Prompt with Meaning

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

(1) Origin Confusion
Meme: 再爱就不礼貌了 (Any More Love Would Be Impolite) **Meaning**: "再爱就不礼貌了" 原本是一段告别文案，用于表达感情的克制和终止。后来被演变成一种网络梗，形成 "再 XX 就不礼貌了" 的格式，用于幽默地表示 "给我到此为止" 的意思。(Originally a farewell message expressing emotional restraint and termination. It later evolved into an internet meme with the format "Any more [X] would be impolite," humorously meaning "that's enough" or "let's stop here.") **Origin**: 该梗起源于日语短视频博主将其制作成日语教学视频，并被网友用恶搞方式演绎成 "塑料日语版"，营造出一种搞笑的氛围。(Originated when a Japanese short video blogger created a Japanese language teaching video featuring this phrase. Internet users then parodied it into a "fake Japanese version," creating a humorous atmosphere.)
Origin by GLM-4-Plus (vague origin): 这个梗的具体出处不太明确，但它在社交媒体和网友的日常交流中逐渐流行起来。可能是源于某个具体的网络事件或段子，后来因其幽默和贴切的表意而被广泛传播和使用。(The exact origin of this meme is unclear, but it gradually gained popularity in social media and everyday online communication. It likely originated from a specific Internet event or joke, and later became widely spread and used due to its humor and apt expression of meaning.) **Origin by Qwen2.5-7B** (incorrect origin): 这句话源于 2018 年 9 月 17 日，演员王一博在《天天向上》节目中的发言。(This phrase originated from actor Wang Yibo's statement on the TV show "Day Day Up" on September 17, 2018.) **Origin by DeepSeek-V3** (incorrect origin): 电视剧《三十而已》中的角色许幻山对林有有所说的话。(A line spoken by the character Xu Huanshan to Lin Youyou in the TV drama "Nothing But Thirty.")
(2) Semantic Shift
Meme: 肾炎 (Shen Yan, literally "nephritis") **Meaning**: "肾炎" 是 "神颜" 的谐音梗，用于嘲讽粉丝形容偶像为 "神颜" 的夸张说法，认为这种夸赞往往带有厚厚的粉丝滤镜，不被大多数人认同。(A homophonic pun for "divine face/godly appearance" in Chinese, used to mock fans who exaggeratedly describe their idols as having "divine beauty." This mockery suggests that such praise often comes with heavy fan bias and isn't generally agreed upon by most people.)
Meaning by Claude 3.5 Sonnet (similar phonetics): "肾炎" 是一个网络谐音梗，源自 "神言"（神的发言/神一般的发言）的谐音。用来形容某人说出了特别有见地、特别精辟或者特别搞笑的话。这个词通常带有一定的赞赏意味。这个梗经常出现在弹幕、评论区等网络交流场合，是比较轻松诙谐的网络用语。(An Internet homophonic pun derived from "divine words" (words of god/godlike statements). It's used to describe when someone says something particularly insightful, profound, or hilarious. This term usually carries a sense of appreciation. This meme frequently appears in bullet comments, comment sections, and other online communication spaces, and is a relatively light-hearted and humorous internet expression.) **Meaning by GLM-4-Plus** (literal meaning): 在中文互联网语境下，"肾炎" 通常指的是一种肾脏疾病，涉及肾脏的炎症反应，可能导致肾功能减退。肾炎可以是急性的或慢性的，病因多样，包括感染、自身免疫疾病、药物反应等。常见的症状包括尿液变化（如泡沫尿、血尿）、水肿、高血压、疲劳、背痛、恶心呕吐等。(In the Chinese Internet context, "nephritis" typically refers to a kidney disease involving inflammatory reactions in the kidneys, which may lead to reduced kidney function. Nephritis can be acute or chronic, with various causes including infections, autoimmune diseases, drug reactions, etc. Common symptoms include changes in urine (such as foamy urine, blood in urine), edema, high blood pressure, fatigue, back pain, nausea, and vomiting.)
(3) Cross-Type Confusion
Meme: 人干事 (Human Doing Things) **Meaning**: "人干事" 源自 "这是人干的事吗" 的缩写，主要用来吐槽不合理或者让人无法接受的事物，多用于表达对某视频内容的不满或出入预期的情况。(Originates from the abbreviation of "Is this something a human would do?" It's mainly used to criticize unreasonable or unacceptable matters, often expressing dissatisfaction with video content or situations that deviate from expectations.)
Meaning by Qwen2.5-7B (incorrect type): "人干事" 是 "人设" 的谐音梗，用于讽刺那些在网上塑造的形象和实际行为不一致的人。(A homophonic pun for "persona" in Chinese, used to mock people whose online image is inconsistent with their actual behavior.)

Table 13: Examples of model outputs for three common error types: (1) Origin confusion, (2) Semantic shift, and (3) Cross-type confusion.

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