AVMeme Exam: A Multimodal Multilingual Multicultural Benchmark for LLMs' Contextual and Cultural Knowledge and Thinking

Xilin Jiang[†♯], **Qiaolin Wang**[†♯], **Junkai Wu**[†♮], **Xiaomin He**[†♯], **Zhongweiyang Xu**[†♭]
Yinghao Ma[‡♩], **Minshuo Piao**[‡♯], **Kaiyi Yang**[‡♭], **Xiuwen Zheng**[‡♭], **Riki Shimizu**[‡♯],
Yicong Chen[‡♮] **Arsalan Firoozi**[‡♯], **Gavin Mischler**[‡♯], **Sukru Samet Dindar**[‡♯],
Richard Antonello[‡♯], **Linyang He**[‡♯], **Tsun-An Hsieh**[‡♭] **Xulin Fan**[‡♭], **Yulun Wu**[‡♭],
Yuesheng Ma[‡♯], **Chaitanya Amballa**[‡♭], **Weixiong Chen**[‡♩], **Jiarui Hai**[‡♪], **Ruisi Li**[‡♪],
Vishal Choudhari[‡♯], **Cong Han**[‡♩], **Yinghao Aaron Li**[‡◦],
Adeen Flinker[‡♪], **Mounya Elhilali**[‡♪], **Emmanouil Benetos**[‡♩],
Mark Hasegawa-Johnson[‡♭], **Romit Roy Choudhury**[‡♭], **Nima Mesgarani**[‡♯]

[♯]Columbia University, [♭]University of Illinois Urbana-Champaign, [♮]University of Washington,
[♪]Johns Hopkins University, [♪]New York University, [♩]Queen Mary University of London,
[♩]Google, [◦]Meta

{xj2289, qw2443, xh2369}@columbia.edu, junkaiwu@uw.edu, zx21@illinois.edu, nima@ee.columbia.edu

Abstract

Internet audio-visual clips convey meaning through time-varying sound and motion, which extend beyond what text alone can represent. To examine whether AI models can understand such signals in human cultural contexts, we introduce AVMeme Exam, a human-curated benchmark of over one thousand iconic Internet sounds and videos spanning speech, songs, music, and sound effects. Each meme is paired with a unique Q&A assessing levels of understanding from surface content to context and emotion to usage and world knowledge, along with metadata such as original year, transcript, summary, and sensitivity. We systematically evaluate state-of-the-art multimodal large language models (MLLMs) alongside human participants using this benchmark. Our results reveal a consistent limitation: current models perform poorly on textless music and sound effects, and struggle to think in context and in culture compared to surface content. These findings highlight a key gap in human-aligned multimodal intelligence and call for models that can perceive contextually and culturally beyond the surface of what they hear and see.
Project page: avmemeexam.github.io/review



Figure 1: A mosaic art of "AVMeme Exam" dotted by video frames from the meme clips.

1 Introduction

Artificial General Intelligence (AGI) Legg & Hutter (2007); Morris et al. (2023); Bubeck et al. (2023); Balepur et al. (2025) demands more than mastery of words. Human perception and communication also rely on sound and vision, along with other sensory inputs. AGI should also understand not only what is said or shown on the surface, but what is actually being communicated: the sarcasm in a voice, the emotion in a song, the sense of triumph or defeat in wordless music, and how billions of people can instantly recognize a second-long

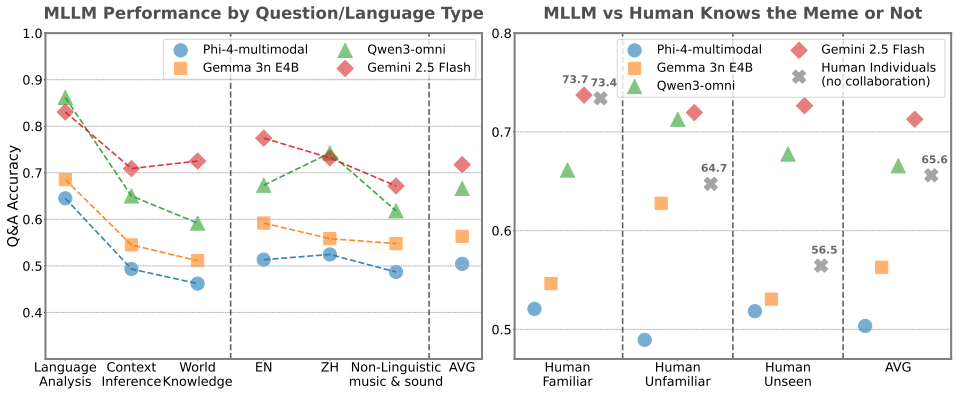


Figure 2: AVMeme Exam tests multimodal understanding beyond language and surface content. **Left:** Multimodal LLMs understand content much better than context, culture, and world knowledge, and media clips with spoken language than those without. **Right:** Open-source LLMs still lose to humans on humans' own familiar memes, while humans remain robust even on their unseen clips, showing that current LLMs are still limited in contextual and cultural thinking and real-world multimedia experience than real humans.

sound cue — from *Beethoven's fate motif* G–G–G–E$^{\flat}$ to *Nokia ringtone* E2–D2–F$^{\sharp}$1–G$^{\sharp}$1. Yet today's multimodal large language models (MLLMs) Alayrac et al. (2022); Liu et al. (2023); OpenAI (2024); Gong et al. (2024b); Tang et al. (2024); Yin et al. (2024) still face key challenges. First, time-varying audio and video signals convey meaning through prosody, melody, pacing, etc, all of which cannot be fully described by language. Second, understanding surface content, who says what or what happens, is only a starting point. Models must also understand why a moment is recognizable, how it is used, and what cultural significance it carries.

To examine whether MLLMs can serve as empathetic assistants that can resonate with human users, we study a pervasive yet understudied medium of online communication: the audio-visual *meme*. We define memes broadly as recognizable audio-visual clips (not necessarily humorous), including movie lines, sound effects, and musical phrases that people reuse with stable communicative purposes to express emotions and intentions. Because their meaning relies on delivery and shared culture besides content itself, they are an effective testbed for multimedia world understanding. Our central question is: *To what extent can MLLMs understand audio-visual memes including their literal content, underlying context, emotion, usage, and cultural grounding. Where do they succeed or fail?*

Yet, no existing audio-visual benchmark dives into aspects of context, culture, usage, and world knowledge. Audio benchmarks such as Dynamic-SUPERB Huang et al. (2024), AudioBench Wang et al. (2025), MMAU Sakshi et al. (2025), ADU-Bench Gao et al. (2025), and MMAR Ma et al. (2025) focus on recognition, captioning, dialogue understanding, and acoustic and semantic reasoning. While challenging, answers still reside in the waveform itself. Audio-visual benchmarks such as AV-SUPERB Tseng et al. (2024), AV-Odyssey Gong et al. (2024a), DAVE Radevski et al. (2025), AV-SpeakerBench Nguyen et al. (2025), and AVQA variants Li et al. (2022); Yang et al. (2022) examine event detection, ordering, alignment, and causality, but are still limited to what appears within the video frames. Meanwhile, a few language and image meme benchmarks indeed explore contextual and cultural implications, including Hateful Memes Challenge Kiela et al. (2020), What-Do-You-Meme Sharma et al. (2023), GOAT-Bench Lin et al. (2024), MemeSafetyBench Lee et al. (2025), MemeReaCon Zhao et al. (2025). But unlike static text and image memes, the time dimension of audio-visual memes carries rich temporal dynamics, emotional shifts, and other culturally meaningful cues that prior meme and cultural benchmarks have not studied.

To bridge these gaps, we introduce **AVMeme Exam**, a manual collection of over one thousand iconic Internet sounds and videos annotated with human-annotated metadata and human-written questions probing surface understanding, contextual inference, emotion,

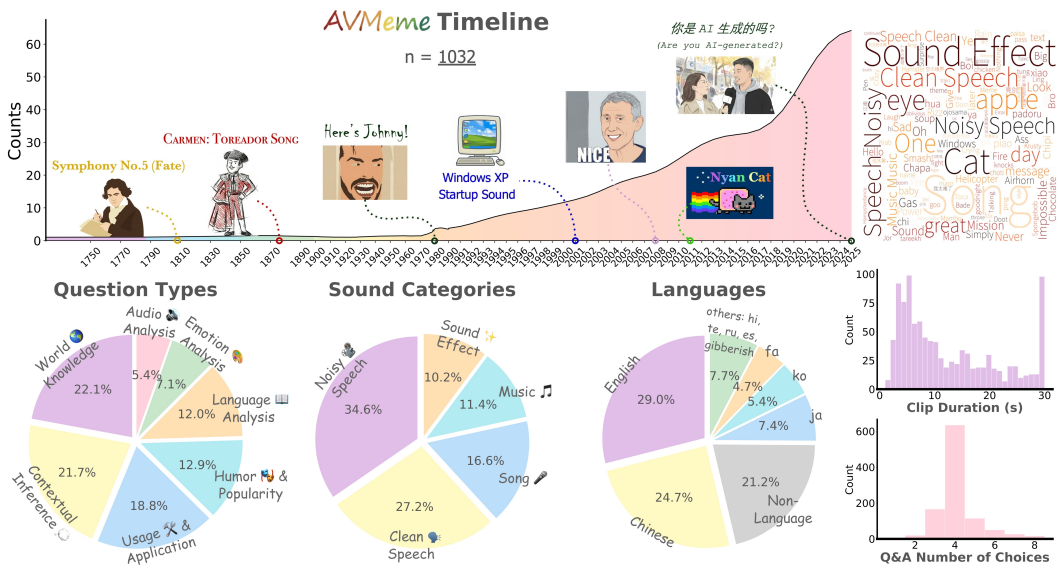


Figure 3: **Top:** Historical timeline of the 1,032 audio-visual memes curated in AVMeme Exam, spotlighting famous music rhythm, movie lines, sound effects, and viral Internet memes. **Bottom:** Pie charts summarize the distributions of question types, sound categories, and languages, highlighting the data diversity. **Right:** Frequent words in the memes' names and distributions of clip durations and number of choices. The duration is cut to 30 seconds, which is the maximum input audio length for most models.

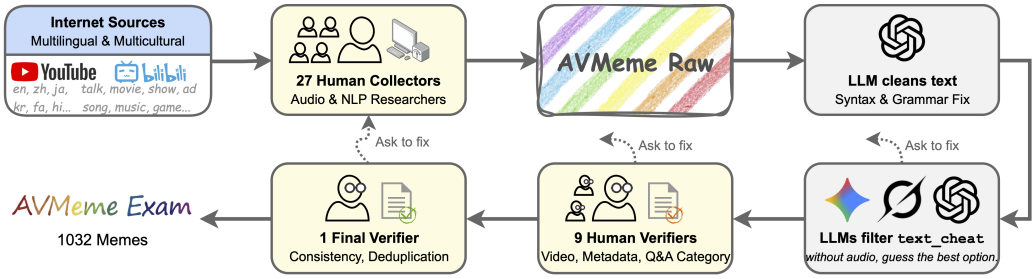


Figure 4: AVMeme Exam collection & verification pipeline. Videos and Q&As are human collected and verified (yellow). LLMs (gray) are used for text cleanup and to detect questions easily answered by text without audio given.

humor, usage, and world knowledge. Evaluation of state-of-the-art MLLMs and human participants reveals consistent limitations in models' ability to recognize, interpret, and culturally situate such clips. Together, our contributions establish AVMeme Exam as a comprehensive resource for diagnosing contextual and cultural weaknesses of AIs and guiding future progress in human-aligned multimodal intelligence.

We will first describe the AVMeme Exam in Section 2, and then present and analyze model performance in Section 3, along with key insights.

2 AVMeme Exam

Figure 4 illustrates the overall collection and verification pipeline of AVMeme Exam. More details are documented in Appendix A.

You shall not pass



Category: noisy speech
Year: 2001
Language: English

Transcription: You shall not pass.
Summary: A powerful wizard confronts a fiery monster on a narrow bridge, striking his staff and shouting defiantly, "You shall not pass!".

Usage: Used to jokingly or seriously deny someone access or prevent them from proceeding.
Emotion: fearful/anxious
Sensitivity: violence/gore
Question: *Which of the following is a humorous way a college student might use this meme?*

A. When talking about the spread of the flu among students.
B. **When discussing how difficult the final exam was in a class.**
C. When presenting a final project in an art class.
D. When asking the professor a question during a lecture.

Question Type: Usage & Application

Never Gonna Give You Up



Category: song
Year: 1987
Language: English
Transcription: We're no strangers to love. You know the rules and...
Summary: The start of music video "Never Gonna Give You Up".

Usage: A bait and switch is usually done using a disguised hyperlink leading to this music video.
Emotion: surprised/shocked, nostalgic
Sensitivity: none
Question: *What is the top reason this song became so popular as a meme?*

A. **Unexpected appearance on the Internet.**
B. Often got edited and remixed for parody.
C. A lot of references from other songs and movies.
D. The singer has sung it a lot.
E. It became a sports anthem.
F. It is popularized from forums such as 4chan and reddit.

Question Type: Humor & Popularity

Figure 5: Meme examples. More examples are provided in Appendix D.

2.1 Collection

The collection of AVMeme Exam is guided by three pillars: **Human-collected**: all clips are selected and annotated by 27 audio and NLP researchers, raised in the U.S., China, Japan, India, Middle East, and etc. They personally recognize and use these memes, ensuring that each entry is chosen with human judgment rather than automatic web scraping. **Audio-centric**: sound serves as the primary media of meaning, complemented by visual, spanning speech, songs, music, and wordless sound effects. **Multicultural-grounded**: the diverse linguistic and cultural backgrounds of contributors enables coverage of memes originating from East and South Asia, Middle East, Europe, and North America. Collectively, these pillars ensure that AVMeme Exam reflects authentic, culturally embedded, real-world multimodal communication.

In total, we collect **1,032** audio-visual memes spanning more than ten languages and five sound categories, with their distributions shown in Figure 3. Each meme consists of a video segment (URL, onset, offset) sourced from YouTube (86.1%) and Bilibili (13.9%), most exceeding one million views. It includes human-annotated metadata (summary, transcript, year, etc) and a multiple-choice Q&A. To maximize diversity, we place no restrictions on the videos (topic, region, etc), but safeguards still apply. We prohibit political materials and explicit depictions of sexual, violent, hateful, criminal, or drug-related content. Implicit depictions, verbal or contextual, are annotated using a sensitivity attribute drawn from {*sex, violence/gore, drug/alcohol, race/gender/geography/identity, disadvantaged/disability*}. Despite these safeguards, ***some clips may remain sensitive for certain people, and discretion is advised.*** The emotion attribute takes one or multiple from {*happy, sad, angry/annoyed, fearful/anxious, surprised/shocked, sarcastic/ironic, cringe/embarrassed, nostalgic, neutral*}. The question type is one of the seven (elaborated later). Figure 5 shows two examples.

Finally, we prompt GPT-5.1 OpenAI (2025) for syntax and grammar fix of the summary, usage, and Q&A fields, *"Do minimal or no change if possible. Try not to add or change words unless necessary."* to preserve the human annotations as original as possible.

2.2 Verification

Nine human verifiers check every entry, reviewing all metadata fields: transcription, summary, usage, sensitivity, language, video integrity, and the multiple-choice Q&A, on a clip-by-clip basis, and return to contributors to fix if any issues are identified.

While this process ensures syntactic correctness and annotation quality, it is not sufficient to guarantee that the Q&A genuinely tests multimodal understanding, as some questions may still be solvable through unintended shortcuts Zheng et al. (2024); Balepur et al. (2025). We therefore begin with **text-cheat** detection. We run three LLMs in text-only modes: Gemini 2.5 Flash Comanici et al. (2025), Grok 4 xAI (2025), and GPT-5.1 OpenAI (2025), with prompt "*without audio, guess the best option*." Q&As that all three models guess correctly are flagged as `text_cheat`, for which the strong text priors allow the LLMs to infer the clip or the answer without listening or watching. For example, *a famous sci-fi movie from the late 1970s* almost inevitably points to *Star Wars*. Aware of this problem, we explicitly prohibit contributors from mentioning names, authors, historical events, or other giveaway keywords in the Q&A. A small number of Q&As correctly guessed by all three LLMs are still retained, since they may be guessed by chance or by choosing the most well-known option. For example, when *John Williams*, the most famous film composer, appears among other composers, it is typically selected regardless of question phrasing. The full dataset is denoted `meme-full` (1032 memes), and removing all `text_cheat` items yields `meme-main` (846 memes). From `meme-full` to `meme-main`, the deaf-and-blind guessing accuracy drops from 34.7→20.4, 37.9→24.3, and 40.9→27.9 for Gemini 2.5 Flash, Grok 4, and GPT-5.1, respectively, showing that `meme-main` represents a more exclusive test of multimodal understanding.

Next, we assess **visual-cheat**. While we treat visual as a supplemental signal, it may trivially reveal the answer, especially through *on-screen text*. We manually inspect each video and assign a `visual_hint` attribute with one or multiple values: {*no text*, *irrelevant text*, *transcription*, *title/name*, *visual contains solution*}. The last and the most severe category (labeled as `visual_cheat`) denotes cases where the visual stream collapses the intended reasoning problem into simple OCR or object detection instead of genuine multimodal understanding. For `visual_cheat` clips, we omit the visual input when evaluating audio–visual models.

These measures together ensure that we are evaluating MLLMs' genuine multimodal understanding abilities without text or visual shortcuts.

2.3 Question Types

Q&As are categorized into seven question types by human verifiers, ensuring consistent labeling while allowing contributors to draft Q&As creatively without prior constraints.

「The first two types ask for the immediate information of the clip: what is heard or said. It reflects the most basic level of multimodal comprehension.」

Audio Analysis focuses on what the sound alone reveals: prosody, rhythm, style, or other audible patterns. They check whether a model can genuinely analyze audio rather than rely on the text. Example: *What kind of musical processing was applied to the human voice?*

Language Analysis tests recognition and parsing of the spoken words, including what is said and how it functions within a sentence or conversation. The answer comes from the linguistic content of the clip. Example: *What one thing does the speaker claim about himself?*

「The middle three types move beyond literal content to how humans grasp meaning, feeling, and humor in a clip. It often requires educated thinking.」

Contextual Inference evaluates whether a model can understand the situation behind the clip: what the speaker intends, what they mean, or what is happening in the scene. The correct answer comes from interpreting the context, not just rephrasing the words. Example: *Which situation best matches the tone and intent of this moment?*

Model \ QT	meme-full								meme-main							
	A	L	C	E	H	U	W	AVG	A	L	C	E	H	U	W	AVG
Audio LLMs																
SALMONN	32.1	34.7	34.4	42.5	35.3	28.2	19.8	30.6	31.1	30.1	28.1	36.2	23.5	27.2	20.7	26.6
Qwen2-Audio	51.8	52.0	41.7	49.3	46.6	41.8	25.1	41.0	42.2	47.6	35.9	44.8	28.2	36.6	22.1	34.2
Audio Reasoner	39.3	42.6	33.2	47.9	38.2	38.1	23.5	35.1	35.6	44.6	29.0	39.7	32.1	34.8	23.7	32.1
Kimi-Audio	50.0	60.5	47.8	61.6	52.6	49.2	35.2	48.5	44.4	56.3	41.6	55.2	31.8	44.4	33.7	42.1
Audio Flamingo 3	46.4	62.9	47.8	**65.8**	55.6	**53.8**	28.6	48.7	40.0	58.3	40.5	**62.1**	36.5	**48.1**	26.4	41.7
Step-Audio-2	48.2	34.7	42.9	38.4	52.6	42.1	26.4	39.3	37.8	24.3	34.1	27.6	31.8	35.2	24.5	30.3
MiDashengLM	57.1	66.9	50.9	60.3	**58.6**	44.6	30.4	49.1	**48.9**	63.1	43.8	55.2	**38.8**	39.5	29.8	42.4
MiMo-Audio	**60.7**	75.8	50.0	60.3	57.9	51.8	37.9	53.1	51.1	74.8	43.8	51.7	**38.8**	45.1	33.7	45.7
Music Flamingo	55.4	63.7	50.9	64.4	55.6	51.3	31.3	50.0	44.4	58.3	43.2	58.6	**38.8**	46.3	28.8	42.8
GPT-4o Audio †	58.9	**77.4**	**54.9**	**65.8**	57.1	52.3	**48.9**	**57.1**	**48.9**	**75.7**	**46.5**	60.3	37.6	46.9	**45.6**	**50.1**
Audio-Visual LLMs																
Baichuan-Omni	38.5	50.0	45.2	52.8	47.4	38.7	26.0	40.9	31.7	43.7	37.6	45.6	27.1	32.7	23.0	33.1
+ visual input	34.6	58.7	47.3	54.8	47.7	46.7	34.2	45.7	29.3	54.0	41.8	51.7	29.8	40.1	32.7	39.6
MiniCPM	55.4	48.4	45.5	46.6	48.9	45.6	43.4	46.5	46.7	42.7	38.4	36.2	29.4	40.1	41.1	39.3
+ visual input	57.1	53.2	45.1	56.2	51.1	49.0	44.2	48.8	48.9	47.6	36.8	48.3	28.2	45.3	42.0	41.6
Phi-4-Multimodal	44.6	57.3	43.8	60.3	46.6	39.0	25.6	42.1	40.0	54.4	37.3	51.7	29.4	32.1	24.5	35.6
+ visual input	42.9	62.1	48.4	61.6	48.5	41.8	31.4	45.8	33.3	55.3	41.8	55.2	28.6	34.2	29.9	38.3
Qwen2.5-Omni	50.0	66.9	45.5	72.6	63.9	45.1	32.6	49.7	40.0	64.1	37.8	67.2	48.2	40.7	32.7	43.5
+ visual input	62.5	85.4	62.3	75.3	63.9	59.3	45.3	62.0	53.3	83.3	58.7	70.7	48.2	55.3	43.1	56.7
Gemma 3n E4B IT	46.4	61.3	47.8	56.2	56.4	40.0	23.3	44.2	35.6	56.3	42.7	48.3	36.5	34.0	19.7	36.4
+ visual input	51.8	66.9	50.0	60.3	60.6	46.9	35.0	50.4	42.2	65.0	44.3	56.9	45.2	41.0	31.4	43.9
Qwen3-Omni	64.3	79.0	59.6	75.3	65.4	59.5	45.8	61.0	55.6	76.7	56.5	69.0	50.6	54.3	42.8	55.4
+ visual input	64.3	84.6	61.9	75.3	64.7	60.3	48.0	62.7	55.6	82.4	58.7	70.7	49.4	56.5	44.6	57.4
Gemini 2.5 Flash †	51.8	85.4	61.6	65.8	68.2	60.0	53.3	62.9	42.2	83.3	55.7	58.6	53.6	56.2	51.0	57.2
+ visual input	56.4	82.3	69.2	78.1	72.2	60.6	63.6	68.3	47.7	79.6	66.1	72.4	60.0	57.1	62.1	63.9
- audio (text-only)	37.5	33.9	36.8	31.5	48.9	35.4	25.1	34.8	22.2	20.4	23.4	13.8	20.0	22.2	18.3	20.5
Gemini 3 Flash †	64.3	86.3	75.4	75.3	77.4	67.7	67.4	73.2	55.6	84.5	71.4	72.4	67.1	64.2	66.3	69.1
+ visual input	64.3	88.7	80.6	82.2	**84.2**	**74.4**	78.9	79.7	55.6	87.4	78.8	79.3	**76.5**	70.4	77.9	76.6
Gemini 3 Pro †	76.8	**94.4**	77.8	80.8	78.2	73.7	78.9	79.5	71.1	**94.2**	74.7	77.6	70.6	70.2	77.9	76.6
+ visual input	**80.4**	91.1	**81.1**	**86.3**	83.5	72.8	**83.6**	**81.9**	**75.6**	90.3	**80.4**	**84.5**	75.3	**71.0**	**83.1**	**80.0**

Table 1: Model performance across question types on `meme-full` and `meme-main`.

Emotion Analysis asks the model to identify the feeling expressed, based on tone, delivery, pacing, or the effect on the audience. They check whether the model can pick up how the moment is meant to feel. Example: *What will people likely feel when they first listen to this piece?*

Humor & Popularity asks why a clip became funny, iconic, or widely shared. They involve noticing unexpected reactions, exaggerated behavior, or other traits that made the moment memorable online. Example: *Which of the following is not a reason that the clip is humorous?*

「The last two types cover the human conventions and external knowledge that anchor the clip in its communicative and cultural setting.」

Usage & Application focuses on how people actually use the clip as a meme: the situations it fits, the message it conveys, or the reaction it arises. They test whether the model understands the application of it. Example: *When do people use this meme?*

World Knowledge requires information beyond the clip, such as who the speaker is, the media it comes from, or what event or reference it points to. Correct answers rely on cultural familiarity or factual background. Example: *Who performs the original version of this track?*

While we define seven question types for clarity of analysis, their boundaries are not absolute. Human interpretation is naturally subjective: for instance, an experienced musician may infer the genre or composer from a first listen (contextual inference), whereas an ordinary listener might need a program book to learn it (world knowledge). These categories should therefore be viewed as guiding notions rather than strict divisions.

3 Main Results

We evaluate 19 state-of-the-art MLLMs: ten audio-only Tang et al. (2024); Chu et al. (2024); Zhifei et al. (2025); Ding et al. (2025); Ghosh et al. (2025a); Wu et al. (2025); Dinkel et al. (2025); Xiaomi (2025); Ghosh et al. (2025b) including GPT-4o Audio Hurst et al. (2024) and nine audio-visual Li et al. (2024); Hu et al. (2024); Abouelenin et al. (2025); Xu et al. (2025a); Team et al. (2025); Xu et al. (2025b) including Gemini 2.5 Flash Comanici et al. (2025), Gemini

Model \ Category	meme-full: ST				meme-main: ST				meme-full: Lang						meme-main: Lang					
	Sp	So	Mu	Sfx	Sp	So	Mu	Sfx	en	zh	ja	ko	fa	nan	en	zh	ja	ko	fa	nan
Audio LLMs																				
SALMONN	32.6	25.1	25.4	33.3	26.7	22.8	25.5	33.0	30.0	32.3	25.0	35.7	35.4	27.2	26.7	23.8	20.3	36.7	26.8	27.5
Qwen2-Audio	43.5	40.8	37.3	30.5	35.9	34.7	32.4	26.8	40.1	51.1	33.8	51.8	29.2	33.6	34.3	39.7	30.2	51.0	26.8	29.0
Audio Reasoner	36.6	35.1	33.9	27.6	32.7	34.5	30.4	26.8	35.8	37.2	32.9	41.1	31.9	29.5	34.1	31.2	29.7	38.8	32.5	26.9
Kimi-Audio	51.3	48.0	42.4	**40.0**	42.8	44.1	38.2	**39.2**	51.3	50.8	42.1	46.4	**47.9**	39.6	46.3	37.0	40.6	42.9	**46.3**	36.3
Audio Flamingo 3	53.1	45.6	43.2	33.3	44.6	41.4	**40.2**	28.9	50.0	54.1	43.4	51.8	35.4	40.1	45.5	40.7	37.5	46.9	34.1	35.8
Step-Audio-2	42.5	38.6	28.8	33.3	31.3	32.4	22.5	29.9	39.3	42.9	35.5	37.5	43.8	32.3	32.9	24.3	28.1	32.7	36.6	27.5
MiDashengLM	54.4	44.4	39.0	36.2	46.2	40.0	35.3	34.0	51.0	57.9	31.6	53.6	41.7	38.2	46.3	44.4	28.1	49.0	41.5	35.2
MiMo-Audio	58.0	52.6	**44.1**	34.3	49.8	48.3	36.3	30.9	57.0	**60.2**	46.1	48.2	39.6	40.1	52.2	**47.6**	39.1	44.9	39.0	33.7
Music Flamingo	55.5	48.5	35.6	35.2	46.8	43.4	32.4	32.0	51.0	57.9	48.7	**60.7**	39.6	35.5	47.1	44.4	40.6	**57.1**	36.6	31.6
GPT-4o Audio †	**62.5**	**59.4**	42.4	37.1	**54.5**	**55.6**	36.3	34.0	**65.0**	59.0	**57.3**	57.1	41.7	**41.2**	**60.0**	46.0	**52.4**	55.1	34.1	**36.5**
Audio-Visual LLMs																				
Baichuan-Omni	45.7	38.6	28.4	29.0	36.8	32.9	22.0	25.0	40.7	51.3	40.3	40.0	38.3	28.3	34.6	38.5	34.4	41.7	31.7	22.9
+ visual input	49.3	45.3	35.9	35.0	41.8	42.4	30.7	33.3	47.9	50.4	42.7	50.0	34.0	35.4	44.4	37.6	39.1	51.0	32.5	31.2
MiniCPM	48.6	42.7	44.4	41.9	39.2	37.2	42.6	39.2	44.0	50.8	39.5	35.7	39.6	46.8	38.4	36.5	32.8	30.6	36.6	44.8
+ visual input	51.6	43.9	45.7	43.8	42.8	37.2	43.0	40.2	49.0	52.6	38.2	41.1	41.7	47.0	43.9	37.6	31.2	38.8	39.0	44.0
Phi-4-Multimodal	47.3	37.4	31.4	29.5	40.0	32.4	26.5	26.8	45.0	48.5	39.5	42.9	33.3	30.9	40.8	36.0	34.4	40.8	34.1	26.4
+ visual input	51.0	40.4	35.0	34.3	41.9	36.6	31.7	28.7	46.7	51.7	42.1	39.3	42.6	34.3	43.1	36.2	37.5	36.7	35.0	29.6
Qwen2.5-Omni	54.4	46.8	37.3	40.0	46.4	43.4	35.3	37.1	50.7	56.4	43.4	58.9	35.4	38.7	45.9	43.9	42.2	55.1	34.1	36.3
+ visual input	67.1	59.1	50.4	48.0	60.9	55.9	47.5	45.7	63.7	72.0	59.2	66.1	44.7	48.8	60.8	62.6	53.1	69.4	37.5	46.0
Gemma 3n E4B IT	52.8	33.9	29.7	24.8	44.0	27.6	28.4	18.6	48.3	50.8	31.6	55.4	45.8	25.8	43.5	34.9	21.9	53.1	46.3	21.8
+ visual input	57.6	43.3	37.6	32.4	49.6	39.3	35.6	29.8	54.2	53.6	47.4	64.3	48.9	34.3	49.2	39.9	40.6	63.3	52.5	31.2
Qwen3-Omni	65.9	61.4	46.6	46.7	59.9	57.2	41.2	44.3	63.2	70.3	51.3	71.4	47.9	45.6	59.8	61.9	45.3	71.4	41.5	41.5
+ visual input	67.4	59.1	54.7	49.0	61.5	55.2	50.5	46.8	64.3	72.3	56.6	67.9	44.7	51.2	61.2	63.1	51.6	71.4	37.5	47.6
Gemini 2.5 Flash †	68.1	58.5	47.5	55.8	61.9	53.8	42.2	54.2	71.3	64.2	55.3	65.5	47.9	51.6	67.8	52.1	51.6	64.6	43.9	48.2
+ visual input	72.4	67.8	55.7	58.1	67.3	65.5	53.0	55.7	75.8	71.8	68.0	67.9	42.6	56.1	72.7	63.0	66.7	71.4	37.5	53.4
- *audio (text-only)*	38.3	30.4	28.8	27.6	21.6	17.9	17.6	21.6	31.1	44.0	35.5	32.1	37.5	25.8	18.9	21.2	23.4	22.4	26.8	16.6
Gemini 3 Flash †	78.7	74.3	57.6	55.2	74.3	69.7	57.8	53.6	83.0	78.6	77.6	67.9	43.8	57.1	80.4	71.4	73.4	65.3	39.0	56.5
+ visual input	82.7	78.4	74.6	69.5	79.4	74.5	74.5	67.0	84.9	82.7	85.5	**78.6**	56.2	71.9	83.9	76.7	82.8	**75.5**	51.2	70.5
Gemini 3 Pro †	83.2	**82.5**	66.9	65.7	80.5	80.0	65.7	62.9	**88.3**	82.0	82.9	67.9	**59.6**	68.2	**87.5**	76.3	82.8	65.3	55.0	66.3
+ visual input	**84.7**	81.3	**76.1**	**72.4**	**82.6**	**79.3**	**76.2**	**71.1**	86.6	**84.6**	**90.8**	71.4	58.3	**75.0**	85.8	**79.4**	**92.2**	71.4	**56.1**	**74.5**

Table 2: Model performance across sound types and languages for `meme-full` and `meme-main`.

3 Flash DeepMind (2025a), and Gemini 3 Pro DeepMind (2025b). The exact checkpoints and code repositories if public are documented in Table 4. All models receive the same prompt: question with randomly shuffled options starting from letter *A*. Audio is converted to monaural 16kHz, and video is converted to 360p resolution (aspect ratio preserved) at one frame per second. For GPT and Gemini, we create temporary copies of the audio and video files (*tmp.wav/mp4*) and upload them as bytes, preventing the models from cheating with file names. AVMeme Exam is organized that each Q&A belongs to a specific question type, and each clip is associated with both a sound type and a language. Leveraging this structure, we report results by question type in Table 1 and by sound category and language in Table 2, enabling a multi-axis analysis of model behavior and relative strengths and weaknesses. We summarize the key findings below. Supplementary results are shown in Appendix C.

i. Overall performance. For both test splits, more recent models (lower in the tables) achieve higher performance. Closed-source commercial models (†) significantly outperform open-source models. Gemini 3 Pro is overall the best model in our benchmark, which achieves an average accuracy of 76.6 (audio-only) and 80.0 (audio–visual) on `meme-main`. The strongest open-source model, Qwen3-Omni, reaches only 55.4 and 57.4, respectively, and still lags behind Gemini 2.5 Flash, despite being considerably stronger than Gemma 3n, Qwen2.5-Omni, and phi-4-Multimodal. For the audio-only, GPT-4o Audio remains the strongest, achieving an average accuracy of 50.1 on `meme-main`. Across all model families, audio-visual models consistently outperform their audio-only counterparts, confirming the benefit of visual signals for interpreting complex Internet clips, where video provides cues about object motion, scene setting, speaker identity, and interactions that are difficult to infer from audio alone. Finally, accuracies on `meme-main` are often 5–10% lower than on `meme-full` across all models and categories, showing that removing easy questions yields a more challenging evaluation that better tests genuine multimodal understanding.

ii. Content versus context and culture understanding. Across nearly all models, Language Analysis (L) is the easiest category, with leading systems such as Gemini 2.5 Flash, Qwen3-Omni, and GPT-4o Audio reaching 76–83% on `meme-main`. Audio Analysis (A) is noticeably harder: even some strongest models are around or below 60%, reflecting the challenge of interpreting prosody, musical style, and other non-linguistic acoustic cues

not captured in speech transcription. Performance drops further for higher-level question types such as Contextual Inference (C), Humor & Popularity (H), Usage & Application (U), and World Knowledge (W). Among these, W is the most difficult for most models, typically falling in the 20–55% range, while U also remains low across both audio-only and audio-visual models. Even top audio-visual systems show 15–30% declines when moving from L to these deeper understanding tasks. Interestingly, Gemini 3 Flash and Pro show significantly higher accuracy on World Knowledge than all other models including 2.5 Flash, with 3 Pro exceeding Qwen3-Omni by over 30% on W. This advantage may be related to broader exposure to the multimedia world, e.g. YouTube videos and associated metadata, during training. Nevertheless, Context and Usage questions still remain the hardest, with performance declining sharply from L (above 90%). Taken together, a finding is consistent across all models: while MLLMs excel in surface semantic parsing, contextual, cultural, and pragmatic understanding still remains challenging.

iii. Challenge of textless sounds. Table 2 highlights another consistent bottleneck. Models perform best on speech (Sp), good on songs (So), worse on music (Mu), and worst on sound effects (Sfx). The last two types typically have no spoken words. Even the strongest audio LLMs, such as GPT-4o and Music Flamingo, reach only 35 to 45% on music and sound effects, compared with around 60 to 65% on speech and songs. Certain models exhibit specialized strengths; for example, Kimi-Audio and Audio Flamingo 3 achieve higher accuracy than GPT-4 on sound effects or music despite being weaker overall. These patterns suggest that current MLLMs handle textual semantics well, but their ability to interpret music and sound effects without linguistic structure falls largely behind.

iv. Challenge of lesser-known languages. Performance also varies significantly across languages. English and Chinese are scored best for nearly all models. Qwen3-Omni is particularly strong in Chinese, outperforming Gemini 2.5 Flash, while the two models perform similarly in English. Japanese, Korean, and Persian are substantially more difficult, with leading models frequently dropping to the 35–55% range and audio-only LLMs sometimes falling below 30% . Even Gemini 3 pro only scores 56.1% in Persian on `meme-main`. The *nan* columns in Table 2 denote clips without language, mainly music and sound effects, whose performance is comparable to or even lower than Persian depending on the model. Adding visual input improves accuracy only marginally for these lesser-known languages and non-verbal sounds.

v. MLLMs vs humans in meme comprehension. To provide models a human reference, we recruit 20 participants, 10 native English speakers and 10 native Chinese speakers aged 18–35, all self-reported frequent Internet users. Each participant answers only clips in their native language and culture, or clips without language, without web search or collaboration. This narrows to 750 memes in total and 37.5 memes per person. Notice that this setting ***disadvantages*** humans, as unlike LLMs trained on web data, no individual has seen all videos. Yet the strong performance of individual humans despite this disadvantage ***strengthens*** our takeaway. Before showing them Q&A, we first ask participants whether they have seen and are familiar with each meme video, and report results grouped by human familiarity level in Figure 2 (More details in Appendix B). As expected, humans perform much better on their familiar memes, outperforming ***all*** open-source MLLMs and comparable to Gemini 2.5 Flash. Qwen3-Omni, the strongest open-source model, only slightly outperforms a human individual when we count the memes they have not seen. But even on unseen memes, humans still outperform the remaining open-source MLLMs, showing the continued advantage of human cultural grounding in meme understanding.

vi. Does thinking help? We compare Gemini 2.5/3 Flash with minimal thinking and Gemini 3.0 Pro with low and high thinking levels (Table 5 in the Appendix). At an aggregate level, thinking appears beneficial: high thinking improves AVG accuracy by 5.4% in the audio-only setting and 1.7% in the audio-visual setting on `meme-main`. However, category-level trends are highly uneven, with gains concentrated in explicit recognition tasks. In audio-only mode, Audio improves from 57.8 to 71.1 and Language from 83.5 to 94.2, while World Knowledge also increases from 68.8 to 77.9, with more *"Aha, I know the correct video!"* in longer thinking. However, longer thinking offers little or even adverse benefit for pragmatic and cultural tasks. In the audio-visual setting, accuracy stagnates for Context

Model	Evaluation Settings			
	"This is a meme."+Q	Meme Name+Q	Keep videos for `visual_cheat`	Default (minimal hint or cheat)
Qwen2-Audio	35.4	44.6	*n.a.*	34.4
Audio Flamingo 3	41.5	50.0	*n.a.*	41.7
Music Flamingo	42.9	51.3	*n.a.*	42.8
Baichuan-Omni	34.1	46.1	*n.a.*	33.1
+ visual input	41.3	49.4	45.3 (80.3)	39.6 (40.0)
Phi-4-Multimodal	34.9	42.5	*n.a.*	35.6
+ visual input	42.4	49.5	43.6 (74.4)	38.3 (35.0)
Gemma 3n E4B IT	38.0	53.6	*n.a.*	36.4
+ visual input	44.3	57.1	50.4 (84.6)	43.9 (38.5)
Qwen3-Omni	53.6	67.1	*n.a.*	55.4
+ visual input	58.7	67.8	61.7 (81.2)	57.4 (49.6)

Table 3: The effect of text or visual *hint* and *cheat* in multimodal evaluation. The right-most column is our default setting with minimal answer shortcuts. Results from `meme-main`.

(80.4 → 80.4) and Humor (75.3 → 75.3), while Usage drops (72.2 → 71.0). In audio-only mode, Emotion (81.0 → 77.6) and Humor (71.8 → 70.6) also decline. Appendix C.3 shows examples of both effective and misguided thinking.

vii. Ablations. In Table 3, we ablate the evaluation setting to show how easily multimodal evaluation can drift away from genuine understanding. First, when we provide the name of the meme (written by the contributor, which often contains the song title, character name, or the meme itself), model accuracy increases significantly (≈10%) because many solutions become recallable from the LLM training text, bypassing audio/video understanding. Second, adding the soft hint *This is a meme* before the question slightly boosts accuracy (1–2%) for most models, as they begin to treat the clip like a familiar Internet template and to think more (especially in ironic ways). While not cheating, we avoid this prefix so that models do not mistake ordinary videos as memes. Finally, allowing visual-cheat conditions, i.e., keeping videos that disclose key information, causes substantial accuracy inflation, 40% or more on the visual-cheatable subset (numbers in parentheses), because on-screen text, titles, or visual objects directly reveal the answer. These observations justify our strict removal of both text and visual shortcuts for fair multimodal evaluation.

Discussion. Taken together, our results and analyses reveal a consistent pattern: MLLMs perform reliably on content-level questions, particularly when rich linguistic cues are present. Their performance drops sharply when understanding requires contextual thinking or cultural grounding, and on textless audio or less common languages. Even the strongest models like Gemini 3 Pro exhibit the largest errors in these categories, and longer thinking often does not help. This underscores that pragmatic and cultural comprehension remain open challenges for multimodal intelligence.

Lesson. Beyond increasing exposure to Internet audio and video, our findings suggest that multimodal intelligence should be more firmly grounded in the human world. Understanding audio–visual signals is not merely about perceiving isolated sounds or images, but about connecting them to how people interpret, feel about, and use them within context and culture. This points to the importance of richer, human-centered annotations that capture emotion, usage, and cultural meaning, along with training objectives that prioritize alignment with human interpretation over surface-level recognition.

4 Conclusion

We presented AVMeme Exam, a multimodal, multilingual, and multicultural benchmark that evaluates whether MLLMs can understand not only what is heard or seen, but also how meaning is constructed through context, emotion, usage, and shared cultural knowledge. Our analysis reveals consistent limitations: models perform substantially worse on textless audio, struggle with contextual and cultural reasoning, and often fail to align with human interpretations. These findings highlight a key gap between current multimodal capabilities and the human expectations of intent-aware, culturally grounded understanding. In future work, we plan to expand AVMeme Exam to broader cultures, linguistic communities, and media ecosystems, creating a more globally representative testbed for multimodal understanding. We also aim to develop methods that could address the challenges revealed by our benchmark, advancing progress toward human-aligned multimodal intelligence.

Limitations. While our AVMeme Exam spans diverse languages and media, several limitations remain. First, cultural coverage is inevitably shaped by our contributors, who are highly-educated researchers aged 22–35 and whose perspectives do not fully represent the global, intergenerational landscape of meme culture. Second, meme meaning shifts over time. Our annotations capture contemporary (by the end of 2025) interpretations and cannot anticipate future cultural drift (which is impossible). Third, most current MLLMs impose strict audio and video length limits, requiring clips to be truncated to 30 seconds, which may omit context essential for real-world understanding. Fourth, our evaluation focuses on controlled multiple-choice Q&A on single clips, whereas real-world applications often require multi-turn dialogue, personalization, and open-ended scenarios. Fifth, meme interpretation is inherently subjective. Although viewers may converge on a majority understanding, alternative readings are equally valid and cannot be fully reflected in a standardized benchmark. Taken together, these considerations suggest that AVMeme Exam should be viewed as a reference benchmark, diagnostic and comparative, rather than an absolute ground truth for human multimedia culture. Nevertheless, the lesson from AVMeme Exam, that current multimodal AIs remain weaker at audio–visual understanding than text and at contextual & cultural thinking than surface content, still holds.

Ethical Considerations

AVMeme Exam is designed as a research benchmark to study multimodal understanding of audio–visual memes, with careful attention to ethical considerations in data collection, annotation, and verification. All audio–visual clips are human-curated by contributors who personally recognize and use these memes, rather than being automatically scraped, and are drawn exclusively from publicly available online videos, with no private, paywalled, or confidential content included. To mitigate potential harm, we prohibit political materials and explicit depictions of sexual, violent, hateful, criminal, or drug-related content, and clearly annotate clips with implicit or contextual sensitivity using sensitivity tags to enable filtering and informed use. The benchmark is intended for research and evaluation, not for safety-critical or high-stakes deployment, and any biases or cultural limitations it reveals should be understood as limitations of current models and of the dataset itself, motivating further work toward more human-aligned, culturally grounded multimodal systems.

Human evaluation was conducted under an Institutional Review Board (IRB) protocol. All participants completed the evaluation online in approximately 30 minutes and were compensated $15.

Acknowledgement

We acknowledge the use of Nano Banana Pro to draw meme figures for illustrative purposes. LLMs were also used in the text-cheat detection procedure described in the paper, and for syntax and grammar fixes. All final data verification and paper writing decisions were made by the human authors.

This work is funded by the National Institutes of Health (NIH-NIDCD) and a grant from Marie-Josee and Henry R. Kravis.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3394–3418, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.169. URL https://aclanthology.org/2025.acl-long.169/.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Gemini 3 flash best for frontier intelligence at speed. https://deepmind.google/models/gemini/flash, Dec 2025a.

Google DeepMind. Gemini 3 pro best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro, Nov 2025b.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4763–4784, 2025.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=FjByDpDVIO.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, et al. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*, 2025b.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information?, 2024a. URL https://arxiv.org/abs/2412.02611.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=nBZBPXdJlC.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao

Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12136–12140. IEEE, 2024.

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

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. Are vision-language models safe in the wild? a meme-based benchmark study. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30533–30576, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1555. URL https://aclanthology.org/2025.emnlp-main.1555/.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19108–19118, 2022.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Transactions on Intelligent Systems and Technology*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, tianrui wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, EngSiong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=fgmrBJemlQ.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*, 2025.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card, 2024.

OpenAI. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/, 11 2025.

Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=4ZAX1NT0ms.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9763–9771, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6890–6894. IEEE, 2024.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

xAI. Grok 4. https://x.ai/news/grok-4, Feb 2025.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://github.com/XiaomiMiMo/MiMo-Audio.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. MemeReaCon: Probing contextual meme understanding in large vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3559–3582, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.176. URL https://aclanthology.org/2025.emnlp-main.176/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23840–23862, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1216. URL https://aclanthology.org/2025.emnlp-main.1216/.

A Collection & Verification Procedures

AVMeme Exam was curated by 27 audio and NLP researchers affiliated with universities and companies in the US and the UK, aged between 22 and 35, the majority of whom are current PhD students. To coordinate collection, all raw entries are first recorded in a shared Google Sheet, organized into separate sheets by language and sound category, which helps prevent duplicate submissions across different contributors. The complete raw data collection process took approximately three months.

All contributors participated voluntarily and with informed consent. They are listed as co-authors of the paper. To mitigate potential ethical risks in data curation, contributors were instructed to rely only on publicly available content, avoid political or religious material, and exclude explicit sexual, violent, or hateful content. No personal data of contributors was collected or released.

All contributors were given and agreed to follow the collection and annotation guidelines:

- **Meme selection:** Memes are chosen based on broad public recognition and sourced from original or highly circulated online videos, preferably from YouTube. A view count of at least one million is a soft requirement.
- **Clip extraction:** Each meme is trimmed to a short segment (1–30 seconds). At most two distinct memes may be extracted from the same source video.
- **Audio-centric requirement:** Memes must be identifiable primarily from audio alone. Clips whose meaning relies heavily on visual information are excluded.
- **Multilingual and cross-cultural coverage:** Memes may span multiple languages and cultures, and are explicitly labeled while retaining their original sound category.
- **Standardized metadata:** Each entry includes structured annotations such as source URLs, temporal boundaries, original appearance date, language, sound category, transcription (when applicable), summary, emotion, sensitivity, and typical usage.
- **Content safeguards:** Political and religious content, explicit sexual material, and graphic violence are excluded. Mild or implicit sensitive content is allowed only with explicit sensitivity labels.
- **Question design:** Each meme is paired with a unique multiple-choice question targeting different levels of understanding. Contributors are encouraged to design challenging and creative questions, and are not required to assign questions to any predefined categories. Question types are labeled later during verification by independent annotators.
- **Leakage prevention:** Questions and answer choices must not disclose identifying information (e.g., titles, names, slogans, or sources) to prevent textual shortcuts and ensure genuine audio-visual evaluation.

For leakage prevention, we label Q&As that can be correctly guessed by all of Gemini 2.5 Flash, Grok 4, and GPT-5.1 in text-only settings, with prompt "*without audio, guess the best option*.", and return them to the original contributors for revision. This process is repeated for two rounds.

AVMeme Exam is verified by nine human verifiers. To reduce bias and confirmation effects, we assign each meme to verifiers who did not author the corresponding entry whenever possible. For less common languages (e.g., Persian and Telugu), where linguistic expertise is limited, original contributors assist in verification to ensure correctness.

The verification process is conducted in a distributed manner through a web-based interface. For each meme, verifiers review all annotated attributes and the multiple-choice Q&A. For each attribute, verifiers select one of several actions: *Ok*, *Wrong*, or *Wrong, but I can fix*. Identified issues are either corrected directly by the verifier (if just a simple typo or video-question mismatch) or returned to the original contributor for revision.

After validating all metadata fields, verifiers assign the question to one of the predefined question types. Only after completing all checks and assigning a question type can a verifier

Model	Checkpoint / Version	Size	Code Repository
SALMONN	SALMONN-13B	13B	https://huggingface.co/tsinghua-ee/SALMONN
Qwen2-Audio	Qwen2-Audio-7B	7B	https://huggingface.co/Qwen/Qwen2-Audio-7B
Audio-Reasoner	Audio-Reasoner-7B	7B	https://huggingface.co/zhifeixie/Audio-Reasoner
Kimi-Audio	Kimi-Audio-7B-Instruct	7B	https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct
Audio Flamingo 3	audio-flamingo-3-hf	7B	https://huggingface.co/nvidia/audio-flamingo-3
Step-Audio-2	Step-Audio-2-mini	8B	https://huggingface.co/stepfun-ai/Step-Audio-2-mini
MiDashengLM	midashenglm-7b-0804-fp32	7B	https://huggingface.co/mispeech/midashenglm-7b-0804-fp32
MiMo-Audio	MiMo-Audio-7B-Instruct	7B	https://huggingface.co/XiaomiMiMo/MiMo-Audio-7B-Instruct
Music Flamingo	music-flamingo-hf	7B	https://huggingface.co/nvidia/music-flamingo-hf
GPT-4o Audio	gpt-4o-audio-preview-2025-06-03	-	-
Baichuan-Omni	Baichuan-Omni-1.5	7B	https://github.com/baichuan-inc/Baichuan-Omni-1.5
MiniCPM	MiniCPM-o-2_6	8B	https://huggingface.co/openbmb/MiniCPM-o-2_6
phi-4-Multimodal	Phi-4-multimodal-instruct	5.6B	https://huggingface.co/microsoft/Phi-4-multimodal-instruct
Qwen2.5-Omni	Qwen2.5-Omni-7B	7B	https://huggingface.co/Qwen/Qwen2.5-Omni-7B
Gemma 3n E4B IT	gemma-3n-E4B-it	8B	https://huggingface.co/google/gemma-3n-E4B-it
Qwen3-Omni	Qwen3-Omni-30B-A3B-Instruct	30B	https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
Gemini 2.5 Flash	gemini-2.5-flash	-	-
Gemini 3 Flash	gemini-3-flash-preview	-	-
Gemini 3 Pro	gemini-3-pro-preview	-	-

Table 4: Details about model checkpoints and repositories. We follow the provided inference code with default configurations for each model.

Gemini (thinking_level)	meme-full								meme-main							
	A	L	C	E	H	U	W	AVG	A	L	C	E	H	U	W	AVG
2.5 Flash (minimal)	51.8	85.4	61.6	65.8	68.2	60.0	53.3	62.9	42.2	83.3	55.7	58.6	53.6	56.2	51.0	57.2
+ visual input	56.4	82.3	69.2	78.1	72.2	60.6	63.6	68.3	47.7	79.6	66.1	72.4	60.0	57.1	62.1	63.9
3 Flash (minimal)	64.3	86.3	75.4	75.3	77.4	67.7	67.4	73.2	55.6	84.5	71.4	72.4	67.1	64.2	66.3	69.1
+ visual input	64.3	88.7	80.6	82.2	84.2	74.4	78.9	79.7	55.6	87.4	78.8	79.3	76.5	70.4	77.9	76.6
3 Pro (low)	66.1	85.5	74.1	84.9	79.7	70.3	70.0	74.9	57.8	83.5	71.4	81.0	71.8	66.0	68.8	71.2
+ visual input	76.8	89.5	82.4	86.3	82.7	74.9	77.9	80.9	71.1	89.3	80.4	82.8	75.3	72.2	77.3	78.3
3 Pro (high)	76.8	94.4	77.8	80.8	78.2	73.7	78.9	79.5	71.1	94.2	74.7	77.6	70.6	70.2	77.9	76.6
+ visual input	80.4	91.1	81.1	86.3	83.5	72.8	83.6	81.9	75.6	90.3	80.4	84.5	75.3	71.0	83.1	80.0

Gemini (thinking_level)	meme-full: ST				meme-main: ST				meme-full: Lang						meme-main: Lang					
	Sp	So	Mu	Sfx	Sp	So	Mu	Sfx	en	zh	ja	ko	fa	nan	en	zh	ja	ko	fa	nan
2.5 Flash (minimal)	68.1	58.5	47.5	55.8	61.9	53.8	42.2	54.2	71.3	64.2	55.3	65.5	47.9	51.6	67.8	52.1	51.6	64.6	43.9	48.2
+ visual input	72.4	67.8	55.7	58.1	67.3	65.5	53.0	55.7	75.8	71.8	68.0	67.9	42.6	56.1	72.7	63.0	66.7	71.4	37.5	53.4
3 Flash (minimal)	78.7	74.3	57.6	55.2	74.3	69.7	57.8	53.6	83.0	78.6	77.6	67.9	43.8	57.1	80.4	71.4	73.4	65.3	39.0	56.5
+ visual input	82.7	78.4	74.6	69.5	79.4	74.5	74.5	67.0	84.9	82.7	85.5	78.6	56.2	71.9	83.9	76.7	82.8	75.5	51.2	70.5
3 Pro (low)	79.2	73.1	64.4	63.8	75.5	70.3	60.8	60.8	83.3	79.3	68.4	67.9	62.5	63.1	82.4	72.0	67.2	67.3	58.5	59.6
+ visual input	83.5	79.5	78.6	69.5	80.6	79.3	76.2	67.0	86.9	85.0	82.9	71.4	52.1	73.6	86.6	80.4	84.4	69.4	46.3	70.8
3 Pro (high)	83.2	82.5	66.9	65.7	80.5	80.0	65.7	62.9	88.3	82.0	82.9	67.9	59.6	68.2	87.5	76.3	82.8	65.3	55.0	66.3
+ visual input	84.7	81.3	76.1	72.4	82.6	79.3	76.2	71.1	86.6	84.6	90.8	71.4	58.3	75.0	85.8	79.4	92.2	71.4	56.1	74.5

Table 5: Gemini family performance across question types, sound types, and languages.

proceed to the next sample. This structured, attribute-level verification ensures annotation consistency, correctness, and alignment with the intended evaluation goals.

Finally, one human verifier aggregates all verified entries into a single file, manually reviews the merged results to remove duplicate entries, and performs a final pass to ensure consistency in question type assignments across the dataset.

B Human Evaluation Details

We conduct a controlled human evaluation to establish a human reference for the AVMeme Exam, following the University's Institutional Review Board (IRB) protocol. A total of 20 participants were recruited, including 10 native English speakers and 10 native Chinese speakers, all aged 18–35 and self-reported frequent users of online video platforms (e.g., YouTube, Bilibili, TikTok). Participants grew up in the U.S. or China and were residing in the U.S. at the time of the study, due to practical recruitment constraints. All participants completed the study individually, without web search, collaboration, or external assistance.

Each evaluation session lasted approximately 30 minutes, and participants were compensated with a $15 prepaid card upon completion. Each participant was assigned 37 or 38

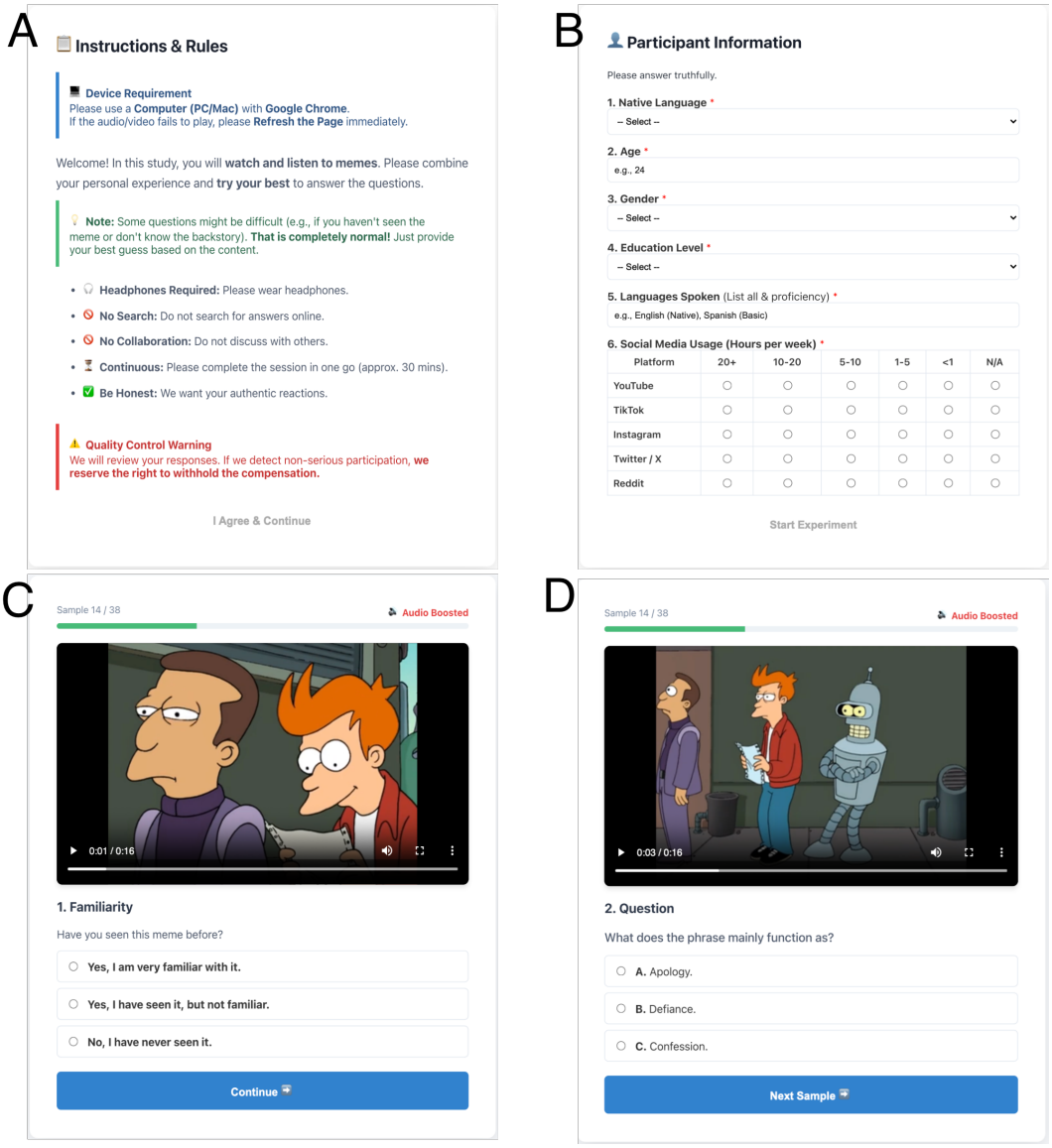


Figure 6: Human evaluation interface. **A**. Instructions and rules shown before the study, specifying device requirements, headphone use, no-search/no-collaboration constraints, and quality control policies to ensure reliable human judgments. **B**. Participant information and background survey collected anonymously and linked only via a participant ID, including demographics, language background, and social media usage. **C**. Familiarity check presented before showing the Q&A, where participants indicate whether they have previously seen the meme clip. **D**. Q&A, in which participants watch and listen to the meme clip and answer a multiple-choice question.

audio-visual meme clips, resulting in 750 total human-evaluated samples. To ensure linguistic and cultural appropriateness, participants only evaluated clips in their native language and cultural context, as well as clips without spoken language (e.g., music or sound effects). Screenshots of the session are shown in Figure 6.

Before the main evaluation, participants completed a background survey collecting demographic information (age range, native language), Internet usage frequency, and self-reported familiarity with online memes and viral media. Each participant was assigned a unique anonymous identifier, and no personally identifiable information was collected. Survey responses were linked only to these anonymous identifiers and were used solely to describe the participant pool and report aggregate evaluation metadata.

For each sample, participants were first shown the video clip alone (or the audio clip for the visual cheating condition, as we enforce for LLMs), without the associated question or answer options. They were asked whether they had previously seen and were familiar with the clip. Only after recording this response were participants presented with the corresponding multiple-choice question and answer options. This two-stage design prevents participants' familiarity judgments from being influenced by question difficulty.

Participants then answered the same multiple-choice questions used for model evaluation, under identical answer formats and constraints. Human responses were recorded and later aggregated by familiarity category for comparison with model performance, as illustrated in Fig. 2. This protocol ensures that human evaluation is conducted under controlled, reproducible conditions aligned with the LLM evaluation setting.

C Supplementary Results

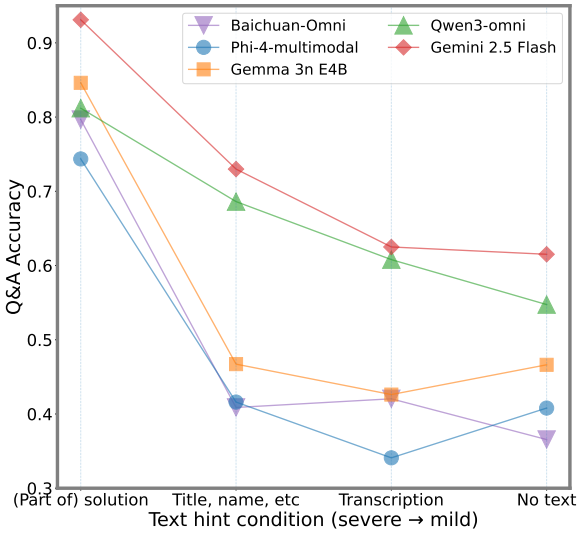


Figure 7: The effect of different levels of text hint from the visual. Results from `meme-main`.

C.1 Study on On-screen Text

Real-world videos often contain on-screen text that may partially assist or hack multimodal understanding. To analyze how such visual text affects model performance, we conduct a focused study on clips with different degrees of on-screen textual hints.

We (9 human verifiers) manually label each clip based on the strongest type of text appearing in the visual stream. Specifically, we categorize clips into five groups: (1) *visual contains (part of) solution* (`visual_cheat`) (2) *title, name, or other keywords*, which reveal identifiable entities (3) *transcription*, where the spoken content or its translation is displayed as subtitles (4) *no*

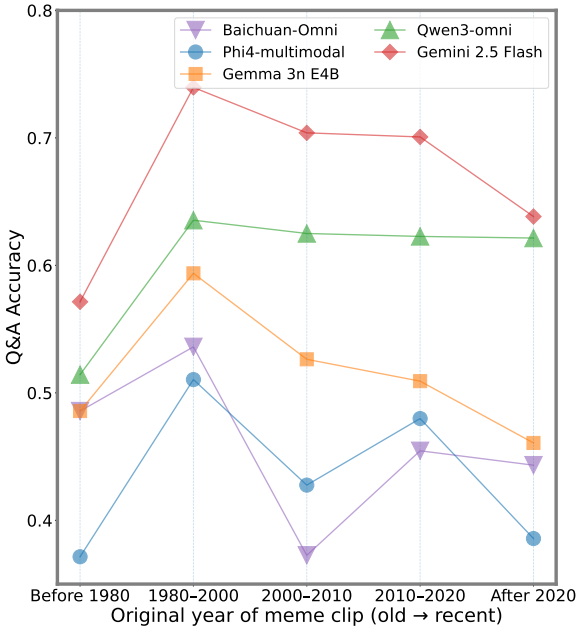


Figure 8: Model performance vs. the original year of the clips. Results from `meme-main`.

text. These categories form a spectrum of visual hints, ranging from the most severe to the mildest.

For each category, we aggregate model accuracy across all applicable clips and report the results in Figure 7. Accuracy peaks when the visual stream directly reveals the solution and decreases monotonically as text hints weaken from titles and keywords to transcription and ultimately to no text. This pattern is consistent across models, indicating that on-screen text acts as a strong shortcut. For some models, the no-text condition is slightly worse than transcription, as many textless clips involve music or sound effects, which are inherently more challenging.

C.2 Study on Meme's Year

We additionally annotate each meme clip with its original year, defined as the upload year of the source YouTube or Bilibili video, or the debut or premiere year for music, songs, and movies. We group clips into coarse temporal spans and aggregate model accuracy within each group, as shown in Figure 8.

Across models, performance peaks on memes originating between 1980 and 2000, while accuracy drops for both older (<1980) and more recent (>2020) clips. We hypothesize that this pattern reflects uneven Internet coverage: middle-era memes are well represented and repeatedly circulated online, whereas very old or very recent content is less consistently captured in training data.

C.3 Correct and Wrong Thinking Example

In addition to the quantitative results of Gemini 3 Flash and Gemini 3 Pro under low and high thinking levels in Table 5, we present examples below illustrating both effective and misguided reasoning (all in audio-visual mode), in supplement to the discussion in Section 3.vi.*Does thinking help?*

GTA wasted

Category: sound effect
Year: 2013
Language: N/A
Transcription: N/A



Summary: The game over sound effect in Grand Theft Auto.

Usage: Used when someone fails something.
Emotion: neutral
Sensitivity: none
Question: *Where will you likely be after you hear this sound?*

A. Home.
B. **Hospital.**
C. Police office.
D. Jail.
E. Another world.
F. Flight landing.

Question Type: World Knowledge

杀死那个石家庄人 (卡祖笛版本)

Category: music
Year: 2010
Language: N/A
Transcription: N/A



Summary: A short instrumental break mixing distorted guitar and kazoo, creating an ironic, mock-heroic feeling in the middle of a melancholic Chinese rock song.

Usage: Used to express absurd humor, irony, or mock-solemn moments in Chinese internet videos and memes.
Emotion: sarcastic/ironic, nostalgic, sad
Sensitivity: none
Question: *What type of character or figure would most likely appear when this distinctive instrumental interlude plays?*

A. A cartoon superhero saving the world in slow motion with exaggerated seriousness.
B. **A middle-aged man performing mundane tasks.**
C. A young couple celebrating a wedding under bright fireworks.
D. A musician playing a kazoo with a jazz song.

Question Type: World Knowledge

Gemini 3 Flash: Jail. ×

Gemini 3 Pro (low thinking): <think>**My Initial Analysis of the Sound** Okay, so the user wants me to pinpoint the typical context of this sound effect and what it usually means. That's interesting. My first instinct is to break down the sound itself. It sounds like it kicks off with a really sharp, heavy impact – think slamming a heavy door or a metal gate slamming shut. That initial "thud" really sets the stage. Then, there's that metallic ringing or clanking that follows. It's almost like the sound of metal bars shifting or settling after the initial impact. It's a very distinctive sound, almost archetypal in my experience, and that's the key to identifying its typical use and cultural significance. <\think> Another world.×

Gemini 3 Pro (high thinking): <think>**My Game Sense Kicked In** Okay, first things first, I need to figure out where we're at based on this audio clip. *Thud, hum... wait a second...* That's it! That's the *Wasted* sound effect from GTA 5! That low, ominous drone is unmistakable. Instantly, the game mechanics flood my brain. When you get "wasted" in GTA 5, the screen goes grey, the big red letters pop up, and this glorious sound confirms your digital demise. So, where do you respawn? Not in some metaphorical "other world," and definitely not at home unless you saved the game. Definitely not at the police office or jail; those are for getting "busted," not "wasted." Flight landing? Nope, no connection there. Ah! The hospital. That's it. Every time you die in GTA 5, they slap you with a medical bill and you wake up outside the nearest hospital. <\think> Hospital.✓

Gemini 3 Flash:
A middle-aged man performing mundane tasks.✓

Gemini 3 Pro (low thinking): <think>**Analysis of a Comedic Audio Cue** Alright, let's break this down. We've got a video with some music, and we need to figure out what kind of character or situation would *most likely* be associated with it. The key is the audio. It's a kazoo-like rendition of "The Final Countdown" – instantly recognizable, but intentionally goofy. My thought process immediately went to the contrast. "The Final Countdown" is an epic, dramatic song. But here, it's

being played on a ridiculous instrument. The best fit is a scenario that mirrors this incongruity, this playful subversion of expectations. Option A, "A cartoon superhero saving the world in slow motion with exaggerated seriousness," perfectly encapsulates this mock-epic vibe. It's that classic comedic juxtaposition: a serious situation rendered absurd by silly sound effects. We see this all the time in memes and parodies. This is where internet culture comes in. The sound itself is part of a *specific* meme format! Option B, a middle-aged man doing mundane tasks, is a common ironic choice for music but misses the specific energy of this song choice. It's a great example, but "The Final Countdown" has an inherent sense of building to something, so a middle-aged man washing dishes doesn't fit the 'countdown' or high-energy vibe. Option C is a straight-faced application of the audio, and there's nothing sincere about a kazoo playing at a wedding. Finally, Option D, while the video has a "musician" in it, the style is wrong (it's rock and not jazz). It's all about the tone! The goal is to maximize the humorous dissonance between the audio and the visual. If this *were* a question on a music or meme culture test, I'd say that there is only one correct answer - it has to be A. The "cartoon superhero with exaggerated seriousness" provides an immediate sense of the *mock-heroic* nature of the track. If the music *is* a key indicator, as is often the case with musical or meme questions, then A is the only choice that makes sense. Therefore, the answer must be A. <\think> A cartoon superhero saving the world in slow motion with exaggerated seriousness..×

Gemini 3 Pro (high thinking): <think>**Alright, let's break this down.** The video's playing, and my mind's immediately going: "That's *Shooting Stars* by Bag Raiders! But, wait... it's a kazoo?! And... strumming? Ok, let's analyze the question, the options, and the context here." My first thought? "This is a meme reference, but which one?" I rule out the basic f̈alling through spacevïsuals because the question's about the instrumental interlude. I quickly scan the options. Superheroes don't quite fit the meme's usual silliness. Middle-aged man? Nope, wrong vibe. Couple at a wedding? Too romantic. Then, it hits me: *Musician playing a kazoo*. Bingo! That's what the video is *literally* showing. Though there's a slight pause, "jazz song"? *Shooting Stars* isn't jazz, but maybe the question author is misinterpreting the solo nature of the kazoo as "jazzy" or just using that as a descriptor.I check myself, re-listening, and making sure the song is actually *Shooting Stars* by comparing the melody and the notes. It is! Double-checking for other meme possibilities, but it's that kazoo cover! The question's phrasing, "What type of character or figure would most likely appear..." is a bit ambiguous. However, I'm confident that it points to a literal description, and the best literal match is definitely D, even if the "jazz" bit is slightly off. The other options are distractors that do not relate to the video content. My logic? The video *shows* a musician playing a kazoo, even if the genre isn't quite right. That kazoo is the key here. Ultimately, this boils down to which option best describes the video content. And that's D, no question. <\think> A musician playing a kazoo with a jazz song.×

D More AVMeme Samples

I am your father.



Category: noisy speech
Year: 1980
Language: English

Transcription: Luke: He told me you killed him. Vader: No. I am your father. Luke: No. No. It's not true.
Summary: A man reveals to a younger man that he is his father.

Usage: Used as a parody of the famous reveal scene, often remixed as "No, I am your X."
Emotion: sad, surprised/shocked
Sensitivity: violence/gore
Question: *What will the man most likely do after he knows the truth?*
A. Fight.
B. **Escape.**
C. Surrender.
D. Stay put.
Question Type: Contextual Inference

Liszt's Hungarian Rhapsody No.2



Category: music
Year: 1851
Language: N/A
Transcription: N/A

Summary: A fast playful dance-like solo piano music.

Usage: Used as the background music often in cartoons for chase and slapstick scenes.
Emotion: happy
Sensitivity: none
Question: *On which of the following occasions will you most likely hear this music?*
A. TV ads.
B. **Cartoon.**
C. Daily news.
D. Talk show.
Question Type: Usage & Application

Windows Error



Category: sound effect
Year: 2001
Language: N/A

Transcription: N/A
Summary: Infinite pop-up errors in Windows.

Usage: The sound effect of the error is widely recognized and used for comic "fail" beats
Emotion: angry/annoyed
Sensitivity: none
Question: *Which operating system makes this sound?*
A. **Windows**
B. macOS
C. Linux
D. Android
E. iOS
Question Type: World Knowledge
visual_cheat: **True**

Start digging in your butt twin



Category: clean speech
Year: 2025
Language: en
Transcription: A: And I want you to make me a Krabby Patty. B: Start diggin' in your butt twin. A: What? Wha–
Summary: A crude absurdist catchphrase from TikTok, telling someone to "start digging in your butt, twin." Often used with SpongeBob edits or as nonsense humor.

Usage: Used in TikTok edits or friend chats as absurd humor. Sometimes implies "don't be uptight" or "stop talking nonsense," but often it's simply meaningless "brain-rot" humor that exaggerates absurdity.
Emotion: sarcastic/ironic, cringe/embarrassed
Sensitivity: sex
Question: *What is the most likely purpose of this clip?*
A. Seeking clarity.
B. **Nudge playfully.**
C. Warn seriously.
D. Request support.
Question Type: Usage & Application

Miku Hatsune's Ievan Polkka



Category: song
Year: 2006
Language: Finnish

Transcription: Ratsatsaa ja ripidabi dilla, Beritstan dillan dellan doo, A baribbattaa baribbariiba, Ribiribi distan dellan doo, Ja barillas dillan deia dooa, Daba daba daba daba daba duvja vuu.
Summary: The famous Leekspin originated from the Finnish song "Ievan Polkka," sung by the Finnish quartet Loituma. It first went viral after being paired with a looping animation of Orihime Inoue twirling a leek from the anime Bleach. Later, it became even more popular thanks to the Hatsune Miku version, along with the creation of the "Loituma Girl" (Negi Miku) character based on her.

Usage: Used for cover performances, dance routines, fan remixes (such as other anime characters spinning leeks), and as background music in cheerful or speed up videos.
Emotion: happy
Sensitivity: none
Question: *Who provided the voice for the first ACG character associated with this song?*
A. Saki Fujita.
B. **Yuki Matsuoka.**
C. Rina Satō.
D. Fumiko Orikasa.

Question Type: World Knowledge

I don't like sand



Category: noisy speech
Year: 2002
Language: English
Transcription: I don't like sand.
Summary: Anakin Skywalker, in an awkward attempt at flirting, complains about sand while comparing it to the person he is with.

Usage: Used to mock awkward dialogue, complain about minor inconveniences in a melodramatic way, or as a general reference to the Star Wars prequel trilogy.
Emotion: cringe/embarrassed
Sensitivity: none
Question: *What is the speaker's primary motivation for making this comment about sand?*
A. To provide a detailed geological analysis.
B. To share a fond childhood memory.
C. **As an awkward attempt to compliment his companion.**
D. To change the subject to his favorite topic, desert planets.

Question Type: Contextual Inference

Look at all those chickens



Category: clean speech
Year: 2013
Language: English

Transcription: Look at all those chickens
Summary: A young girl, awestruck, exclaims "Look at all those chickens¡' while looking at a flock of what are actually geese or ducks.

Usage: Used to express excitement or amazement at a large number of things, often while misidentifying them.
Emotion: happy, surprised/shocked
Sensitivity: none
Question: *What is the most likely reason for the person's exclamation?*
A. She is scared of the animals.
B. **She is excited and misidentifying the animals.**
C. She is angry at the animals for making a mess.
D. She is trying to sell the animals.

Question Type: Contextual Inference

霸王洗发水



Category: noisy speech
Year: 2004
Language: Chinese

Transcription: 当我第一次知道要拍那个洗发水广告的时候,其实我是拒绝的。因为我觉得,你不能叫我拍我就马上去拍,第一我要试一下,不然我又不想说,一个广告拍完后加上好多特技,那个头发,Duang,很黑,很亮,很柔,结果观众出来一定会骂我,根本没有这种头发,证明上面这个是假的。我说先给我试一下。后来我经过也知道它们是中药的,而且没有那种化学成分的。
Summary: An actor says he first refused to shoot the shampoo ad with the worry that special effects would make the hair look unreal, and only agreed after testing it.

Usage: "Duang" is often used for a flashy, sudden, over-the-top sound or visual effect.
Emotion: neutral
Sensitivity: none
Question: *Which one of the following is the least suitable to be described by the onomatopoeic word in this clip?*
A. Luxury race car.
B. Silky-shiny hair.
C. **Dramatic sea view.**
D. New game with good visuals.

Question Type: Usage & Application

超威蓝猫



Category: clean speech
Year: 2009
Language: Chinese

Transcription: 我说小朋友们谁愿意跟我蓝猫学一句配音呐? 只听见下面有一个：我来！这时候我一看！小朋友们，人群之中钻出来一个光头！哇！他全身穿的都是我们蓝猫品牌的衣服，特别显眼。我一把就把他抱了起来。

Summary: The voice actor calls for a voice-dubbing volunteer, and a bald kid in full Blue Cat clothes eagerly shouts out.

Usage: Used as a prosody when someone with a bald head suddenly appears.

Emotion: happy, surprised/shocked

Sensitivity: none

Question: *What is the funniest thing here?*
A. Cartoon suit.
B. Shouting out.
C. The voice acting.
D. **Bare hair.**

Question Type: Humor & Popularity

野々村議員



Category: noisy speech
Year: 2014
Language: Japanese

Transcription: この世の中を! ウグッブーン!! ゴノ ゴノ世の ブッヒィフエエエーーーンン!! ヒィェーーッフウンン!! ウゥ……ウゥ…… アﾞーーーーーアﾞッアﾞーー!!! ゴノぉよぉっきゃガッハッハアン!! アﾞーー 世の中を! うぅ変エダイ! その一心でえぇ!! ィヒー フーッ ハゥ 一生懸命訴えて 西宮$¥#%& 縁もゆかりもない 西宮ッヘエ市民の 皆さまに選出されて! やっと! 議員に! なったんですうぅー!!!

Summary: Mr.Nonomura is holding a press conference in which he is expressing his apology for misusing publidc funds.

Usage: Used as a reaction image to express fake or exaggerated sadness
Emotion: cringe/embarrassed
Sensitivity: none
Question: *What is the man expressing that is so ironic?*
A. **Desire to better the city.**
B. Desire to sincerely apologize.
C. Desire to keep his job.

Question Type: Language Analysis

They're taking the hobbits to Isengard



Category: song
Year: 2005
Language: English

Transcription: They're taking the Hobbits to Isengard.
Summary: A looping remix of Legolas from the Lord of the Rings repeatedly shouting "They're taking the hobbits to Isengard" edited to a catchy electronic beat.

Usage: Used to humorously announce that someone is being taken somewhere, or as a nostalgic reference to early viral internet video culture.
Emotion: sarcastic/ironic, happy
Sensitivity: none
Question: *How does the popular meme version of this clip transform the mood of the original dramatic scene?*
A. It makes the scene even more serious and tragic.
B. It removes the audio and adds a laugh track.
C. **It turns the scene into a catchy and repetitive electronic dance song.**
D. It slows down the dialogue for emphasis.

Question Type: Humor & Popularity

僕は新世界の神となる



Category: clean speech
Year: 2003
Language: Japanese

Transcription: いいか、僕はキラ。
そして、新世界の神！
Summary: A man claims he is God.

Usage: Used as a way of parodying Light's god complex by dropping his serious line into completely unserious, usually funny situations.
Emotion: nostalgic, sarcastic/ironic, cringe/embarrassed
Sensitivity: none
Question: *Why is he claiming that he is a god?*
A. Because he is a god.
B. **Because he has supernatural power.**
C. Because he is mentally ill and delusional.

Question Type: Contextual Inference