KwaiChat: A Large-Scale Video-Driven Multilingual Mixed-Type Dialogue Corpus

Anonymous ACL submission

Abstract

Video-based dialogue systems have compelling application value, such as education assistants, thereby garnering growing interest. However, the current video-based dialogue systems are limited by their reliance on a single dialogue type, which hinders their versatility in practical applications across a range of scenarios, including question-answering, emotional dialog, etc. In this paper, we identify this challenge as how to generate video-driven multilingual mixed-type dialogues. To mitigate this challenge, we propose a novel task and create a human-to-human video-driven multilingual mixed-type dialogue corpus, termed KwaiChat, containing a total of 93,209 videos and 246,080 dialogues, across 4 dialogue types, 30 domains, 4 languages, and 13 topics. Additionally, we establish baseline models on KwaiChat. An extensive analysis of 7 distinct LLMs on KwaiChat reveals that GPT-4o achieves the best performance but still cannot perform well in this situation even with the help of in-context learning and fine-tuning, which indicates that the task is not trivial and needs further research.[1]

1 Introduction

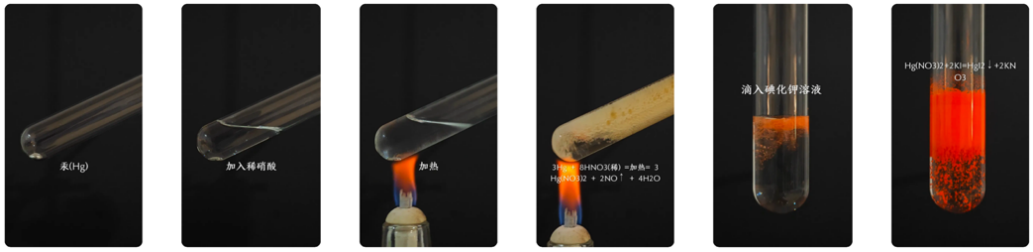
Recently, groundbreaking advancements in large language models (LLMs) (Zeng et al., 2022; OpenAI, 2023a; Touvron et al., 2023; Bao et al., 2023), have ignited significant technological shifts within the realm of dialogue systems. These LLMs are capable of producing responses that closely mimic human conversation, achieving remarkable performance, and fundamentally transforming the underpinnings of dialogue systems. Consequently, they are being integrated into various applications, such as healthcare (Laranjo et al., 2018; He et al., 2023; Hadi et al., 2023; Li et al., 2024b), education (Lan and Chen, 2024; Lieb and Goel, 2024), and law (Shu et al., 2024). Among these applications, multi-participant video-based dialogue systems stand out for their substantial contributions across various scenarios, such as collaborative learning (Anjewierden et al., 2007; Trausan-Matu et al., 2007) and entertainment (Herring et al., 2009; Cadilhac et al., 2012), which hold immense practical value and have garnered considerable interest from both academia and industry.

The current video-based dialogue systems (Lalanne et al., 2023) for multiple participants assume that dialogues are performed in a single dialogue type, predominantly focusing on chitchat. However, in many real-world applications, there are multiple dialogue types in human-bot conversations (termed as mixed-type dialogs), such as chitchat, emotional dialogue, and even question-answering (Ram et al., 2018; Zhou et al., 2020). As illustrated in Figure 1, the conversation encompasses a variety of dialogue types, including chitchat, question-answering, knowledge-grounded dialogue, and emotional dialogue. Therefore, how to generate video-driven mixed-type multi-participant dialogues is important and non-trivial. However, the challenge remains unexplored in previous works.

To facilitate the study of video-driven mixed-type multi-participant dialogue generation, we first construct a new human-to-human video-driven mixed-type multi-participant dialogue corpus by collecting videos, comments, replies of comments, domains, and topics from a video-sharing platform Kwai[2], owing to the rich video-driven mixed-type multi-participant dialogue resource in Kwai. For ethical considerations, we initially secure permission to access the data and meticulously remove content that could compromise privacy. Additionally, in our pursuit to advance multilingual research and enhance the robustness of dialogue systems, we collect a diverse corpus encompassing a variety of languages, including Portuguese, Chinese,

[1] Our dataset and code will be publicly available.

[2] https://www.kuaishou.com



Chinese	
Root Comment	热知识：硝酸钾不与任何物质发生反应 (Hot knowledge: Potassium nitrate does not react with any substance)
Topic	**Scientific Supplement**
Response 1	有硝酸根的大部分都不会 (Most substances with nitrate ions do not)
Response 2	冷知识：人家放的是稀硝酸 (Cold knowledge: They put dilute nitric acid)
Portuguese	
Root Comment	完了没学化学，我爱上化学了 (I didn't study chemistry, I fell in love with it)
Topic	**Simple Interaction**
Response 1	等你学起来你就会知道多么痛苦了 (When you start learning, you will know how painful it is)
Spanish	
Root Comment	突击提问：为什么加热前要把火焰摇几下？ (Assault question: Why do we need to shake the flame a few times before heating?)
Topic	**Seeking answers**
Response 1	预热，初中知识 (Preheating, junior high school knowledge)
Response 2	防止试管受热不均匀爆炸 (Prevent uneven heating and explosion of test tubes)
Indonesian	
Root Comment	一看到化学就头疼，脑海里都是痛苦的回忆。 (Seeing chemistry gives me a headache, and my mind is filled with painful memories.)
Topic	**Emotional Sharing**
Response 1	回首向来萧瑟处，也无风雨也无晴。 (Looking back at the desolate place, there was neither wind nor rain nor sunshine.)

Knowledge-grounded Dialogue | Chitchat | Question & Answering | Emotional Dialogue

Figure 1: An example of KwaiChat. The image above is captured from a video. Below the video, there are comments in four languages, and a Chinese dialogue is shown, with annotated topics and corresponding dialogue types.

Indonesian, and Spanish. Besides, to ensure the dataset quality and address the long-tail distribution of video topics, three filtering strategies and an adaptive video balancing method are utilized, respectively. Finally, the dataset is contained, termed as KwaiChat, with 93,209 videos and 246,080 dialogues, covering 4 dialogue types, 30 domains, 4 languages, and 13 topics. As the example shown in Figure 1, multiple participants engage in conversations around a given video in 4 different languages, and make their responses in different dialogue types based on the video, context, and topics.

To promote research on video-driven mixed-type multi-participant dialogue, we conduct benchmarking experiments on KwaiChat for end-to-end dialogue generation. Specifically, we conduct an extensive analysis of 7 LLMs' zero-shot (directly utilizing LLMs for inference and show no case in input), in-context learning (show 3, 4, and 5 case in input, respectively), fine-tuning performance on KwaiChat. Experimental results reveal that the existing state-of-the-art LLMs still cannot perform well in this situation even with the help of in-context learning and fine-tuning.

There are three main contributions of this work:

- We identify a new challenge as how to generate video-driven multilingual mixed-type multi-participant dialogues.
- To promote research on the challenge, we collect a new dataset, termed KwaiChat, containing a total of 93,209 videos and 246,080 dialogues, across 4 dialogue types, 30 domains, 4 languages, and 13 topics.
- We conduct extensive experiments of 7 LLMs on KwaiChat. Experiment results show that current LLMs can not perform well in generating video-driven mixed-type multi-participant dialogues even with in-context learning and fine-tuning, which indicates that the task is not trivial and needs further research.

2 Related Work

This work is closely related to image-based dialogue, video-based dialogue, and mixed-type dialogue. A summary is shown in Table 1.

Dataset	Mixed-type	Multilingual	Multimodal	Multi-parti	# of Turns
VisDial (Das et al., 2017)	✗	✗ (EN)	✓ (image,text)	✗	2.47M
IGC (Mostafazadeh et al., 2017)	✗	✗ (EN)	✓ (image,text)	✗	25.3K
Image-Chat (Shuster et al., 2018)	✗	✗ (EN)	✓ (image,text)	✗	401K
PhotoChat (Zang et al., 2021)	✗	✗ (EN)	✓ (image,text)	✗	156K
MMDD (Lee et al., 2021)	✗	✗ (EN)	✓ (image,text)	✗	346K
DialogCC (Lee et al., 2022)	✗	✗ (EN)	✓ (image,text)	✗	929K
MMChat (Zheng et al., 2021)	✗	✗ (ZH)	✓ (image,text)	✗	314K
MMDialog (Feng et al., 2022)	✗	✗ (EN)	✓ (image,text)	✗	4.92M
OpenViDial (Meng et al., 2020)	✗	✗ (EN)	✓ (image,text)	✗	41.10M
OpenViDial 2.0 (Wang et al., 2021)	✗	✗ (EN)	✓ (image,text)	✗	5.60M
AVSD (Alamri et al., 2019)	✗	✗ (EN)	✓ (video,text)	✗	236K
YTD-18M (Han et al., 2023)	✗	✗ (EN)	✓ (video,text)	✗	54M
MLED (Poria et al., 2018)	✗	✗ (EN)	✓ (video,text)	✗	13.7K
M^3ED (Poria et al., 2018)	✗	✗ (ZH)	✓ (video,text)	✗	24.4K
TikTalk (Lin et al., 2023)	✗	✗ (ZH)	✓ (video,text)	✗	827K
CRD3 (Rameshkumar and Bailey, 2020)	✗	✗ (EN)	✗ (text)	✓	399K
MediaSum (Zhu et al., 2021)	✗	✗ (EN)	✗ (text)	✓	463K
Reddit (Schrading et al., 2015)	✗	✗ (EN)	✗ (text)	✓	628K
DuRecDial 2.0 (Liu et al., 2021)	✓	✓ (ZH, EN)	✗ (text)	✗	255K
COD (Majewska et al., 2023)	✗	✓ (RU, AR, ID, KIS)	✗ (text)	✗	800
CCNet (Hung et al., 2022)	✗	✓ (AR, ZH, RU, DE)	✗ (text)	✗	100K
BiToD (Lin et al., 2021)	✗	✓ (ZH, EN)	✗ (text)	✗	144K
MedDialog (Zeng et al., 2020)	✗	✓ (ZH, EN)	✗ (text)	✗	4M
LiveChat (Lalanne et al., 2023)	✗	✗ (JPN)	✓ (video,text)	✓	3.2M
KwaiChat (Our)	✓	✓ (ZH, ES, ID, POR)	✓ (video, text)	✓	0.8M

Table 1: Comparison of KwaiChat with other dialogue datasets. "Multi-parti" stands for multi-participant. "DE", "EN", "ZH", "JPN", "ID", "RUS", "AR", "KIS", "ES", and "POR" stand for German, English, Chinese, Japanese, Indonesian, Russian, Arabic, Kiswahili, Spanish and Portuguese.

2.1 Image-based Dialogue

Image-based dialogue systems are required to answer questions accurately, given an image, a dialog context, and a question about the image. Typical image-based dialogue datasets include VisDial (Das et al., 2017), IGC (Mostafazadeh et al., 2017), Image-Chat (Shuster et al., 2018), PhotoChat (Zang et al., 2021), MMDD (Lee et al., 2021), DialogCC (Lee et al., 2022), MMChat (Zheng et al., 2021), MMDialog (Feng et al., 2022), OpenViDial (Meng et al., 2020), and OpenViDial 2.0 (Wang et al., 2021). The dialogues in these datasets are monolingual, image-based, and have two participants. Compared with these studies, KwaiChat is a unique video-driven multi-participant dialogue corpus that supports multiple languages and incorporates various dialogue types.

2.2 Video-based Dialogue

Typical video-based dialogue datasets include AVSD (Alamri et al., 2019), YTD-18M (Han et al., 2023), MLED (Poria et al., 2018), M^3ED (Poria et al., 2018), TikTalk (Lin et al., 2023), LiveChat (Japan) (Lalanne et al., 2023), LiveChat (Chinese) (Gao et al., 2023), and VideoChat (Li et al., 2023). Dialogues in these datasets are monolingual, video-based, and have two participants. Compared with the existing literature, KwaiChat stands out as a comprehensive, multi-participant dialogue dataset that boasts multilingual support and diverse dialogue types.

2.3 Mixed-type Dialogue

Recently, research on the mixed-type dialogue has increased significantly. These researches fall into two categories: (1) train an all-in-one conversation model by using multiple single-skill conversation datasets, such as persona-chat, task-oriented dialogue, to bind multiple dialogue skills (Madotto et al., 2020; Roller et al., 2021; Madotto et al., 2021); (2) collect mixed-type dialog datasets (Shuster et al., 2020; Smith et al., 2020; Sun et al., 2021; Liu et al., 2021; Chiu et al., 2022; Liu et al., 2022) to train mixed-type dialog models. Those datasets are intended to mix different dialogue skills to meet specific needs, such as recommending movies and songs, and are unable to solve medical consultations. Compared with them, we collect a video-driven multilingual mixed-type dialogue corpus, KwaiChat, to facilitate further study.

3 Dataset Collection

Kwai[3] is a global video website with a large number of users in various languages and various marked topics, which is suitable for collecting dialogue corpus. Thus, Kwai is utilized as the data source. This section describes the two main steps for KwaiChat construction: (1) dialogue selection; (2) data merging and balancing. The dialogue selection is designed to select data of high quality. Then, to address the long-tail distribution of video topics, data merging, and balancing is applied with an adaptive video balancing method.

3.1 Dialogue Selection

To ensure better data quality, two data filtering strategies are applied, including preliminary filtering based on SQL tags, and heuristics rules. Besides, to remove data related to violence, illegal activities, and ethical issues, LLMs-based method is utilized. Below is a detailed introduction.

SQL Tags based Filtering. The filtering method based on SQL tags is designed to conduct a preliminary screening of data based on SQL tags. Main comments and reply comments are handled with different strategies. For main comments, both likes and direct comments must reach a certain threshold, which are set as 500, and 200, respectively. Besides, those deleted comments, or comments with "@", or AI comments are all not kept. For those reply comments, likes must reach a certain threshold of 50, with those deleted, or with "@", or AI comments not kept. Besides, multiparticipant dialogues are extracted with the requirement that *reply_to_comment_id=root_comment_id*. Finally, a total of 691,282 Chinese dialogues and 623,934 dialogues in other languages are obtained.

Heuristic Rule-based Filtering. Five heuristic rules are utilized for the data filtering. First, main comments with empty replies are deleted. If the main comment has no reply comments, delete the main comment. The rule is designed to obtain multiparticipant, multi-turn dialogues. Second, short comments are deleted. If the reply comment has fewer than 3 characters, delete the reply comment. The rule is designed based on the observation that short replies are mostly tone words with little information content, which is low in quality, such as "Haha", and "like". To ensure the dialogue quality, these dialogues are deleted. Third, comments containing specific keywords are deleted. Reply comments deleted, which contain the specified keywords ["Feeling will be popular", "Will be popular", "Occupying the floor", "Occupying a floor", "Front row", "First floor", "Second floor"]. Then, data deduplication is applied. Reply comments with the same content are only kept once. Finally, a total of 440,666 Chinese dialogues and 303,062 dialogues in other languages are kept.

LLM based Filtering. LLM prompting technique is utilized to filter the dataset. Specifically, dialogues with the following six aspects are removed by ChatGPT (OpenAI, 2022). 1. Illegal Content **Violence**: Content describing or inciting violent behavior. **Terrorism**: Content related to terrorist activities. **Illegal Activities**: Including but not limited to drug trafficking, hacking attacks, copyright infringement, etc. 2. Harmful Content **Hate Speech**: Attacks or defamation against groups of race, religion, gender, sexual orientation, nationality, etc. **Discrimination**: Any form of discriminatory content. **Harassment**: Harassment, threats, or bullying against individuals or groups. 3. Inappropriate Content **Adult Content**: Including pornography, obscenity, or explicit sexual content. **Extreme Violence**: Content that is too bloody or cruel. **Sensitive Topics**: Topics such as suicide, self-harm, or eating disorders. 4. False Information **Fake News**: Spreading misleading or false news information. **Misleading Information**: Spreading unverified rumors or misleading information. 5. Morality and Ethics **Moral Standards**: Content that violates the moral standards recognized by society. **Ethical Issues**: Content involving ethical controversies, such as human experimentation or animal abuse. 6. Privacy **Personal Information**: Disclosing personal privacy information, such as addresses, phone numbers, ID numbers, etc. **Sensitive Data**: Involving sensitive data such as financial information, medical records, etc.

Finally, a total of 391,036 dialogues are kept.

3.2 Data Merging and Balancing

To address the long-tail distribution of video topics, an adaptive video topic balancing method based on the number of likes is designed to balance the topics of videos. It is observed that an abundance of videos on topics tends to attract a larger audience, which in turn generates a higher volume of views. Consequently, this leads to an increase in the number of likes received. Given this correlation, it becomes evident that a higher threshold for content curation is necessary to ensure qual-

[3] https://www.kuaishou.com/

Figure 2: Domains of KwaiChat.

Split	Language	# of Videos	Avg-v	Avg-u
Train	POR	41,400	83.05	3.04
	ZH	21,021	58.22	11.28
	ID	11,419	57.33	3.00
	ES	3,824	35.12	3.14
Validation	POR	4,116	80.40	3.14
	ZH	2,117	58.43	11.28
	ID	1,073	59.25	2.97
	ES	371	31.90	3.00
Test	POR	4,209	84.24	3.66
	ZH	2,119	59.49	10.94
	ID	1,147	58.80	3.08
	ES	393	34.09	3.32

Table 2: Data division of KwaiChat. "Avg-v", and "Avg-u" denote the average duration of videos, the average utterance of dialogues.

ity and relevance. Based on this hypothesis, an adaptive algorithm is designed to calculate the like count threshold *Like_threshold* based on the number of videos in a topic. The parameters of the algorithm include (1) *base_threshold* (θ_b): The base threshold value, which defaults to 500. (2) *scaling_factor* (θ_s): The scaling factor, which defaults to 0.5. (3) *min_threshold and max_threshold*: The minimum and maximum limits for the threshold value, which are 50 and 10,000, respectively. (4) *exponent* (β): The power exponent, is used to control the nonlinear scaling, which defaults to 0.5. (5) *scaling_factor* (γ): The scaling factor. (6) *quantile_threshold* (θ_q): The quantile of video like number is set as 75%.

The scaled threshold is calculated as

$$\theta_s = \theta_b \times \left(\frac{N}{\theta_q}\right)^{\beta} \times \gamma,$$

where N is the number of videos. The *Like_threshold* (θ_l) is calculated as

$$\theta_l = \begin{cases} \text{min_threshold}, & \text{if } \theta_s < \text{min_threshold} \\ \text{max_threshold}, & \text{if } \theta_b > \text{max_threshold} \\ \theta_s, & \text{otherwise} \end{cases}.$$

Videos with likes below the specified threshold are regarded as low-frequency. To mitigate the long tail issue, only those videos that surpass this threshold are preserved.

3.3 Data Statistics and Quality

Data Statistics. Table 2 provides statistics of KwaiChat. Statistics show that dialogues in KwaiChat are rich in dialogue numbers and long in video duration, which is much larger than other multi-modal dialogue datasets.

Figure 2 illustrates the domains of KwaiChat. The dataset is expansive, encompassing a diverse array of 30 domains, underscoring its richness and providing a robust foundation for the advancement of mixed-type dialogue systems. There are six main domains, including lifestyle and interests, health and lifestyle, education and culture, entertainment and arts, technology and innovation, and news and current affairs. There are a total of 30 domains in the dataset, with 5 domains for health and lifestyle, 6 domains for education and culture, 9 domains for entertainment and arts, 2 domains for technology and innovation, 3 domains for news and current affairs, 5 domains for lifestyle and interests.

Figure 3 illustrates the domains, languages, topics, and dialogue types of KwaiChat. The table elegantly delineates the intricate connections between diverse elements, which shows that our filtering algorithm ensures a balanced and uniform distribution across all categories. Besides, the dataset contains 30 domains, underscoring its richness and providing a robust foundation for the advancement of mixed-type dialogue systems.

Data Quality. Following the evaluation method in previous work (Liu et al., 2020), we conduct human evaluations for data quality.[4] Finally, we obtain an average score of 0.90 on this evaluation set, indicating that the dataset has high quality.

[4]A dialog is rated with a score "1" if it is coherent, logical, grammatically correct, otherwise "0". Then, we ask three experts to judge the quality of 200 randomly sampled dialogs

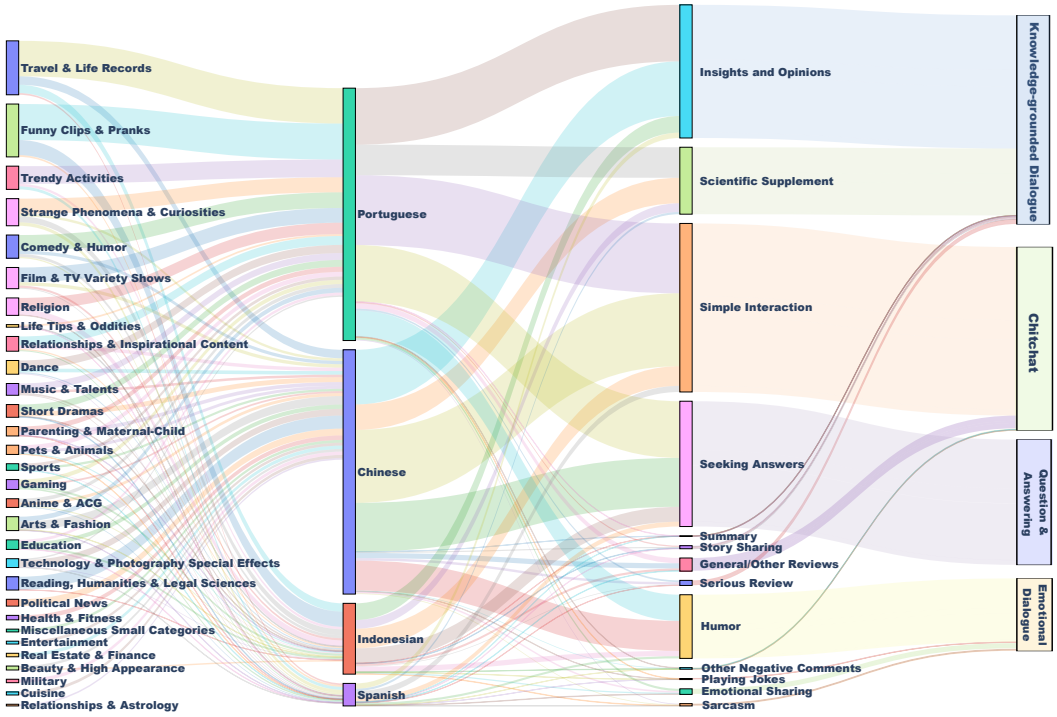


Figure 3: The domains, languages, topics, and dialogue types of KwaiChat. The first column lists the domains. The second column lists the four languages. The third column lists the topics. The fourth column lists the dialogue types.

4 Experiments and Results

This section introduces the experimental setting, data and evaluation metrics, baselines, automatic evaluations, human evaluations, and the case study.

4.1 Experimental Setting

Experimental Setting. For Transformer, the implementation by HuggingFace [5] is utilized, where the hyperparameters follow the default settings in the original Transformer (Vaswani et al., 2017).

Computing Platform. Our experiments are conducted on the workstation with an Intel Xeon E5 2.40 GHz CPU, 128 GB memory, an NVIDIA A100 GPU, and CentOS 7.2.

4.2 Data and Evaluation Metrics

4.2.1 Data

Following Tiktalk (Lin et al., 2023), we split KwaiChat into the training set, the validation set, and the test set by randomly sampling data with the ratio of 10:1:1, and ensure that random and identical videos do not appear in different sets.

[5] https://github.com/huggingface/transformers

4.2.2 Automatic Evaluation Metrics

Following Tiktalk (Lin et al., 2023), we adopt two types of automatic metrics to evaluate the performance of models from different perspectives, token similarity and diversity.

Similarity. Three basic automatic evaluation metrics for generation tasks are utilized in this work, including BLUE1 (Papineni et al., 2002), BLUE2 (Papineni et al., 2002), and ROUGE-L (Lin, 2004). These metrics all measure the similarity between the generated responses and the ground truth via n-gram matching.

Diversity. Following previous work(Liu et al., 2021), this paper adopts DIST-2 and DIST-3 (Li et al., 2015) to measure the diversity of generated responses.

4.2.3 Human Evaluation Metrics

Following Lin et al. (2023), five human evaluation metrics are utilized in this work:

Sensibleness measures whether the response is meaningful and coherent, and conforms to logic and common sense;

Specificity measures whether the response is relevant to a specific context, provides sufficient details, and avoids being too vague or general;

	Model	Language	BLEU1	BLEU2	ROUGE-L	DIST-2	DIST-3
Text Model	Qwen2-72B-Instruct	POR	2.25	0.21	**5.00**	99.89	99.99
		ID	1.16	0.14	2.95	**99.95**	**100.00**
		ES	2.67	0.25	4.68	99.94	99.99
		ZH	**5.85**	**0.57**	0.23	99.62	99.90
		POR,ID,ES,ZH	4.21	0.41	2.11	99.74	99.94
	LLaMA3.1-8B	POR	1.98	0.20	**5.20**	99.10	99.30
		ID	0.94	0.14	1.77	95.86	94.84
		ES	2.67	0.30	4.12	**99.61**	99.14
		ZH	**5.16**	**0.61**	0.16	99.49	**99.57**
		POR,ID,ES,ZH	3.72	0.43	2.02	99.05	99.06
Multi-modal	Qwen-VL	POR	2.42	0.27	**4.84**	90.84	92.40
		ID	0.93	0.14	2.23	79.31	81.78
		ES	2.00	0.28	3.65	76.29	78.59
		ZH	**6.35**	**0.93**	0.37	**96.30**	**97.51**
		POR,ID,ES,ZH	4.56	0.64	1.99	92.58	94.04
	Video-LLaMA2-7B	POR	2.56	0.35	**5.39**	77.97	80.21
		ID	1.49	0.22	3.32	79.24	82.38
		ES	2.29	0.27	4.49	80.15	82.32
		ZH	**6.90**	**1.11**	0.34	**94.96**	**96.20**
		POR,ID,ES,ZH	4.92	0.77	2.31	87.82	89.56
	InternVL2-26B	POR	1.86	0.23	**4.04**	80.54	83.86
		ID	1.40	0.20	3.00	91.48	93.16
		ES	2.00	0.26	3.71	78.18	81.11
		ZH	**6.30**	**1.04**	0.45	**94.82**	**96.33**
		POR,ID,ES,ZH	4.41	0.70	1.85	89.72	91.83
	LLaVA-Next-Video-7B	POR	0.48	0.05	1.18	36.80	38.16
		ID	0.49	0.07	1.22	**66.13**	**67.23**
		ES	0.77	0.08	**1.63**	51.17	54.15
		ZH	**1.98**	**0.26**	0.30	62.60	63.30
		POR,ID,ES,ZH	1.46	0.18	0.64	56.36	57.33
	GPT-4o	POR	2.68	0.27	**5.71**	99.83	**99.98**
		ID	1.23	0.16	3.05	**99.87**	99.97
		ES	3.01	0.34	5.65	99.75	99.96
		ZH	**6.34**	**0.81**	0.26	99.52	99.79
		POR,ID,ES,ZH	4.70	0.58	2.31	99.65	99.87

Table 3: Results of 7 models on KwaiChat in the zero-shot setting. Results are reported in percentage (%). "POR", "ID", "ES", and "ZH" stand for Portuguese, Indonesian, Spanish, and Chinese.

Vision Relevance examines the degree of correlation between the response and the multi-modal information in the video;

Dialog Relevance measures fluency, relevancy, and logical consistency of each response when given the global context;

Human-likeness examines the similarity between generated response and with corresponding human response from the perspectives of appropriateness, fluency, and proactivity.

Each metric has 7 levels, ranging from point 1 (worst) to point 7 (best). We randomly select the generation results of 100 dialogues from the test set and ask three annotators who are experts in this task to independently evaluate them. The final score is calculated as the mean of their scores.

4.3 Baselines

We carefully select a few strong baselines for comparison. Specifically, two baselines for text-only dialogue generation (Qwen2-Instruct (Yang et al., 2024), LLaMA3.1 (Dubey et al., 2024)) and five baselines for multi-modal dialogue generation (Qwen-VL (Bai et al., 2023), Video-LLaMA2 (Zhang et al., 2023), InternVL2 (Chen et al., 2024), LLaVA-Next-Video (Li et al., 2024a), GPT-4o (OpenAI, 2023b)).

4.4 Automatic Evaluation

We conduct experiments in three settings, zero-shot, in-context, in-context learning, and fine-tuning.

Zero-shot. The results of zero-shot on automatic evaluation metrics are shown in Table 3. The results show the following conclusions. First, video-based LLMs perform better than text LLMs, which indi-

Model	Sensibleness	Specificity	Vision-R	Dialog-R	Human-likeness	Average
Human comment	**6.00**	5.25	4.71	**5.96**	**6.54**	**5.70**
Qwen2-7B-Instruct	2.34	3.38	1.96	3.46	3.09	2.84
Qwen2-7B-Instruct-LoRA	3.71	3.04	2.30	4.00	5.29	3.67
Qwen2-7B-Instruct-Full	2.59	1.79	1.71	2.54	4.46	2.62
Qwen2-72B-Instruct	3.88	4.54	2.96	4.54	4.00	3.98
Qwen2-72B-Instruct-3-shot	3.05	3.54	2.63	3.67	3.46	3.27
Qwen2-72B-Instruct-4-shot	3.25	3.96	2.38	4.50	3.63	3.54
Qwen2-72B-Instruct-5-shot	3.25	3.84	3.00	4.09	3.46	3.53
InternVL2-8B	3.63	4.25	4.46	4.71	3.58	4.12
InternVL2-8B-LoRA	2.84	2.17	2.33	3.00	4.58	2.98
InternVL2-8B-Full	4.50	4.04	3.63	4.63	5.75	4.51
GPT-4o	4.29	4.25	4.38	5.00	4.71	4.53
GPT-4o-3shot	4.25	4.42	3.79	5.21	5.29	4.59
GPT-4o-4shot	5.13	5.04	4.13	5.84	5.54	5.13
GPT-4o-5shot	5.58	**5.38**	**4.79**	5.71	6.25	5.54

Table 4: Human evaluation results on five aspects. The average scores are reported. "Vision-R" and "Dialog-R" stand for vision relevance and dialog relevance, respectively.

cates that LLMs cannot make accurate responses without video information. Second, LLMs perform better in Chinese than in other languages. The reason is that during the pre-training phase, there is an abundance of high-quality Chinese training corpora, which allows LLMs to achieve better pre-training in Chinese. Third, the performance of LLMs is high on DIST, and the reason is that LLMs can generate diverse and fluent responses.

In-context Learning. The results of LLMs' in-context learning performance are shown in Table 5. The results show the following conclusions. First, showing more cases performs better. The reason is that more cases help LLMs to better understand the questions and answers, thus LLMs can make better responses. Second, GPT-4o performs better than Qwen2 even with the few-shot setting, which shows strong power of GPT-4o.

Fine-tuning. The results of LLMs' fine-tuning performance are shown in Table 6. The results show that fine-tuning with Lora performs better than fine-tuning full parameters. The reason may be that fine-tuning of all parameters may result in overfitting to the training data, leading to low generalization.

4.5 Human Evaluation

Table 4 shows the human evaluation results on the five evaluation metrics. Three conclusions can be obtained from the results. First, these models perform better on human-likeness and dialog relevance than the other three evaluation metrics. This shows that LLMs are good at generating fluency and context-aware responses. Second, multi-modal models outperform text-only models due to their enhanced ability to leverage video information, which results in superior dialogue outcomes. Third, GPT-4o achieves the best performance, which shows that GPT-4o is currently the best LLM on this task.

4.6 Case Study

To further investigate the performance of the baselines, we present a response case in Table 7. Results show that LLaMA3.1 refuses to reply to the second comment. The possible reason is that LLaMA3.1 regards the comment as an illegal event. Besides, Video-LLaMA2 gives a wrong answer, which shows that it contains serious hallucinations.

5 Conclusion

This work first identified the issue of how to generate multi-participant video-driven mixed-type dialogue. To address this issue, we introduced a novel task and built a corpus of human-to-human multi-participant video-driven multilingual mix-typed corpus, named KwaiChat. To facilitate further research, we established baselines on KwaiChat. Next, we will promote the cross-lingual research.

Limitations

This research is limited to an analysis of four major languages, focusing on LLMs' performance towards video-driven mixed-type multi-party conversations. Moving forward, we are eager to delve into the realm of languages in low resources. By studying LLMs' performance on minority languages, we aim to enhance LLMs' ability to handle a broader range of linguistic inputs, ultimately leading to more inclusive and sophisticated conversational AI systems.

Ethics Statement

Regarding safety, we acknowledge the limitations of the current framework proposed and the potential risks associated with deploying them directly for users. We do not recommend the direct deployment of the fine-tuned models from this work into real-life situations; currently, they are only suitable for academic research. While we intend to develop models for the greater good of society, it is crucial to recognize that the dataset contains potentially problematic content, including toxic or biased material that could be used to generate negative or offensive content.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Anjo Anjewierden, Bas Kolloffel, and Casper Hulshof. 2007. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *International Workshop on Applying Data Mining in e-Learning (ADML 2007)*.

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*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Anaïs Cadilhac, Nicholas Asher, and Farah Benamara. 2012. Annotating preferences in chats for strategic games. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 139–143.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158, Dublin, Ireland. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

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*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *arXiv preprint arXiv:2306.08401*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. *arXiv preprint arXiv:2303.09713*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Susan C Herring, Daniel O Kutz, John C Paolillo, and Asta Zelenkauskaite. 2009. Fast talking, fast shooting: Text chat in an online first-person game. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022.

Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. *arXiv preprint arXiv:2205.10400*.

Julien Lalanne, Raphael Bournet, and Yi Yu. 2023. Livechat: Video comment generation from audio-visual multimodal contexts. *arXiv preprint arXiv:2311.12826*.

Yu-Ju Lan and Nian-Shing Chen. 2024. Teachers' agency in the era of llm and generative ai. *Educational Technology & Society*, 27(1):I–XVIII.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, and Ho-Jin Choi. 2022. Dialogcc: Large-scale multi-modal dialogue dataset. *arXiv preprint arXiv:2212.04119*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024b. Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine: 108013*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Anna Lieb and Toshali Goel. 2024. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, et al. 2023. Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1303–1313.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*.

Qianchu Liu, Diana McCarthy, and Anna Korhonen. 2020. Towards better context-aware lexical semantics: Adjusting contextualized representations through static anchors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4066–4075.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1024–1034, Dublin, Ireland. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *AAAI*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2020. Attention over parameters for dialogue systems. *Preprint*, arXiv:2001.01871.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *Available: https://openai.com/blog/chatgpt/*.

OpenAI. 2023a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023b. Gpt-4: technical report. *https://arxiv.org/pdf/2303.08774.pdf*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and et al. 2021. Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2577–2583.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. *arXiv preprint arXiv:2407.21065*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Kurt Shuster, Da JU, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *NAACL*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Stefan Trausan-Matu, Traian Rebedea, Alexandru Dragan, and Catalin Alexandru. 2007. Visualisation of learners' contributions in chat conversations. *Blended learning*, pages 217–226.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

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

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

A In-context Learning Results

The results of LLMs' in-context learning performance are shown in Table 5. The results show the following conclusions. First, showing more cases performs better. The reason is that more cases help LLMs to better understand the questions and answers. Second, GPT-4o performs better than Qwen2 with the few-shot setting, which shows the strong performance of GPT-4o. Third, current models still fall short in terms of BLEU and ROUGE metrics, indicating that further improvement is necessary. This also underscores the significance of this dataset.

B Fine-tuning Results

The results of LLMs' fine-tuning performance are shown in Table 6, where the performance of LLMs fine-tuning with the Lora technique and fine-tuning full parameters are reported. The results show that fine-tuning with Lora performs better than fine-tuning full parameters. The reason may be that fine-tuning of all parameters may result in overfitting to the training data, leading to low generalization. Besides,

C Case Study

To further investigate the performance of the baselines, we present a response case in Table 7. Results show that LLaMA3.1 provides inconsistent answers to the main comment and refuses to reply to the second comment. The possible reason is that LLaMA3.1 regards the second main comment as an illegal event. QwenVL provides inconsistent answer to the first main comment but provides correct answer to the second one. Video-LLaMA2 gives a wrong answer for both the two main comments, which shows that it contains serious hallucinations. InternVL2 also contains hallucination in the first comment. Besides, only GPT-4o makes correct replies for these two comments. The results show that current LLMs fall short in knowledge-based question-and-answering.

Model	Setting	Language	BLEU1	BLEU2	ROUGE-L	DIST-2	DIST-3
Qwen2-72B-Instruct-GPTQ-Int4	Base	POR	2.25	0.21	**5.00**	99.89	99.99
		ID	1.16	0.14	2.95	**99.95**	**100.0**
		ES	2.67	0.25	4.68	99.94	99.99
		ZH	**5.85**	**0.57**	0.23	99.62	99.90
		POR,ID,ES,ZH	4.21	0.41	2.11	99.74	99.94
	3-shot	POR	2.41	0.26	**5.36**	**99.93**	99.97
		ID	1.31	0.19	3.35	99.91	**99.99**
		ES	2.72	0.27	5.10	99.89	**99.99**
		ZH	**6.86**	**0.78**	0.18	99.11	99.60
		POR,ID,ES,ZH	4.82	0.55	2.20	99.46	99.76
	4-shot	POR	2.31	0.24	**5.29**	**99.94**	99.99
		ID	1.35	0.17	3.40	99.86	99.96
		ES	2.44	0.23	4.65	99.93	**100.0**
		ZH	**6.97**	**0.83**	0.29	99.11	99.59
		POR,ID,ES,ZH	4.87	0.57	2.27	99.46	99.76
	5-shot	POR	2.25	0.25	**5.15**	**99.91**	**99.98**
		ID	1.35	0.17	3.43	99.89	**99.98**
		ES	2.73	0.29	4.88	99.84	99.94
		ZH	**7.06**	**0.83**	0.17	99.13	99.61
		POR,ID,ES,ZH	4.91	0.57	2.17	99.46	99.77
GPT-4o	Base	POR	2.68	0.27	**5.71**	99.83	**99.98**
		ID	1.23	0.16	3.05	**99.87**	99.97
		ES	3.01	0.34	5.65	99.75	99.96
		ZH	**6.34**	**0.81**	0.26	99.52	99.79
		POR, ID, ES, ZH	4.70	0.58	2.31	99.65	99.87
	3-shot	POR	2.72	0.38	**6.55**	**99.97**	**99.89**
		ID	1.73	0.29	4.33	99.93	99.76
		ES	2.74	0.41	5.95	99.92	99.81
		ZH	**8.49**	**1.47**	0.22	98.30	98.96
		POR, ID, ES, ZH	5.95	1.00	2.68	98.99	99.29
	4-shot	POR	2.50	0.35	**6.63**	**99.97**	99.75
		ID	1.92	0.28	4.33	99.95	99.74
		ES	3.16	0.57	**6.63**	**99.97**	**99.83**
		ZH	**8.58**	**1.55**	0.21	98.07	98.82
		POR, ID, ES, ZH	6.01	1.05	2.69	98.86	99.21
	5-shot	POR	2.61	0.41	**6.76**	**99.96**	99.80
		ID	1.91	0.28	4.53	99.93	99.70
		ES	2.89	0.47	6.17	99.95	**99.82**
		ZH	**8.83**	**1.61**	0.29	97.96	98.70
		POR, ID, ES, ZH	6.18	1.10	2.78	98.79	99.15

Table 5: Results of Qwen2-72B-Instruct-GPTQ-Int4 and GPT-4o in the 3-shot, 4-shot, and 5-shot setting.

	Model	Language	BLEU1	BLEU2	ROUGE-L	DIST-2	DIST-3
Text Model	Qwen2-7B-Instruct-Lora	POR	4.58	1.10	**10.55**	86.17	71.37
		ID	3.85	0.89	7.70	88.07	72.82
		ES	5.58	1.33	10.21	**94.01**	87.49
		ZH	**9.93**	**2.76**	0.27	92.44	**93.27**
		POR,ID,ES,ZH	7.58	2.03	4.47	90.13	84.40
	Qwen2-7B-Instruct-Full	POR	2.57	0.46	**6.91**	90.75	69.99
		ID	1.77	0.33	3.96	82.32	63.64
		ES	3.65	0.52	5.82	**94.10**	89.54
		ZH	**9.39**	**2.26**	0.05	91.93	**92.99**
		POR,ID,ES,ZH	6.40	1.47	2.73	90.79	83.07
Multi-modal	InternVL2-8B-LoRA	POR	3.23	0.51	**7.49**	91.81	81.02
		ID	2.22	0.43	4.70	86.34	72.45
		ES	3.56	0.58	6.22	**93.93**	86.17
		ZH	**8.76**	**2.02**	0.27	93.36	**92.23**
		POR,ID,ES,ZH	6.38	1.39	3.01	92.31	86.98
	InternVL2-8B-Full	POR	3.17	0.60	**7.75**	89.28	75.71
		ID	2.02	0.32	4.65	83.63	66.26
		ES	3.36	0.76	6.62	91.30	78.94
		ZH	**8.29**	**1.99**	0.31	**93.56**	**92.00**
		POR,ID,ES,ZH	6.07	1.39	3.12	91.36	84.50

Table 6: Results of Qwen2-7B and InternVL2-8B with Lora fine-tuning and full-parameter fine-tuning.

Main Comment	热知识：硝酸钾不与任何物质反应 (Hot knowledge: Potassium nitrate does not react with any substance)
Ground Truth	有硝酸根的大多不会 (The majority of nitrate ions are not present)
LLaMA3.1	很好奇哪些物质能和硝酸钾反应呢? (I am curious about which substances can react with potassium nitrate?)
QwenVL	硝酸钾可不，你试试把土豆丢进硝酸钾里，保证一个土豆变两个(Is potassium nitrate okay? You can try throwing the rice mud into potassium nitrate to make sure you can turn one into two)
Video-LLaMA2	硝酸钾确实不会与任何物质发生反应，这是一种非常稳定的化学物质。 (Potassium nitrate does not react with any substance, it is an extremely stable chemical substance.)
InternVL2	哈哈，看来你对化学知识了解的挺多啊！不过，一种非常稳定的化学物质，它通常不会与大多数物质发生反应。但是，在高温下，它会与氢气反应生成氮氧化物和水。 (Haha, it seems that you have a good understanding of chemistry! However, potassium nitrate is indeed a highly stable compound that typically does not react with most substances. However, at high temperatures, it does react with hydrogen gas to produce nitrogen oxides and water.)
GPT-4o	哈哈，这个热知识简直太搞笑了，事实摆在眼前，硝酸钾可是剧烈反应的好手啊！(Haha, this hot knowledge is simply hilarious! The fact is that potassium nitrate is a good hand for violent reactions!)
Main Comment	为什么加热前要把火焰摇几下？(Why do we need to lower the flame before heating?)
Ground Truth	防止试管受热不均匀爆炸(Prevent uneven heating and explosion of test tubes)
LLaMA3.1	不好意思，无法为这个问题提供回复。(Sorry, we are unable to provide a response to this question.)
QwenVL	这一步骤是为了让硝酸甘油和硝酸钾混合物均匀受热，避免局部过热导致硝酸甘油分解产生气体。 (This step is to evenly heat the mixture of nitric acid white oil and potassium nitrate, avoiding local overheating that may cause the decomposition of nitric acid white oil to produce gas.)
Video-LLaMA2	加热前把火焰摇几下可能是为了增加火焰的热量，从而加快加热的速度。 (Shaking the flame before heating can increase the heat of the flame and accelerate the heating process.)
InternVL2	因为这样可以让火焰更均匀，避免局部过热。 (Because this makes the flame uniform and avoid local overheating.)
GPT-4o	嘿，看来你也是个化学爱好者！加热前摇几下火焰是为了让火焰更均匀，这样加热的时候不会因为突然的温差导致试管破裂哦。科学实验真是处处有玄机呢！(Hey, it seems like you're also a chemistry enthusiast! Shaking the flame before heating is to make the flame more stable and uniform, so that the test tube will not rupture due to sudden temperature differences during heating. Scientific experiments are full of opportunities!)

Table 7: Two generation cases of 5 LLMs for the given video and context.