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

Hangxiao Zhu
hangxiao@tamu.edu
Texas A&M University
College Station, USA

Suliu Qin
Sophie508727@gmail.com
Independent Researcher
College Station, USA

Zhuoyan Li
li4178@purdue.edu
Purdue University
West Lafayette, USA

Ming Jiang
ming.jiang@wisc.edu
University of Wisconsin–Madison
Madison, USA

Yu Zhang
yuzhang@tamu.edu
Texas A&M University
College Station, USA

Meng Xia
mengxia@tamu.edu
Texas A&M University
College Station, USA

## Abstract

Communicating across cultures is inherently challenging, especially through culturally dense and ambiguous formats like memes. While people expect large language models (LLMs) hold promise for bridging such gaps, existing benchmark datasets often fail to capture the cultural context necessary for accurate interpretation. To address this, we introduce MemeBridge, a curated dataset centered on U.S.-originated memes, designed to capture two complementary perspectives: (1) how Chinese participants interpret these memes, and (2) how U.S. participants anticipate how people from other cultures might misunderstand them. Here, context refers to implicit cultural knowledge—background beliefs, norms, and shared assumptions that shape meme comprehension. The dataset was constructed via a multi-stage crowdsourcing pipeline with rigorous validation, including human agreement checks and GPT-based classification verification. Each meme is annotated with sentiment, emotion, cultural significance, and knowledge type, providing rich supervision for downstream tasks. Notably, we observe that the anticipated misunderstandings from U.S. participants are often inaccurate, highlighting the asymmetries in cultural understanding and the challenges of adopting perspectives beyond one's own. This bidirectional framing—focusing on both expression and perception—enables more nuanced benchmarking of cross-cultural comprehension. Our probing of multiple LLMs reveals that while models developed in different cultural contexts exhibit partial cross-cultural understanding, they often struggle with sophisticated interpretations. By contrast, fine-tuning with MemeBridge improves model performance, underscoring the value of culturally grounded resources for training and evaluating LLMs in globally diverse settings.

## CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing → Sociology**.

## Keywords

Human Behavior Analysis, NLP Tools for Social Analysis

**ACM Reference Format:**
Hangxiao Zhu, Suliu Qin, Zhuoyan Li, Ming Jiang, Yu Zhang, and Meng Xia. 2026. MemeBridge: A Dataset for Benchmarking and Mitigating the Bidirectional Cultural Gap in Meme Interpretation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770854.3785691

## 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 [7]. 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 [18]. 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.

> *"... 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 [29]. 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.

*KDD '26, Jeju Island, Republic of Korea*

ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3785691

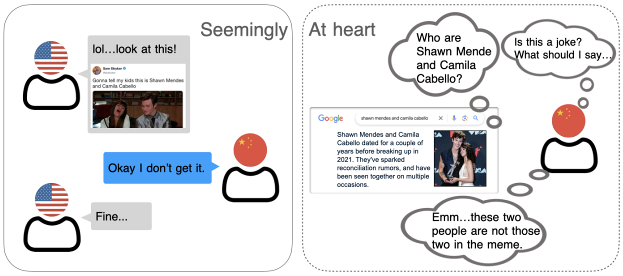

Seemingly
lol...look at this!
Sam Stryker
Gonna tell my kids this is Shawn Mendes and Camila Cabello
Okay I don't get it.
Fine...
At heart
Who are Shawn Mende and Camila Cabello?
Is this a joke? What should I say...
Google
shawn mendes and camila cabello
Shawn Mendes and Camila Cabello dated for a couple of years before breaking up in 2021. They've sparked reconciliation rumors, and have been seen together on multiple occasions.
Emm...these two people are not those two in the meme.

**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.**

| Feature | FigMemes [16] | MemeCap [11] | Multi3Hate [4] | MemeBridge |
|---|---|---|---|---|
| Bidirectional Framing | ✗ | ✗ | ✗ | ✓ |
| Sentiment Annotation | ✗ | ✗ | ✗ | ✓ |
| Emotion Annotation | ✗ | ✗ | ✗ | ✓ |
| Interpretation Caption | ✗ | ✓ | ✗ | ✓ |
| Native-Speaker Annotation | ✗ | ✓ | ✓ | ✓ |

**Table 1: Comparison between MemeBridge and existing meme datasets.**

As large language models (LLMs) continue to advance, multimodal variants such as GPT-4o [10], Llama-3.2-Vision [6], GLM-4V [28], and Qwen-VL [3] have become increasingly 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 [27] 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 [34]. 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.

Several existing meme datasets have supported progress in meme understanding, captioning, and harmful content detection. However, these datasets are not designed to explicitly address cross-cultural interpretation. In contrast, our work focuses on bridging cultural gaps in meme understanding by introducing a bidirectional framing that captures both native-speaker interpretations and potential cross-cultural misunderstandings. In addition, we incorporate sentiment and emotion annotations grounded in cultural context, along with native-speaker verification to ensure interpretive fidelity. Table 1 summarizes these key distinctions, highlighting how MemeBridge complements prior resources rather than serving as a direct substitute.

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) Model performance is not strongly tied to the country of origin. Despite being developed by organizations based in either China or the United States, all evaluated LLMs exhibited comparable performance and behavioral patterns across tasks. This suggests that contemporary alignment and safety-tuning practices mitigate overt, origin-specific cultural biases, rather than embedding strong national perspectives. (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 [21, 22]. 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 [30]. For example, Shi et al. [27] 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 [14, 27, 33] 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. [20] 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 [13], 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 [15, 17]. 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 [9, 12]. 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 [4, 19]. 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. [32] 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 [16] focuses on the identification of figurative language in political memes. MCC [26] contains 3,400 memes and their contexts focusing on detecting explanatory evidence for memes. MemeCap [11] enables the evaluation of visual language models in the meme captioning task. SemanticMemes [35] highlights semantic clustering. MemeMQA [2] offers a multimodal question-answering framework for a better semantic explanation of memes. Multi3hate [4] 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. Participant demographics are provided in Appendix A, while recruitment and compensation details are available in Appendix B. The full research consent form is included in Appendix C. The MemeBridge dataset is publicly available[1].

### 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. To ensure consistent interpretation, annotators were provided with detailed instructions, including an illustrative example: how non-U.S. participants might misinterpret the phrase "Netflix and Chill" as a literal movie invitation, not realizing its euphemistic meaning. This encouraged contributors to reflect on how culturally specific cues might be missed or misread. The full annotation prompt is included in Appendix D. 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} [25]. 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. [31], 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 a binary scheme, i.e. classifying each explanation and misunderstanding as either "good" or "bad" quality. Majority voting was used to determine the final label for each data point. This labeled dataset was then used to train a BERT-based classifier [5] to filter out low-quality meme interpretations. To justify the classifier's reliability and training sample sufficiency, our results in Appendix E show that 200 training samples are enough to achieve strong classification performance. 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 [24] 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 [1] to rewrite the original data, standardize the format, and improve grammatical accuracy while preserving semantic integrity. To ensure consistency, we applied a similarity scoring 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 F. 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.

[1] https://drive.google.com/drive/folders/152AN3iREfi71WThArmr8OcUWM5cV8YZy

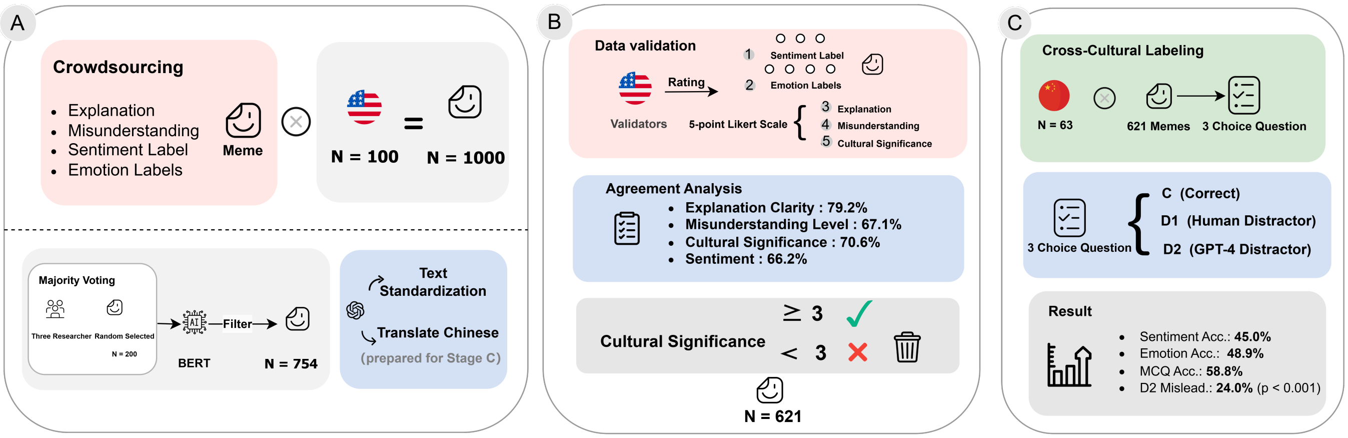


**Figure 2: (A) Data Collection & Cleaning: 200 memes randomly selected from 1000 collected and labeled by three researchers, followed by BERT-based refinement (N=754). (B) Data Validation: Participants labeled memes for sentiment and emotion. Explanation text, misunderstandings, and cultural significance 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.**

| | Positive | Neutral | Negative | Total |
|---|---|---|---|---|
| **Sarcastic** | 31 | 91 | 60 | 182 |
| **Humorous** | 167 | 285 | 87 | 539 |
| **Motivational** | 24 | 9 | 2 | 35 |
| **Offensive** | 3 | 0 | 24 | 27 |

**Table 2: Distribution of sentiment labels within each emotion category.**

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.

We summarize the final distribution of sentiment and emotion labels in Table 2. This shows a reasonably balanced sentiment distribution, with a diverse range of emotions well-represented.

To further analyze the dataset, we employed GPT-4 to classify each meme along two dimensions. The prompts used for these classification tasks are provided in Appendix G. For topic labeling, we adopted five categories inspired by prior work such as MEMEX [26]: Political Satire, Entertainment, Films and Cartoons, Pop Culture, and Social Events. The distribution of our dataset across these categories is shown in Figure 3. Each meme was also categorized as requiring either cultural or general knowledge for correct interpretation. This binary classification was again performed using GPT-4o, with one label per meme. Table 3 shows the resulting distribution. Together, these results confirm that the dataset spans diverse topical domains and is rich in culture-specific content—consistent with our goal of facilitating research on cross-cultural meme understanding.

| Knowledge Type | Number of Memes |
|---|---|
| Cultural Knowledge | 529 |
| General Knowledge | 92 |

**Table 3: Distribution of memes by required knowledge type.**

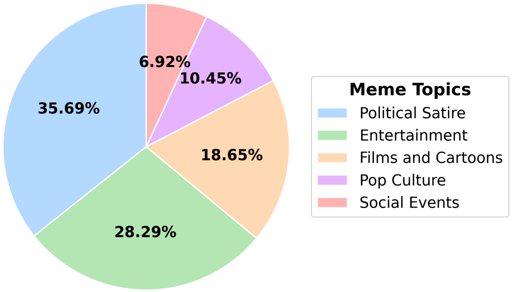


**Figure 3: Distribution of meme topics in the dataset.**

We conducted a separate human validation study to confirm the agreement level and assess the reliability of our automatic topic and knowledge-type annotations (general vs. cultural knowledge). We randomly sampled 200 memes from the dataset, and one author manually labeled each meme for both topic and knowledge type. These human-provided labels were then compared with GPT-4o's automatic classifications. The agreement rates by category and Cohen's Kappa are shown in Table 4. These results demonstrate strong

(Topic Classification)

| Category | Agreement (%) |
|---|---|
| Political Satire | 91.3 |
| Entertainment | 93.4 |
| Films and Cartoons | 93.3 |
| Pop Culture | 94.1 |
| Social Events | 83.3 |
| **Cohen's Kappa** | **0.89** |

(Knowledge Type Classification)

| Label | Agreement (%) |
|---|---|
| General vs. Cultural | 95.5 |
| **Cohen's Kappa** | **0.85** |

**Table 4: Human-GPT-4o agreement on topic and knowledge-type classification (N=200).**

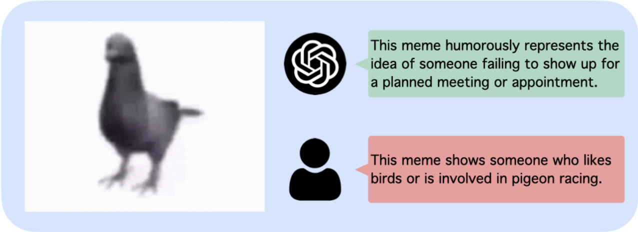

This meme humorously represents the idea of someone failing to show up for a planned meeting or appointment.
This meme shows someone who likes birds or is involved in pigeon racing.

**Figure 4: Example comparison between GPT-4-generated and human-assumed distractors for two U.S. memes.**

alignment between human and GPT-4o annotations, with Cohen's Kappa values of 0.89 for topic classification and 0.85 for knowledge-type classification. This high level of agreement provides strong empirical evidence for the reliability of automatic labeling in our dataset.

## 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×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 cultural 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. The prompt used for generating $D_2$ is included in Appendix I. 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 all responses selected the human-assumed distractor ($D_1$), while 24.0% selected the LLM-generated distractor ($D_2$), indicating that Chinese participants were significantly more misled by the LLM-generated distractor than the human-assumed distractor ($p < 0.001$). This supports observation 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. To contextualize this difference, we include a comparison of LLM-generated and human-assumed distractors in Figure 4. Additional representative examples from the dataset are included in Appendix H.

| Task | Cultural Memes | General Memes |
|---|---|---|
| Multiple Choice (MCQ) | 59.69% (0.491) | 54.04% (0.500) |
| Sentiment Labeling | 43.54% (0.496) | 53.03% (0.500)$^*$ |
| Emotion Labeling | 47.76% (0.500) | 55.56% (0.498) |

**Table 5: Chinese participant performance on memes requiring cultural vs. general knowledge. Values are mean accuracy with standard deviation in parentheses. $^*$ indicates statistically significant difference at $p < 0.05$.**

To investigate whether misunderstandings by Chinese participants stem from a lack of cultural or general knowledge, we compared their performance across memes classified as requiring cultural versus general knowledge. Accuracy rates for three key tasks - multiple choice (MCQ), sentiment labeling, and emotion labeling - are shown in Table 5. A significance marker (*) indicates $p < 0.05$. The sentiment classification task showed a significant performance gap, with participants performing worse on memes requiring cultural knowledge. Emotion labeling performance was also lower for cultural memes, with marginal significance. These results indicate that observed misunderstandings are more likely due to cultural barriers rather than a lack of general knowledge.

# 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 [3], GLM-4V [28], Llama-3.2-11B-Vision [6], and GPT-4o [10][2]. 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. All prompts used for model evaluation are provided in Appendix J.

## 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

[2] 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 6: Comparing the performance of different LLMs with Chinese participants on Meme Understanding. This table presents the accuracy of 4 different LLMs and Chinese annotators across three tasks: Multiple choice questions selection (MCQ), Sentiment labeling (Sent), and Emotion labeling (Emo). Underlined values indicate statistically significant differences from Chinese annotators (p < 0.05).**

in Table 6. 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 [23].

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 5. 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.

Further analysis of the multiple-choice responses revealed a consistent association: both LLM judges and our Chinese participants more frequently selected LLM-generated distractors than human-assumed distractors, as shown in Figure 5. Importantly, this result is descriptive—we do not claim that LLM-generated distractors are inherently better or that their higher selection rate reflects a causal effect. In light of findings from the LLM-as-a-Judge literature on potential self-enhancement or source bias and other confounds (e.g., option position, verbosity/length, or implicit model cues) [8], we interpret the pattern conservatively. Nonetheless, the alignment suggests that when LLMs make errors, they may do so in ways that resemble misunderstandings observed among Chinese participants, supporting the view that cultural gaps are bidirectional—not only in interpreting memes, but also in anticipating how they may be misread across cultures.

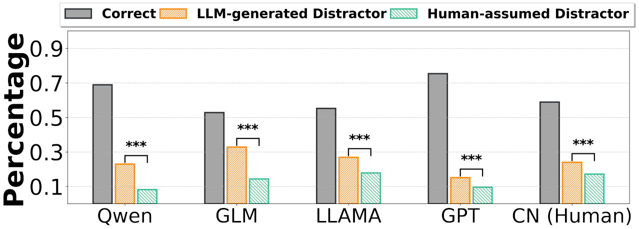


**Figure 5: Comparing the distribution of answer choices across different LLMs and Chinese participants (CN (Human)).Significance indicators (*: p < 0.05, **: p < 0.01, ***: p < 0.001) above pairs of distractor bars show whether participants significantly favored one type of distractor over the other 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-RP , and another instructing it to respond as a native Chinese CN-RP [3]. 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 7. 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 summarize overall model performance across five evaluation metrics, we introduce a simple aggregate measure called the *Performance Score* (PS), shown in Table 7. 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):

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

Here, $\mathrm{Sim}_{\mathrm{original}}$ and $\mathrm{Sim}_{\mathrm{formatted}}$ represent the cosine similarity between the model-generated explanation and the original crowdworker explanation, and the GPT-4 generated explanation, respectively. $\mathrm{Sim}_{\mathrm{formatted}}$ serves as a complementary, style-controlled similarity measure. Both similarity terms are computed in parallel and equally weighted for all models, ensuring that the aggregate

[3] RP denotes *role-playing*.

| | 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 |
| $\text{Acc}_{\text{MCQ}}$ | 68.8% | 67.9%↓ | 69.5%↑ | 52.8% | 46.7%↓ | 52.9%↑ | 55.2% | 47.8%↓ | 44.8%↓ | 75.4% | 72.7%↓ | 72.3%↓ |
| $\text{Acc}_{\text{Sent}}$ | 40.4% | 46.2%↑ | 47.0%↑ | 37.7% | 33.2%↓ | 38.3%↑ | 35.3% | 40.1%↑ | 39.0%↑ | 54.2% | 55.0%↑ | 51.9%↓ |
| $\text{Acc}_{\text{Emo}}$ | 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 7: Performance of LLMs across cultural settings and evaluation metrics. This table shows the performance of different LLMs under corresponding prompting strategies: DEF (Default Setting), US-RP (US RolePlaying), and CN-RP (Chinese RolePlaying). Performance is evaluated using metrics including similarity with original data ($\text{Sim}_{\text{original}}$) and formatted ($\text{Sim}_{\text{formatted}}$) text, Accuracy on Multiple choice questions selection ($\text{Acc}_{\text{MCQ}}$), Sentiment labeling ($\text{Acc}_{\text{Sent}}$), and Emotion labeling ($\text{Acc}_{\text{Emo}}$) tasks, along with a *Performance Score* (PS). For statistical significance test and pairwise comparisons across cultural settings, please refer the details in Table 3.**

score remains model-fair while capturing semantic alignment under both original and controlled textual forms. $\text{PS}_{\text{MCQ}}$, $\text{PS}_{\text{Sent}}$, and $\text{PS}_{\text{Emo}}$ correspond to the normalized performance scores for the multiple-choice, sentiment, and emotion classification tasks. Each component is scaled to ensure comparable weight, allowing the PS to serve as a compact but interpretable summary of overall model behavior across both generative and classification tasks. Full details on score normalization and weighting are provided in Appendix K.

Note that for emotion labeling, a prediction is marked correct if the ground-truth labels are a subset of the predicted labels. We acknowledge that the subset-based evaluation rule for emotion labeling could raise concerns about degenerate solutions—such as models predicting all available emotion labels to maximize correctness. However, this risk was mitigated by design: the evaluation criteria were not disclosed to either models or human participants, making strategic label inflation unlikely. Furthermore, our empirical analysis (see Appendix L) confirmed that such behavior did not occur. Most predictions included only 1–2 emotion labels, and no response—model-generated or human—selected all four. The expected number of labels per response ranged from 1.27 to 1.63, far below the theoretical maximum of 4. These results validate the realism of model behavior under the subset-based metric and support its robustness for evaluating multi-label emotion classification.

While we report all sub-metrics individually in Table 7, we found that examining them in isolation can make it tedious to interpret trends across models and prompt settings. For example, the GLM model shows improved similarity scores under the US-RP condition but performs worse on multiple-choice and sentiment classification compared to the default setting. Under CN-RP , the pattern reverses: accuracy scores improve, while similarity scores decline. This contradictory behavior makes it difficult to determine which persona setting is most effective for the model overall. The PS provides a compact, interpretable summary to address this challenge. It enabled us to observe consistent trends—such as models performing best under one of the role-play conditions—without obscuring the underlying metric-level variation. Importantly, the PS is not intended to replace individual metrics but to serve as a complementary tool for high-level comparison.

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 8) 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 indicate that persona-conditioned prompting can influence model behavior in ways that reflect alignment with cultural context. When the prompted identity matches the cultural background of the meme content—as in the US-RP setting—models tend to show improved interpretability. In contrast, prompts that are less aligned with the content's origin—such as CN-RP —are associated with relatively lower performance, though often still above the default setting. This suggests that models exhibit context-sensitive behavior when guided by role-specific instructions.

## 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 three models via their respective APIs: Qwen-2.5-VL-3B (developed by Alibaba), GLM-4V (by Zhipu AI), and GPT-4o (by OpenAI). Qwen-2.5-VL-3B has a publicly known parameter size of 3B. While the exact sizes of GLM-4V and GPT-4o are not disclosed, it is reasonable to assume that both are substantially larger than 9B parameters, especially in the case of GLM-4V, where the Zhipu AI API likely provides access to a more powerful variant than the open-sourced 9B model.

These models were evaluated on the same set of tasks: semantic similarity checking, multiple-choice question answering, sentiment labeling, and emotion classification. 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

| | DEF vs. US-RP | DEF vs. CN-RP | US-RP vs. CN-RP |
|---|---|---|---|
| **Qwen** | $\text{Acc}_{\text{Emo}}$ | $\text{Sim}_{\text{formatted}}$ $\text{Acc}_{\text{Emo}}$ $\text{Acc}_{\text{Sent}}$ | $\text{Sim}_{\text{original}}$ $\text{Sim}_{\text{formatted}}$ |
| **GLM** | $\text{Sim}_{\text{original}}$ $\text{Sim}_{\text{formatted}}$ | $\text{Acc}_{\text{Emo}}$ | $\text{Sim}_{\text{original}}$ $\text{Acc}_{\text{Emo}}$ |
| **LLaMA** | $\text{Sim}_{\text{formatted}}$ $\text{Acc}_{\text{Emo}}$ | $\text{Acc}_{\text{Emo}}$ | $\text{Sim}_{\text{original}}$ $\text{Sim}_{\text{formatted}}$ |
| **GPT** | $\text{Sim}_{\text{original}}$ $\text{Sim}_{\text{formatted}}$ | $\text{Sim}_{\text{formatted}}$ | $\text{Sim}_{\text{original}}$ $\text{Sim}_{\text{formatted}}$ |

**Table 8: Metrics with significant differences across cultural settings for various LLMs ($p < 0.05$). For example, for the Qwen2.5-VL model, $\text{Acc}_{\text{Emo}}$ in the DEF vs. US-RP column indicates significantly better performance of US-RP on the emotion labeling task compared to DEF .**

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 M.

# 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 6, 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 U.S. participants' assumptions don't always reflect how Chinese individuals interpret 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. In 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 behavior 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, LLMs interpret U.S. memes better when prompted to act as native Chinese speakers compared to their default setting—but not as well as when role-playing as native U.S. speakers. This suggests that alignment has reduced cultural bias, likely to avoid favoring specific groups. While their stronger performance as English speakers reflects their English-heavy training data, alignment appears to suppress these biases. In some cases, this even lowers performance when mimicking non-U.S. perspectives. Overall, LLMs show not just reduced bias, but an ability to adapt culturally when explicitly instructed.

# 6 Limitations

While our study provides valuable insights into the role of LLMs in cross-cultural meme interpretation, several limitations should be acknowledged. The dataset size remains relatively small due to the high cost and logistical complexity of crowdsourcing raw meme data along with rich, structured annotations. As detailed in Appendix B, the entire data collection process spanned approximately 40 days and incurred a cost of around $2000. Despite these constraints, our multi-stage data collection pipeline has proven reliable and scalable for future expansion. 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.

Third, 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.

Finally, we clarify that the "context" mentioned in our work refers to implicit cultural knowledge—the background beliefs, shared assumptions, and interpretive frameworks that shape how individuals understand memes. Although not explicitly embedded in text or image, this form of context plays a critical role in meme comprehension. Related to this, we also acknowledge that the potential misunderstandings collected in Stage 1 have limited predictive accuracy. These misunderstandings—imagined by U.S. participants—frequently failed to match actual responses from Chinese participants. We view this not as a flaw, but as a reflection of a deeper cultural asymmetry: that individuals often struggle to simulate interpretations from another cultural background. This limitation is amplified by the inherently subjective nature of misunderstanding, especially in ambiguous, culturally nuanced formats like memes. As misunderstandings are not objectively verifiable ground truths, it is difficult to collect reliable imagined misunderstandings through one-shot prompts. We hope future work can explore improved or interactive methods for eliciting these cross-cultural mismatches more accurately.

## References

[1] 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* (2023).

[2] Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing. In *Findings of ACL*. 5042–5078.

[3] 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* (2023).

[4] 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* (2024).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[6] 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* (2024).

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

[8] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594 [cs.CL] https://arxiv.org/abs/2411.15594

[9] Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of EMNLP*. 7591–7609.

[10] 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* (2024).

[11] EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In *EMNLP*. 1433–1445.

[12] Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and Multi-cultural Figurative Language Understanding. In *Findings of ACL*. 8269–8284.

[13] 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*. 10258–10279.

[14] 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. In *COLM*.

[15] Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *EMNLP*. 262–276.

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

[17] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *EMNLP*. 10467–10485.

[18] 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 (2024), 1443–1455.

[19] Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. YesBut: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models. In *EMNLP*. 16878–16895.

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

[21] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics* (2025), 1–96.

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

[23] Neetu Rani, Ranjan Walia, et al. 2024. A Comprehensive Review of Sentiment Analysis: Techniques, Datasets, Limitations, and Future Scope. In *International Conference on Computational Intelligence and Communication Technologies*. 403–409.

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

[25] 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*. 759–773.

[26] 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*. 5272—-5290.

[27] 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*. 4996–5025.

[28] 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* (2024).

[29] 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 (2017), 2–21.

[30] Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking Machine Translation with Cultural Awareness. In *Findings of EMNLP*. 13078–13096.

[31] 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*. 2039–2055.

[32] 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*. 188–202.

[33] 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. In *LREC-COLING*. 17696–17706.

[34] Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal Understanding of Memes with Fair Explanations. In *CVPR*. 2007–2017.

[35] Naitian Zhou, David Jurgens, and David Bamman. 2024. Social Meme-ing: Measuring Linguistic Variation in Memes. In *NAACL-HLT*. 3005–3024.

# A Participant Demographics

To better understand the background of our crowdsourcing participants, we summarize below the key demographic distributions for

| Stage | Sex (F/M) | Ethnicity | Employment |
|---|---|---|---|
| Stage 1 | 38 / 62 | Black (45), White (40), Mixed (12), Asian (2), Other (1) | Full-time (50), Part-time (26), Unemployed (7), Other (15), Not in paid work (2) |
| Stage 2 | 107 / 73 | White (119), Black (35), Mixed (10), Asian (10), Other (6) | Full-time (71), Part-time (26), Unemployed (19), Not in paid work (12), Other (50), Starting soon (2) |

**Table 9: Participant demographics in Stage 1 and Stage 2.**

both Stage 1 (initial data collection) and Stage 2 (validation). The Table 9 highlights sex, ethnicity, and employment status.

## B Participant Recruitment and Compensation

We used different recruitment strategies across stages. **Stage 1** (data collection) and **Stage 2** (validation) participants were recruited via Prolific with U.S.-based screening and demographic diversity controls. **Stage 3** (cross-cultural testing) participants were recruited through digital flyers shared via university mailing lists and social media groups targeting Chinese international communities. All participants were paid at $10/hour. Data collection lasted ~40 days, with total recruitment/compensation costs of ~$2000.

## C Research Study Consent Form

**Purpose:** Study cross-cultural understanding of internet memes.
**Procedure:** Tasks vary by stage and may include demographic questions and meme interpretation activities. Stage 1: upload ≥10 U.S. memes and provide cultural context, potential misunderstandings, sentiment, and emotion labels. Stage 2: validate meme interpretations and rate explanation/misunderstanding quality. Stage 3: answer multiple-choice interpretation questions for U.S. memes and label sentiment/emotions.
**Data Use & Confidentiality:** Responses are anonymous (no personally identifiable information collected) and stored securely for research use, including potential future studies.
**Risks/Benefits & Voluntary Participation:** No anticipated risks beyond everyday online content exposure. Participation is voluntary; you may skip questions or withdraw at any time without penalty.

## D Annotation Instructions

To guide misunderstanding annotations in Stage 1, participants were given the following instruction with an example based on the “Netflix and Chill” meme (Figure 6):

```
Describe how someone from another culture might
     misinterpret this meme if they lack the relevant
     cultural background (e.g., missing implied
     meanings or cultural references).
Example (Netflix and Chill): A non-U.S. viewer may take
     it literally as watching Netflix, not realizing it
     is a euphemism for romantic/sexual activity.
```

## E BERT Classifier Performance on Varying Training Sizes

We evaluated the reliability of our BERT-based classifier by training it on varying sizes of the 200 annotated samples. We reserved 50 samples as a held-out test set and trained the model on 10, 50, 100, and 150 examples. As shown in Table 10, classifier performance improved consistently with more training data and reached near-perfect levels with 150 samples, suggesting the sufficiency of our labeled set for quality filtering.

**Figure 6: Example meme used in the annotation prompt: “Netflix and Chill.”**

| # Training Samples | Accuracy | F1 Score | AUC-ROC |
|---|---|---|---|
| 10 | 0.891 | 0.940 | 0.964 |
| 50 | 0.938 | 0.963 | 0.978 |
| 150 | 0.969 | 0.981 | 0.992 |
| 200 | 0.984 | 0.991 | 1.000 |

**Table 10: BERT classifier performance on held-out test set (50 examples) under varying training sizes.**

## F Additional Details on the Rewriting Process of Original Data

This appendix details the utilization of GPT-4, in several key stages of our data processing pipeline.

```
You are a cultural analyst. Rewrite the input text into
     a standardized format without changing meaning.
Rules:
- Preserve all original keywords, slang, and cultural
     references (do not paraphrase them).
- Add minimal context only if needed for clarity.
- If U.S. cultural context is central, start with "In
     the US, this meme ..."; otherwise start with "This
     meme ...".
- For misunderstandings, keep the original concern but
     standardize phrasing.
Input:
Explanation (raw): [Explanation_Original]
Misunderstanding (raw, optional):
     [Misunderstanding_Original]
Output (follow exactly):
Explanation: [Standardized 1 sentence]
Potential Misunderstanding: [Standardized 1 sentence,
     start with "People might" / "Some viewers might"]
```

## G Prompts for Meme Classification

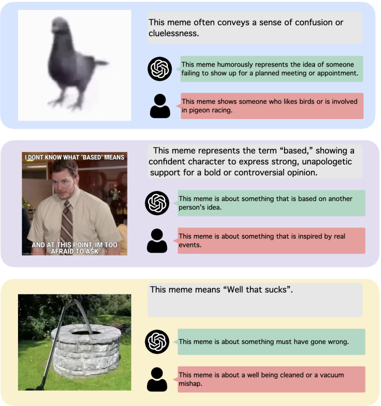
```
You are an expert in meme analysis. Given a meme image,
     classify it using the specified task and output
     format.
Task A (Knowledge Dependency):
Choose ONE label:
```

```
- Cultural Knowledge-Dependent: Requires U.S.-specific
     cultural knowledge (e.g., history, celebrities,
     media, norms).
- General Knowledge-Based: Understandable from
     universal experiences or common reasoning.
Output: Knowledge: [Cultural
     Knowledge-Dependent/General Knowledge-Based]

Task B (Topic Category):
Choose ONE label:
[Political Satire / Entertainment / Films and Cartoons
     / Pop Culture / Social Events]
Output: Topic: [Political Satire/Entertainment/Films
     and Cartoons/Pop Culture/Social Events]
```

## H Example Data Instances

To better illustrate the structure and labeling of our dataset, we present three representative examples below in Figure 7.



This meme often conveys a sense of confusion or cluelessness.

This meme humorously represents the idea of someone failing to show up for a planned meeting or appointment.

This meme shows someone who likes birds or is involved in pigeon racing.

I DONT KNOW WHAT "BASED" MEANS
AND AT THIS POINT, IM TOO AFRAID TO ASK

This meme represents the term “based,” showing a confident character to express strong, unapologetic support for a bold or controversial opinion.

This meme is about something that is based on another person's idea.

This meme is about something that is inspired by real events.

This meme means “Well that sucks”.

This meme is about something must have gone wrong.

This meme is about a well being cleaned or a vacuum mishap.

**Figure 7: Three example memes from the dataset. Each includes the original explanation (gray box), the GPT-4 generated interpretation (green box), and a human-annotated potential misunderstanding (red box).**

## I Multiple Choice Questions Generation

```
Template 1 (Misleading Option):
Input:
- Correct answer: [Rewritten_Explanation]
- Misunderstanding: [Rewritten_Misunderstanding]
Output:
- Misleading option (20-30 words): A plausible but
     incorrect choice consistent with the
     misunderstanding.

Template 2 (Chinese-View Misunderstanding):
Instruction: Assume the role of a native Chinese viewer
     with limited knowledge of U.S. cultural context.
Input:
- Meme image: [Meme]
Output:
- Possible misunderstanding (20-30 words): A plausible
     misinterpretation based only on the image and
     Chinese cultural intuition.
```

## J Prompt for Model Testing

This appendix the structured prompts developed to evaluate how language models interpret memes across cultural and academic frameworks.

```
You will analyze an American meme from a specified
     perspective.

Perspective: [Neutral Academic / American / Chinese]

Follow the output format exactly:

Explanation: (20-30 words) Provide the meme's meaning
     and cultural/contextual relevance as appropriate
     for the selected perspective.
Misunderstanding: (20-30 words) [Include ONLY if
     Perspective is Neutral Academic or American]
     Describe a plausible misinterpretation by
     non-US/non-American audiences due to cultural
     differences.
Sentiment: [Positive/Negative/Neutral]
Emotions: [Sarcastic, Humorous, Motivational, Offensive]
```

## K Performance Score Calculation

To construct PS, we aligned the expected scores of the classification tasks with the similarity metrics by setting $\mathbb{E}[\text{PS}_{\text{MCQ}}] = \mathbb{E}[\text{PS}_{\text{Sent}}] = \mathbb{E}[\text{PS}_{\text{Emo}}] = 1$. For multiple-choice and sentiment classification, where each question has three options, the expected accuracy is approximately 0.33. For emotion labeling, where a prediction is marked correct if the ground-truth labels are a subset of the predicted labels, we calculate expected accuracy as:

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

where $\mathcal{M}$ is the set of memes, PLS denotes the number of possible label sets for each meme, and PCLS is the number of those considered correct.

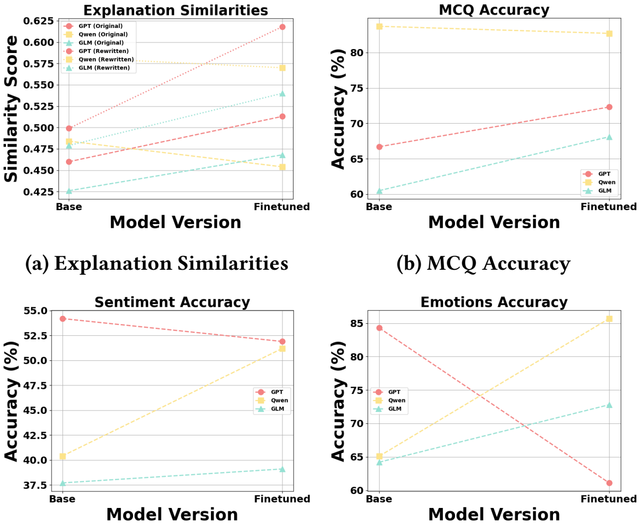
The number of possible label sets (PLS) for each meme is given by:

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

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



**(a) Explanation Similarities**

**(b) MCQ Accuracy**

**(c) Sentiment Accuracy**

**(d) Emotion Accuracy**

**Figure 8: Performance comparison between base models and their fine-tuned versions across different metrics.**

| # Labels | Qwen | GPT | GLM | LLaMA | Human |
|---|---|---|---|---|---|
| 1 | 68.9% | 75.4% | 52.8% | 55.2% | 74.5% |
| 2 | 22.9% | 15.1% | 32.8% | 26.9% | 23.8% |
| 3 | 8.2% | 9.5% | 14.4% | 17.9% | 1.7% |
| 4 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| $\mathbb{E}$[# Labels] | **1.39** | **1.34** | **1.62** | **1.63** | **1.27** |

**Table 11: Distribution of number of emotion labels selected per response.**

($l_m^i \in [0, 3]$). The expected accuracy for emotion classification is then given by Equation 2, 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:

$$\begin{cases} \mathbb{E}_{\text{Emo}}[\text{Acc}] \cdot x = \mathbb{E}_{\text{MCQ}}[\text{Acc}] \cdot y, \\ x + 2y = 3\mathbb{E}[\text{PS}], \end{cases} \tag{5}$$

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:

$$\begin{aligned} \text{PS} &= \text{Sim}_{\text{original}} + \text{Sim}_{\text{formatted}} \\ &\quad + \text{PS}_{\text{MCQ}} + \text{PS}_{\text{Sent}} + \text{PS}_{\text{Emo}} \\ &= \text{Sim}_{\text{original}} + \text{Sim}_{\text{formatted}} \\ &\quad + \text{Acc}_{\text{MCQ}} \cdot y + \text{Acc}_{\text{Sent}} \cdot y + \text{Acc}_{\text{Emo}} \cdot x. \end{aligned} \tag{6}$$

## L Emotion Label Count Distribution

For emotion prediction, we mark a response correct if the ground-truth emotion label(s) are a subset of the predicted labels. This raises a potential concern that a model could inflate accuracy by selecting all labels.

In our study, the grading rule was not disclosed to either models or human participants, making deliberate label inflation unlikely. To verify this empirically, we compare the distribution of the number of emotion labels selected per response across models and humans, summarized in Table 11.

These results show that both models and humans almost always choose 1–2 labels, occasionally 3, and never all 4. The expected number of selected labels ranges from 1.27 (human) to 1.63 (LLaMA), well below the degenerate maximum of 4, indicating no label inflation in practice.

## M Fine-tuned Models Performance

According to the results shown in Figure 8, 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.