MemeBridge: A Dataset for Benchmarking and Mitigating the Bidirectional Cultural Gap in Meme Interpretation

Anonymous ACL submission

Abstract

Communicating with people of different cultures is a complex challenge. Memes, as a prevalent form of online communication, can lead to misunderstandings when used improperly in communication. Large language models (LLMs) can potentially help; however, there is a notable lack of meme datasets that provide context-based explanations and potential misunderstandings for training and evaluating LLMs. To address this gap, we introduce a carefully curated meme dataset MEMEBRIDGE. The accuracy of the dataset was manually examined and quantitative evaluations were performed. Initial probing of various LLMs developed by teams with different cultural backgrounds revealed they have a certain level of cross-cultural understanding and the ability to recognize cultural differences, despite some limitations in meme comprehension. Besides, fine-tuning these LLMs with our dataset led to performance improvements, underscoring the importance of context-rich datasets in enhancing the cultural understanding capacity of LLMs.

1 Introduction

Memes serve as a form of speech act in digital communication, enabling Internet users to engage in social interactions through shared cultural references and semiotic cues (Grundlingh, 2018). However, memes are more than just visual humor; they function as cultural artifacts that reflect societal trends, linguistic variations, and generational differences. Their interpretation is often deeply rooted in a cultural context, making them susceptible to misinterpretation by individuals from different backgrounds (Mukhtar et al., 2024). For example, when Chinese individuals attempt to interpret memes originating in the United States, significant gaps can be seen with respect to humor, historical references, and societal norms. To investigate this issue, we conducted informal interviews with eight Chinese individuals currently residing in the United States. Many participants shared concerns about the potential for miscommunication when using memes, as illustrated by the following quote.

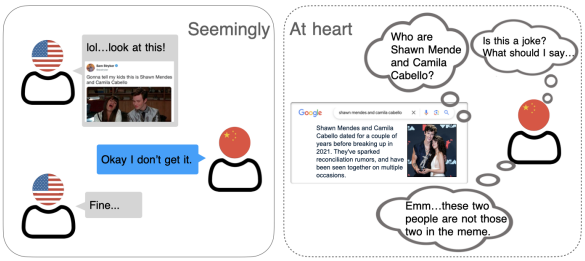


Figure 1: This meme humorously distorts history, exaggerating Mendes and Cabello's relationship. A Chinese individual unfamiliar with the 'Gonna Tell My Kids' meme format or U.S. pop culture may find it confusing, leading to awkward interactions.

> *"... Honestly, sometimes I worry about using memes incorrectly and accidentally causing awkwardness or conflicts with my (American) friends...."*

Moreover, as global social media platforms continue to expand, this problem could become increasingly common, even among Chinese people not living in the United States. Following our interview, although Chinese individuals typically use platforms within their own cultural group, such as WeChat and Weibo, many of them also engage with global platforms like Twitter and Facebook, thereby being exposed to other cultures as well (Tsai and Men, 2017). A lack of cultural familiarity can lead to unintended misunderstandings or misaligned social interactions, as in Figure 1. Addressing these gaps is crucial for improving cross-cultural digital communication and mitigating the risk of misinterpretation in online discourse.

As large language models (LLMs) continue to advance, multimodal variants such as GPT-4o (Hurst et al., 2024), Llama-3.2-Vision (Dubey et al., 2024), GLM-4V (Team et al., 2024), and Qwen-VL (Bai et al., 2023) have become increas-

ingly accessible to the general public. Given their ability to process and generate both textual and visual content, multimodal LLMs present a potential solution to bridge cultural gaps in communication, including the interpretation of memes. However, previous research on cultural knowledge bases (Shi et al., 2024) has demonstrated that LLMs predominantly reflect Western-centric perspectives, making it challenging for non-Western audiences to fully understand culturally embedded content. Additionally, biases in training data and limitations in understanding the relationship between text and image can lead LLMs to generate skewed or inaccurate meme explanations (Zhong and Baghel, 2024). Despite the increasing sophistication of LLMs, these limitations raise concerns about whether they can effectively interpret and contextualize memes across cultures and help people understand memes from different countries, necessitating further investigation into their performance in cross-cultural meme understanding.

In this paper, we investigate the ability of state-of-the-art LLMs to interpret U.S.-based memes, focusing on their capacity to provide explanations, detect sentiment, and identify emotions. To facilitate this study, we constructed MEMEBRIDGE, a carefully curated dataset consisting of memes contributed by native U.S. participants. Each meme entry includes explanations, potential misunderstandings that individuals from different cultural backgrounds might experience, and sentiment and emotion annotations. Using this dataset, we evaluate and fine-tune multiple LLMs to assess their effectiveness in cross-cultural meme interpretation.

Through extensive experiments, we have the following key observations: (1) The cultural gap in meme interpretation is bidirectional. Chinese individuals face challenges in understanding U.S. memes, with 58.8% accuracy in determining their explanations, 45% accuracy in labeling sentiment and 48.9% accuracy in labeling emotions. Similarly, U.S. participants struggled to accurately predict how Chinese individuals misinterpret memes, as the misunderstandings proposed by U.S. participants often did not align with actual misconceptions held by Chinese participants. (2) The origin of an LLM—whether developed by a Chinese or U.S. company—does not inherently lead to significant cultural biases. Instead, these models were well-aligned to minimize explicit cultural tendencies. (3) LLMs demonstrate cultural awareness and adaptability. Explicitly instructing LLMs to adopt a specific cultural perspective significantly impacts their interpretative performance. All tested models exhibited performance differences when role-playing as U.S. participants or role-playing as Chinese participants, compared to the default setting. This suggests their ability to adjust to different cultural identities and perspectives.

2 Related Work

Cultural Awareness in LLMs. The popularity and adoption of LLMs in various domains pose challenges and the need for cultural awareness (Pawar et al., 2024; Ramezani and Xu, 2023). Existing studies have found that LLMs have biases in understanding cultural symbols, have different performances for different regional cultures, and are difficult to reach human levels (Yao et al., 2024). For example, Shi et al. (2024) have demonstrated that LLMs predominantly reflect Western-centric perspectives, making it challenging for non-Western audiences to fully understand culturally embedded content. To address this, a growing number of studies (Shi et al., 2024; Zhao et al., 2024; Li et al., 2024) have explored various aspects of integrating cultural understanding into LLMs, with the aim of bridging communication gaps and facilitating effective cross-cultural exchange. For example, Nguyen et al. (2023) has proposed Candle, an end-to-end approach to extracting cultural common sense knowledge from Web corpora on a large scale. Although these studies offer valuable information, many of them focused on machine translation with text information. More recently, the concept of image transcreation for cultural relevance acknowledges the need to adapt visual content for cultural appropriateness (Khanuja et al., 2024), representing a crucial step towards bridging the gap between visual language understanding and cultural interpretation.

Cross-Cultural Understanding with Multimodal LLMs. Some recent works on multimodal LLMs highlight the challenges of adapting multimodal reasoning across diverse linguistic and cultural contexts. One major focus has been cultural adaptation in multimodal tasks, where researchers explore how models interpret visual and textual information differently across cultures, emphasizing the need for datasets that reflect such diversity (Liu et al., 2021; Li and Zhang, 2023). Another key area is culturally influenced language inference, examining how cultural norms shape reasoning, particularly in tasks like natural language inference

and figurative language understanding (Huang and Yang, 2023; Kabra et al., 2023). Additionally, work on humor, satire, and harmful content detection demonstrates the necessity of culturally aware AI, as humor and hate speech often rely on nuanced cultural context (Nandy et al., 2024; Bui et al., 2024). Collectively, these studies stress the importance of integrating cultural awareness into vision-language models to enhance their robustness and fairness in global applications. These studies inspired us to investigate the cross-cultural understanding of memes, the dynamic and informal media circulating in online communities.

Cross-Cultural Understanding and Evaluation of Memes. Several datasets have been collected to facilitate the understanding of memes. For example, Zannettou et al. (2018) collected and analyzed 160 million images from four major online communities (i.e., Twitter, Reddit, 4chan's /pol/, and Gab), establishing a methodological framework for cross-platform meme tracking and analysis. FigMemes (Liu et al., 2022) focuses on the identification of figurative language in political memes. MCC (Sharma et al., 2023) contains 3,400 memes and their contexts focusing on detecting explanatory evidence for memes. MemeCap (Hwang and Shwartz, 2023) enables the evaluation of visual language models in the meme captioning task. SemanticMemes (Zhou et al., 2024) highlights semantic clustering. MemeMQA (Agarwal et al., 2024) offers a multimodal question-answering framework for a better semantic explanation of memes. Multi3hate (Bui et al., 2024) is designed for specific tasks such as context understanding and hate speech detection. There studies enhanced the understanding and interpretation of memes but did not adequately address their understanding from a cross-cultural perspective. Our work collects data through crowdsourcing and requires participants to provide an explanation and possible misconceptions of each meme, as well as sentiment and emotion tags. Furthermore, we propose a novel cross-cultural evaluation design by prompting LLMs as people of different cultural backgrounds, enabling a more direct and quantitative assessment of cross-cultural misunderstandings.

3 MEMEBRIDGE Dataset Construction

To ensure high-quality data for cross-cultural meme understanding, we designed a three-stage crowdsourcing pipeline for dataset collection, validation, and cross-cultural testing. This structured process aims to systematically refine and validate the dataset while identifying cultural misunderstandings embedded in memes.

3.1 Stage 1: Initial Data Collection

The dataset construction process began with collecting a diverse set of memes and their interpretations from a U.S. crowd group consisting of 100 participants, recruited through Prolific. Each participant was asked to contribute 10 memes along with their personal *explanations* and *potential misunderstandings* they believed could arise for individuals from other cultural backgrounds, yielding a total of 1,000 data points. In addition to these textual inputs, participants were asked to assign each submitted meme a sentiment label from {positive, negative, neutral} and one or more emotion labels from {sarcastic, humorous, offensive, motivational} (Sharma et al., 2020). This approach ensures that the dataset captures not only the explicit meaning of memes but also the emotions and cultural context associated with them. Our crowdsourcing method aligns with prior efforts, such as Yin et al. (2022), which leveraged diverse participant contributions to collect geo-diverse commonsense knowledge.

To enhance data quality, we randomly selected 200 data points and had three researchers label them, using majority voting to determine their quality. This labeled dataset was then used to train a BERT-based classifier (Devlin et al., 2019) to filter out low-quality meme interpretations. Next, we conducted a linguistic complexity check, revealing that explanations contained an average of 26.12 words, while misunderstandings averaged 20.89 words. A Type-Token Ratio (TTR) analysis (Richards, 1987) further showed that explanations had an average TTR of 0.905, whereas misunderstandings had a slightly higher average of 0.928. To ensure high linguistic quality, we filtered out low-diversity data by computing the average TTR of each explanation and misunderstanding. Data points with an average TTR below 0.5 were discarded, retaining only those with sufficient lexical diversity. After applying these filtering steps, 754 data points met the quality criteria and were retained for further analysis.

Following these analyses, we used the GPT-4 API (Achiam et al., 2023) to rewrite the original data, standardize the format, and improve grammatical accuracy while preserving semantic integrity. To ensure consistency, we applied a similarity scor-

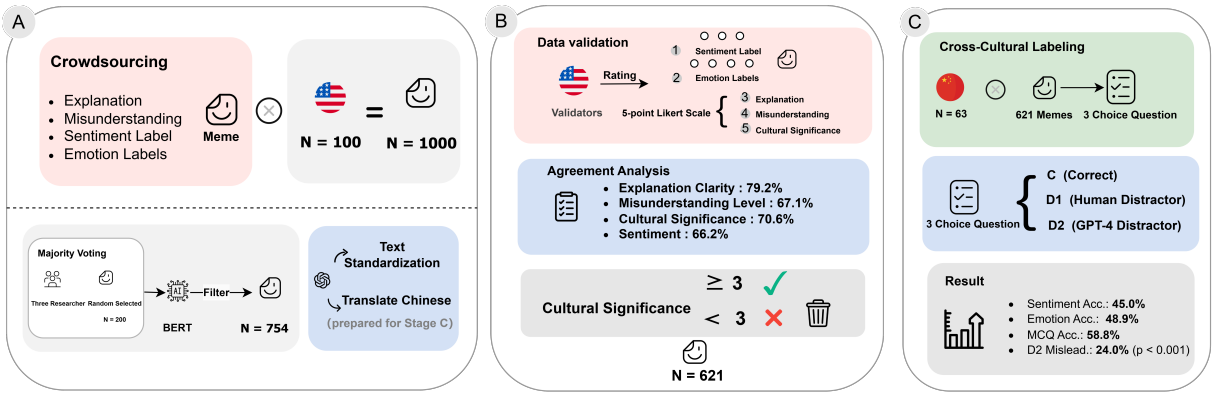


Figure 2: (A) Data Collection & Cleaning: 200 memes randomly selected from 1000 collected and labeled by three researchers, followed by BERT-based refinement (N=754). (B) Data Validation: Participants labeled memes for sentiment and emotion. Explanation text, misunderstandings, and cultural significance were rated on a five-point Likert scale. Resulting in the final meme dataset (N=621) (C) Cross-Cultural Labeling: Chinese participants participated in meme interpretation tests.

ing mechanism to compare the refined text with the original. Details of our prompt design and continuous monitoring of similarity scores to ensure successful rewriting are provided in the Appendix B. Finally, we also leveraged GPT-4 to translate the dataset into Chinese, preparing it for comparative cross-cultural evaluation in Stage 3.

3.2 Stage 2: Data Validation

To ensure robustness and reliability, we conducted a validation process involving another group of 180 U.S. participants. However, some participants left early or failed the attention check. To maintain consistency, we ensured that each meme was reviewed and annotated by exactly four participants. This phase focused on measuring consistency and agreement among annotators to assess the quality of the collected explanations and misunderstandings. Participants were asked to assign sentiment and emotion labels to the memes to gauge consensus and alignment in interpretations. Moreover, the explanatory text, identified misunderstandings and cultural significance were evaluated using a five-point Likert scale, allowing us to quantify recognition and ensure the cultural relevance of our data.

To assess agreement levels, we computed percent agreement for each meme across multiple dimensions: explanation clarity, misunderstanding level, cultural significance (all mapped to a three-level scale derived from the five-point Likert ratings), sentiment, and emotion. Agreement was determined by identifying the modal rating among the four annotators and calculating the proportion of annotators who assigned the same rating. Our results showed agreement rates of 79.2% for explanation clarity, 67.1% for misunderstanding level, 70.6% for cultural significance, 66.2% for sentiment, and between 75% and 90% for the four emotion labels. Based on these results, we filtered out memes with an aggregated cultural significance rating below 3 (on a five-point Likert scale, meaning that most annotators did not perceive them as culturally significant in the U.S. context). After filtering, 621 memes remained in the final dataset, and we assigned the aggregated sentiment and emotion labels to them for further analysis. The distribution of sentiment and emotion labels, shown in Table 1, demonstrates a well-balanced representation of the collected memes.

Sentiment		
Positive	Neutral	Negative
185	310	126

Emotions			
Sarcastic	Humorous	Motivational	Offensive
439	539	586	594

Table 1: Distribution of sentiment and emotion labels.

3.3 Stage 3: Cross-Cultural Assessment

The final stage aimed to evaluate cross-cultural differences in meme interpretation by engaging 84 Chinese participants. After filtering out those who left early or failed the attention check, 63 participants remained, ensuring that each meme was reviewed by two individuals, resulting in 1,242 data points (621 memes$\times$2 reviews/meme). These participants provided sentiment and emotion labels, allowing us to compare their perceptions with those of the U.S. participants and identify potential cul-

tural divergences.

Additionally, participants were asked to complete multiple-choice questions constructed in the following way: *Explanations* were designated as the correct answer C, while *potential misunderstandings* served as one distractor D_1. To introduce further variation, we employed GPT-4 to generate an additional distractor D_2 by providing the meme as input. This resulted in a three-choice question format $\{C, D_1, D_2\}$ for each meme.

Our assessment results demonstrated that Chinese participants struggled to accurately interpret sentiment in U.S.-centric memes, achieving only 45.0% accuracy. Similarly, their ability to correctly identify emotions was limited, with an accuracy of 48.9%. Most notably, their performance on the multiple-choice task was relatively low, with a correctness rate of just 58.8%. As discussed in the introduction, these findings reinforce the existence of a cultural gap affecting meme comprehension. For the multiple-choice task, 17.1% of incorrect answers were attributed to D_1, while 24.0% were attributed to D_2, indicating that Chinese participants were significantly more misled by the LLM-generated distractor than the human-assumed distractor ($p < 0.001$). This supports Hypothesis 1: the cultural gap is bidirectional—just as Chinese participants struggle to interpret U.S. memes, U.S. participants may also have difficulty predicting how others will perceive their memes.

This stage provided a systematic assessment of cross-cultural misinterpretations and quantitative insights into the challenges non-U.S. audiences face in understanding American memes.

4 Evaluating LLM's Cross-Cultural Meme Understanding

With the dataset constructed, we aim to evaluate the performance of LLMs in meme interpretation, focusing specifically on four off-the-shelf multimodal LLMs: Qwen2.5-VL-3B (Bai et al., 2023), GLM-4V (Team et al., 2024), Llama-3.2-11B-Vision (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024)[1]. Our goal is to assess the models' ability to generate human-like interpretations, accurately detect the sentiment and emotions conveyed by memes, and evaluate their adaptability to different cultural perspectives.

[1]In the following discussion, we abbreviate these models to Qwen, GLM, LLaMA, and GPT for the ease of notation.

	Qwen	GLM	LLaMA	GPT	CN (Human)
MCQ	68.8%	52.8%	55.2%	75.4%	58.8%
Sent	40.4%	37.7%	35.3%	54.2%	45.0%
Emo	65.1%	64.2%	32.4%	84.3%	48.9%

Table 2: Comparing the performance of different LLMs with Chinese participants on Meme Understanding. Underlined values indicate statistically significant differences from Chinese annotators ($p < 0.05$).

4.1 Assessment on LLMs

First, we evaluated LLMs under the same test conditions as Chinese participants in Section 3.3. Our findings indicate that while GPT consistently outperforms Chinese participants in interpreting U.S. memes, the other models exhibit specific weaknesses, as shown in Table 2. Notably, for sentiment classification, Qwen, GLM, and LLaMA all performed worse than Chinese participants, indicating that recognizing sentiment is inherently subtle and remains an open problem (Vanshika et al., 2024).

In contrast, for emotion detection, Qwen and GLM significantly outperformed Chinese participants, indicating the potential of these models to assist non-native speakers in understanding emotions conveyed in U.S. memes. However, LLaMA performed consistently worse than both the other models and human participants. This could be attributed to the differences in dataset curation—while Qwen, GLM, and GPT are developed by companies with proprietary, curated training data, LLaMA is trained predominantly on open-source datasets, which may lack diversity, fine-grained annotations, or up-to-date meme-related content. Consequently, its performance on specialized tasks such as sentiment and emotion detection is notably lower.

Further analysis of multiple-choice answers by different models revealed an interesting pattern: LLMs, similar to Chinese participants, tend to prefer LLM-generated distractors over human-assumed distractors, as shown in Figure 3. This observation suggests that while LLMs can effectively understand U.S. memes, their errors align with 'real' misunderstandings experienced by Chinese participants. This finding underscores the potential of LLMs in modeling cultural misinterpretations and highlights the dual nature of the cultural gap—both in interpreting and anticipating meme misunderstandings across cultures.

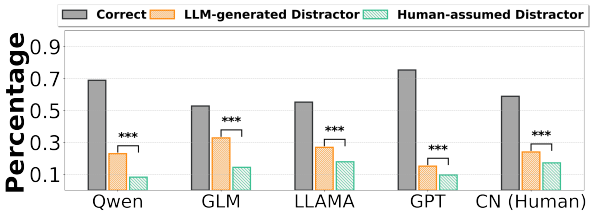


Figure 3: Comparing the distribution of answer choices across different LLMs and Chinese participants (CN (Human)).Significance indicators (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) above pairs of distractor bars show whether participants significantly favored one type of distractor over the other when making incorrect choices.

4.2 Detection of LLMs' Potential Bias

To evaluate cross-cultural adaptation and cultural awareness, we designed experiments to identify potential biases in model training that may result in cultural tendencies. We selected four models for comparison: Qwen2.5-VL and GLM-4V, Chinese developed open-source models; Llama-3.2-Vision, a U.S.-based open-source model; and GPT-4o, a widely regarded state-of-the-art closed-source model. Each model was tested under three conditions: the default setting DEF , an explicit prompt instructing the model to respond as a native US person US-RolePlaying , and another instructing it to respond as a native Chinese CN-RolePlaying . Under each condition, the model first completed the same test as in Section 3.3, and we measured their performance on those tasks. Then, in a fresh session, the models were instructed to generate an explanation for this meme. We compared these explanations with both the original (crowdworker-provided) explanations and the formatted (LLM-rewritten) explanations (both obtained in Section 3.1), using cosine similarity scores to quantify textual alignment. The test results are presented in Table 3. Across all models, performance was highest under the US-RP condition. Additionally, LLM-generated explanations exhibited higher similarity to the formatted explanations than to the original ones. This aligns with expectations, as LLM outputs tend to be more structured and formal, whereas crowdworker-written explanations exhibit more variability in grammar and vocabulary.

To further quantify model performance, we introduce a *Performance Score* (PS) to measure the overall performance of the model across all tasks, as shown in Table 3. The score was computed by grouping the two similarity comparisons into a single task, along with three classification tasks: multiple choice questions selection (MCQ), sentiment labeling (Sent), emotions labeling (Emo):

$$\begin{aligned} \text{PS} = &(\text{Sim}_{\text{original}} + \text{Sim}_{\text{formatted}}) \\ &+ \text{PS}_{\text{MCQ}} + \text{PS}_{\text{Sent}} + \text{PS}_{\text{Emo}} \end{aligned} \quad (1)$$

Since similarities can inherently serve as score metrics, and by observation, the sum of the two similarities approximates 1 for each model in each mode, we would like to achieve a similar expected score of $\mathbb{E}[\text{PS}_{\text{MCQ}}] = \mathbb{E}[\text{PS}_{\text{Sent}}] = \mathbb{E}[\text{PS}_{\text{Emo}}] = 1$ to each of the remaining three tasks. To achieve this, we first estimated the expected accuracy for each task. For multiple-choice question answering, where each question has three answer choices, the expected accuracy is $\mathbb{E}_{\text{MCQ}}[\text{Acc}] \approx 0.33$. Similarly, for sentiment labeling: $\mathbb{E}_{\text{Sent}}[\text{Acc}] \approx 0.33$. For the emotion labeling task, we define a model's submitted label set as correct if the ground truth labels form a subset of the predicted labels. We can compute the expected accuracy of this task as:

$$\mathbb{E}_{\text{Emo}}[\text{Acc}] = \left(\sum_{m \in \mathcal{M}} \frac{\text{PCLS}_m}{\text{PLS}} \right) \Big/ |\mathcal{M}| \quad (2)$$

where $\mathcal{M}$ represents the set of all memes, with $|\mathcal{M}| = 621$, PLS stands for the number of possible label sets for each meme, and PCLS stands for the number of possible correct label sets (PCLS). With the expected accuracy of each task determined, we could assign weighted maximum possible possible scores to each task, then compute PS. See details in Appendix A.1.

Across all models, the performance score improved when models were explicitly instructed to adopt either the US-RP or CN-RP perspective, compared to the DEF condition. This finding suggests that role-playing prompts significantly impact LLMs' interpretative accuracy and their alignment with cultural contexts.

Pairwise significance tests (see Table 4) reveal that US-RP consistently improves on DEF in a statistically significant manner in key metrics, and similar improvements are observed when comparing CN-RP to DEF for most metrics. Notably, when directly comparing US-RP and CN-RP , the US-RP condition generally outperforms. Overall, based on PS, the performance ranking for Qwen, LLaMA, and GPT follows the order: US-RP > CN-RP > DEF . These results suggest that for Qwen, LLaMA, and GPT,

	Qwen			GLM			LLaMA			GPT		
	DEF	US-RP	CN-RP	DEF	US-RP	CN-RP	DEF	US-RP	CN-RP	DEF	US-RP	CN-RP
$\text{Sim}_{\text{original}}$	0.484	0.492↑	0.468↓	0.429	0.449↑	0.428↓	0.455	0.467↑	0.442↓	0.460	0.505↑	0.471↑
$\text{Sim}_{\text{formatted}}$	0.582	0.600↑	0.554↓	0.479	0.536↑	0.475↓	0.546	0.566↑	0.536↓	0.499	0.605↑	0.554
Acc_{MCQ}	68.8%	67.9%↓	69.5%↑	52.8%	46.7%↓	52.9%↑	55.2%	47.8%↓	44.8%↓	75.4%	72.7%↓	72.3%↓
Acc_{Sent}	40.4%	46.2%↑	47.0%↑	37.7%	33.2%↓	38.3%↑	35.3%	40.1%↑	39.0%↑	54.2%	55.0%↑	51.9%↓
Acc_{Emo}	65.1%	79.4%↑	74.3%↑	64.2%	67.5%↑	77.2%↑	32.4%	42.2%↑	41.7%↑	84.3%	84.5%↑	82.6%↓
PS	2.887	3.192↑(0.305)	3.049↑(0.162)	2.595	2.663↑(0.068)	2.821↑(0.226)	2.135	2.321↑(0.186)	2.232↑(0.097)	3.242	3.385↑(0.143)	3.245↑(0.003)

Table 3: Models performances across all tasks, including the *Performance Score* (PS).

they retain an underlying cultural awareness, as their DEF performance is closer to the CN-RP mode. They seem to be able to 'show their full power' when required and 'hide' their ability when they are supposed to hide (i.e. when asked to act like Chinese people to interpret US-culture-related objects).

4.3 Fine-tuning

To further validate the effectiveness of our dataset, we conducted fine-tuning experiments. The dataset was split into 70% for fine-tuning, 15% for validation, and 15% for testing. We fine-tuned Qwen, GLM, and GPT, evaluating their performance across the same tasks: semantic similarity check, multiple choice question answering, sentiment labeling, and emotion labeling. Overall, fine-tuning led to performance improvements across most tasks. However, an exception was observed with GPT in emotion classification, where accuracy dropped significantly from 87.1% to 61.1% on the test set. This decline may be attributed to overfitting, as the base GPT model already demonstrated strong performance prior to fine-tuning. Meanwhile, performance improvements were still observed in other tasks where fine-tuned GPT had not originally excelled. A similar trend was noted for Qwen, where its performance in generating explanations and multiple-choice question answering declined slightly after fine-tuning. Notably, this model initially outperformed the others in these tasks. However, fine-tuning resulted in substantial improvements in sentiment and emotion classification—areas where the base Qwen model had previously struggled.

These results suggest that while our dataset is effective in enhancing LLMs' capabilities in particularly intricate and niche tasks, the extent of improvement may depend on the pre-existing strengths of each model. Models that initially performed poorly, such as those struggling with sentiment classification, exhibited more noticeable improvements after fine-tuning, suggesting that our dataset is particularly beneficial for models with weaker prior knowledge and could possibly enhance their ability to interpret culturally relevant content. Conversely, models that were already strong in specific tasks, such as GPT in emotion classification, may experience diminishing returns or even degradation in performance due to overfitting. The graphs showing performance change are in Appendix A.2.

5 Discussions

5.1 The Bidirectional Cultural Gap and Usage of LLMs

Our findings indicate that Chinese participants exhibited relatively low accuracy in multiple-choice question answering, sentiment labeling, and emotion classification. As shown in Table 2, they were frequently outperformed by LLMs, confirming one direction of the cultural gap: Chinese participants face challenges in understanding U.S. memes.

Additionally, when Chinese participants made errors in the multiple-choice task, they were more likely to select LLM-generated distractors rather than the human-assumed misunderstandings. This pattern was also observed in LLM testing. This suggests that human-assumed misunderstandings, written by U.S. participants during data collection, do not always align with what Chinese participants actually perceive when interpreting the memes. While LLMs can, to some extent, attempt to fathom out Chinese people's thought processes. This confirms the other direction of the cultural gap: U.S. participants may struggle to accurately anticipate how Chinese individuals interpret their memes.

These findings highlight an important application of LLMs beyond assisting Chinese participants in understanding U.S. memes. LLMs can also be leveraged to help U.S. participants anticipate potential misinterpretations of their shared content, allowing them to better understand how their messages might be perceived by individuals from different cultural backgrounds. This suggests that LLMs have the potential to facilitate cross-cultural communication by not only bridging comprehension gaps but also fostering perspective-taking. In

	DEF vs. US-RP	DEF vs. CN-RP	US-RP vs. CN-RP
Qwen	Acc_{Emo}	$Sim_{formatted}$ Acc_{Emo} Acc_{Sent}	$Sim_{original}$ $Sim_{formatted}$
GLM	$Sim_{original}$ $Sim_{formatted}$	Acc_{Emo}	$Sim_{original}$ Acc_{Emo}
LLaMA	$Sim_{formatted}$ Acc_{Emo}	Acc_{Emo}	$Sim_{original}$ $Sim_{formatted}$
GPT	$Sim_{original}$ $Sim_{formatted}$	$Sim_{formatted}$	$Sim_{original}$ $Sim_{formatted}$

Table 4: Metrics with significant differences across cultural settings for various LLMs ($p < 0.05$). For example, for the Qwen2.5-VL model, Acc_{Emo} in the DEF vs. US-RP column indicates significantly better performance of US-RP on the emotion labeling task compared to DEF .

practical applications, this could help reduce misunderstandings, mitigate awkwardness, and prevent unintended conflicts in intercultural exchanges.

5.2 Roleplaying Effects on LLMs

Based on our experiment results, explicitly instructing LLMs to engage in role-playing can significantly enhance their performance on certain metrics, revealing that LLMs are aware of different cultural settings. This suggests that LLMs can adjust their interpretations and responses when guided to adopt a particular cultural perspective. However, the degree of improvement varies across different tasks and models, indicating that LLMs' underlying understanding may still be limited by their training data and pre-existing biases.

Interestingly, even when LLMs are instructed to role-play as native Chinese, their ability to interpret U.S. memes improves compared to the default setting—although the improvement is not as pronounced as when prompted to act as native U.S. people. This suggests that LLMs may be intentionally aligned to suppress their cultural tendencies, possibly to avoid exhibiting explicit biases. Given that their training data are predominantly in English, their stronger performance when acting as English speakers is unsurprising. However, if they were not aligned, they might exhibit strong cultural biases, leading to skewed interpretations. Post-alignment, their cultural biases appear to be substantially suppressed, to the extent that their ability to understand U.S. memes is sometimes lower when acting as a Chinese speaker than in their default setting. This suggests that while cultural bias introduced by training data has been effectively mitigated, LLMs also exhibit an awareness of adjusting their responses under explicit instructions. In essence, this reveals that LLMs are not merely reflecting biases from training data but are also capable of controlled cultural adaptation when explicitly guided.

6 Limitations

While our study provides valuable insights into the role of LLMs in cross-cultural meme interpretation, several limitations should be acknowledged. The size of our dataset is limited due to the challenges associated with crowdsourcing raw meme data, including both images and detailed annotations. However, our proposed data collection pipeline has demonstrated its reliability, and we believe it is feasible to scale up the dataset in future research. The relatively small dataset size may negatively impact LLM fine-tuning, potentially leading to overfitting. While fine-tuning has generally improved model performance, certain tasks, such as emotion classification in GPT-4o, exhibited performance degradation, likely due to overfitting to the limited data.

Besides, although we identified a bidirectional cultural gap, our study did not validate its reversal—where Chinese participants provide memes and U.S. participants attempt to interpret them. It remains an open question whether Chinese participants would also struggle to predict potential misunderstandings by U.S. participants and how challenging U.S. participants would find it to interpret Chinese memes. Investigating this aspect would provide a more comprehensive understanding of cross-cultural meme interpretation.

Lastly, Our study focuses exclusively on Chinese and U.S. cultural contexts, leaving out other linguistic and cultural backgrounds that may exhibit distinct patterns in meme interpretation. Future work should extend this research to a broader range of cultural settings to explore whether similar bidirectional gaps exist across other regions and communities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. Mememqa: Multimodal question answering for memes via rationale-based inferencing. *arXiv preprint arXiv:2405.11215*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Minh Duc Bui, Katharina von der Wense, and Anne Lauscher. 2024. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. *arXiv preprint arXiv:2411.03888*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

L Grundlingh. 2018. Memes as speech acts. *Social Semiotics*, 28(2):147–168.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. In *EMNLP*, pages 1433–1445.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. *arXiv preprint arXiv:2305.16171*.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *EMNLP*, pages 10258–10279.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *EMNLP*, pages 7069–7086.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Shahira Mukhtar, Qurat Ul Ain Ayyaz, Sadaf Khan, Atiya Muhammad Nawaz Bhopali, Muhammad Khalid Mehmood Sajid, Allah Wasaya Babbar, et al. 2024. Memes in the digital age: A sociolinguistic examination of cultural expressions and communicative practices across border. *Educational Administration: Theory and Practice*, 30(6):1443–1455.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. *arXiv preprint arXiv:2409.13592*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *WWW*, pages 1907–1917.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *ACL*, pages 428–446.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of Child Language*, 14(2):201–209.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773.

Shivam Sharma, S Ramaneswaran, Udit Arora, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Memex: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *ACL*, pages 5272—-5290.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of EMNLP*, pages 4996–5025.

GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Wan-Hsiu Sunny Tsai and Linjuan Rita Men. 2017. Consumer engagement with brands on social network sites: A cross-cultural comparison of china and the usa. *Journal of Marketing Communications*, 23(1):2–21.

Vanshika, Neetu Rani, and Ranjan Walia. 2024. A comprehensive review of sentiment analysis: Techniques, datasets, limitations, and future scope. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 403–409.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of EMNLP*, pages 13078–13096.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*, pages 2039–2055.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *IMC*, pages 188–202.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *CVPR*, pages 2007–2017.

Naitian Zhou, David Jurgens, and David Bamman. 2024. Social meme-ing: Measuring linguistic variation in memes. In *NAACL-HLT*, pages 3005–3024.

A Appendix

A.1 Performance Score Calculation

The number of possible label sets (PLS) for each meme is given by:

$$\text{PLS} = \sum_{l=1}^{n} \binom{n}{l} = \sum_{l=1}^{n} \frac{n!}{l!(n-l)!} \tag{3}$$

where n is the number of possible labels (in this case $n = 4$), and l is the number of labels chosen by the model ($l \in [1, 4]$). The number of possible correct label sets (PCLS) is then defined as:

$$\begin{aligned} \text{PCLS}_m &= \sum_{l_m^i=0}^{n-c_m} \binom{n-c_m}{l_m^i} \\ &= \sum_{l_m^i=0}^{n-c_m} \frac{(n-c_m)!}{l_m^i![(n-c_m)-l_m^i]!} \end{aligned} \tag{4}$$

where c_m is the number of true labels for a given meme ($c_m \in [1, 4]$), and l_m^i represents the number of false labels for the current meme ($l_m^i \in [0, 3]$). The expected accuracy for emotion classification is then given by:

$$\mathbb{E}_{\text{Emo}}[\text{Acc}] = \left(\sum_{m \in \mathcal{M}} \frac{\text{PCLS}_m}{\text{PLS}} \right) \Big/ |\mathcal{M}| \tag{5}$$

where $\mathcal{M}$ represents the set of all memes, with $|\mathcal{M}| = 621$. Based on our dataset, 440 memes have one emotion label, 174 have two emotion labels, 7 have three emotion labels, and no meme has all four emotion labels. Therefore, we got $\mathbb{E}_{\text{Emo}}[\text{Acc}] \approx 0.12$. To assign scores, we solve the following system of equations:

$$\begin{cases} \mathbb{E}_{\text{Emo}}[\text{Acc}] \cdot x = \mathbb{E}_{\text{MCQ}}[\text{Acc}] \cdot y \\ x + 2y = 3\mathbb{E}[\text{PS}] \end{cases} \tag{6}$$

, where x is the maximum possible possible score assigned to the emotion labeling, and y is the maximum possible possible score assigned to both the multiple choice question selection and sentiment labeling. Then, the final performance score can be computed as:

$$\begin{aligned} \text{PS} &= \text{Sim}_{\text{original}} + \text{Sim}_{\text{formatted}} \\ &\quad + \text{PS}_{\text{MCQ}} + \text{PS}_{\text{Sent}} + \text{PS}_{\text{Emo}} \\ &= \text{Sim}_{\text{original}} + \text{Sim}_{\text{formatted}} \\ &\quad + \text{Acc}_{\text{MCQ}} \cdot y + \text{Acc}_{\text{Sent}} \cdot y + \text{Acc}_{\text{Emo}} \cdot x \end{aligned} \tag{7}$$

A.2 Fine-tuned Models Performance

According to the results shown in Figure 4, our dataset has the potential to enhance LLMs' ability to interpret memes, provided that overfitting does not occur. To mitigate the risk of overfitting, we recommend following our dataset curation pipeline to ensure the creation of a sufficiently large dataset.

B Additional Details on the Rewriting Process of Original Data

This appendix details the utilization of Large Language Models (LLMs), specifically GPT-4, in several key stages of our data processing pipeline. We employed GPT-4 for content reformatting, translation, and the generation of multiple-choice questions (MCQs) based on the collected meme explanations and potential misunderstandings.

B.1 Content Rewriting

We designed specific prompts to guide GPT-4 in standardizing meme explanations while preserving their original meaning and terminology.

B.1.1 Introduction Prompt

```
Please act as a cultural analyst to
    ↪ standardize explanations of memes
    ↪  while preserving their original
    ↪ meaning and terminology. Your
    ↪ task is to reformat provided meme
    ↪  explanations according to strict
    ↪  guidelines, ensuring all key
    ↪ terms, slang, and cultural
    ↪ references remain unchanged.
```

B.1.2 Key Requirements Prompt

```
Preservation Rules:
Maintain ALL original keywords, phrases,
    ↪  and cultural references.
Ensure the reformatted explanation has
    ↪ the same meaning as the original.
Add contextual framing only where
    ↪ necessary for clarity.

Structural Rules:
Begin explanations with "In the US, this
    ↪  meme" if US cultural context is
    ↪ involved.
For potential misunderstandings, retain
    ↪ the original concern but
    ↪ standardize phrasing.
```

B.1.3 Instruction Prompt

```
For each meme explanation:
Explanation: Start with "This meme" or "
    ↪ In the US, this meme," followed
    ↪ by the original content.
Potential Misunderstanding: Begin with "
    ↪ People might" or "Some viewers
    ↪ might," then state the
    ↪ misunderstanding exactly as
    ↪ described.

Examples:
Original: "The joke is about student
    ↪ loans being expensive"
-> Standardized: "This meme refers to
    ↪ student loans being expensive."
Original: "Missing the reference to
    ↪ SpongeBob"
-> Standardized: "People might miss the
    ↪ specific reference to SpongeBob."
```

B.2 Content Translation

After the rewriting stage, the dataset was translated into Chinese to facilitate cross-cultural comparison. We again utilized the GPT-4 API for this task. The prompts for translation emphasized accuracy and fluency in the target language while preserving the nuances of the meme interpretations.

B.2.1 Instruction Prompt for Translation

```
For each segment of text to be
    ↪ translated:
Translation: Strictly adhere to the
    ↪ following guidelines:
Maintain a professional academic tone,
    ↪ avoiding stiff or overly
    ↪ technical terminology.
Retain the original sentence structure
    ↪ and all numbers/proper nouns.
Translate culturally specific
    ↪ expressions through paraphrasing
    ↪ while preserving the original
    ↪ meaning.
```

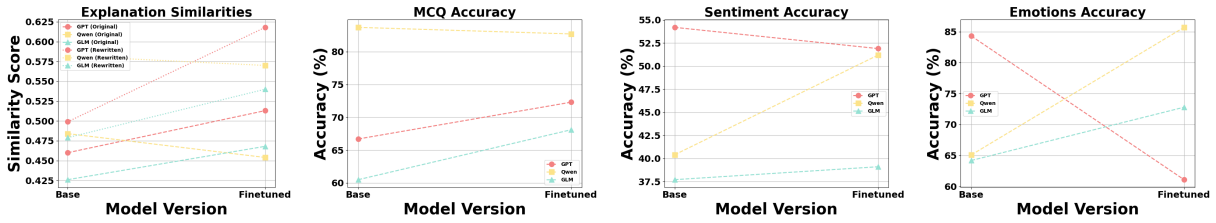
B.3 Multiple Choice Questions Generation

In the cross-cultural test phase, we generated two multiple-choice questions based on explanations and potential misunderstandings. The rewritten explanations were used as the correct answers, while the potential misunderstandings were adapted into distracting choices. The GPT-4 API played an important role in perfecting these transformations. We distributed two distinct prompts to GPT-4.

B.3.1 Prompt 1: Misleading Option Generation

```
Task: Generate a misleading option based
    ↪  on a misunderstanding.

Correct answer: [Rewritten_Explanation]
```



(a) Explanation Similarities (b) MCQ Accuracy (c) Sentiment Accuracy (d) Emotions Accuracy

Figure 4: Performance comparison between base models and their finetuned versions across different metrics.

```
Possible misunderstanding: [
    ↪ Rewritten_Misunderstanding]

Please generate a misleading option
    ↪ based on the misunderstanding
    ↪ based on the above information.

Requirements:
1. The option should be misleading: The
    ↪ option should look reasonable,
    ↪ but actually wrong, and can
    ↪ mislead people who are not
    ↪ familiar with the correct answer.
2. The option should be related to the
    ↪ misunderstanding: The option
    ↪ should be constructed based on
    ↪ the provided misunderstanding and
    ↪  can reflect the specific
    ↪ manifestation of the
    ↪ misunderstanding.
3. The option should be concise and
    ↪ clear: The option should be
    ↪ concise and clear, and avoid
    ↪ using overly complex or ambiguous
    ↪  language. Keep it to about 20-30
    ↪  words.
```

B.3.2 Prompt 2: Chinese Cultural Perspective on Misunderstanding

```
Please play the role of a Chinese
    ↪ culture who lacks in-depth
    ↪ understanding of the American
    ↪ cultural background.

Task description:
1. Meme selection: The model will
    ↪ provide a series of network
    ↪ factors (meme) pictures
    ↪ originating from American culture
    ↪ .
2. Cultural background information
    ↪ return: It is assumed that there
    ↪ is no similarity in the American
    ↪ culture, history, social
    ↪ background and other information
    ↪ involved in the meme.
3. Misunderstanding possibility analysis
    ↪ : Subjectively need to try to
    ↪ infer the meaning of the meme
    ↪ based on the content of the
    ↪ picture and combined with one's
    ↪ own cultural cognition, and
    ↪ record possible misunderstandings
    ↪  or deviations. Keep it to about
    ↪ 20-30 words.
```

C Prompt for Model Testing

This appendix the structured prompts developed to evaluate how language models interpret memes across cultural and academic frameworks. The prompts simulate distinct perspectives (American, Chinese, and neutral academic analysis) to systematically assess cultural bias, interpretability gaps, and alignment with region-specific or objective contexts.

C.1 Prompt For Academic Research Analysis

```
For academic purposes, analyze this meme
    ↪  neutrally. Follow this format:

Explanation: Objectively summarize the
    ↪ memeâĂŹs cultural relevance in
    ↪ the US (20-30 words).
Misunderstanding: Identify a plausible
    ↪ misinterpretation by non-US
    ↪ audiences (20-30 words).
Sentiment: [Positive/Negative/Neutral]
Emotions: [Sarcastic, Humorous,
    ↪ Motivational, Offensive]
```

C.2 Prompt For American Perspective

```
As a native American living in the US,
    ↪ analyze this meme. Follow this
    ↪ format:

Explanation: As someone familiar with US
    ↪  culture, explain the memeâĂŹs
    ↪ meaning to Americans (20-30 words
    ↪ ).
Misunderstanding: How might non-
    ↪ Americans misinterpret this meme
    ↪ due to cultural differences?
    ↪ (20-30 words).
Sentiment: [Positive/Negative/Neutral]
Emotions: [Sarcastic, Humorous,
    ↪ Motivational, Offensive]
```

C.3 Prompt For Chinese Perspective

```
As a native Chinese person, analyze this
    ↪  American meme. Follow this
    ↪ format:

Explanation: From your Chinese cultural
    ↪ viewpoint, interpret the memeâĂŹs
    ↪  intent or symbolism (20-30 words
    ↪ ).
Sentiment: [Positive/Negative/Neutral]
Emotions: [Sarcastic, Humorous,
    ↪ Motivational, Offensive]
```