ConvFaithEval: Evaluating Faithfulness of Large Language Models with Real-World Customer Service Conversations

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in diverse tasks but are prone to hallucinations. Most existing benchmarks primarily focus on evaluating factual hallucinations, while the assessment of faithfulness hallucinations remains underexplored, especially with practical conversations that involve casual language and topic shifts. To bridge this gap, we introduce CONVFAITHEVAL, the first faithfulness hallucination evaluation benchmark built on real-world customer service conversations. CONVFAITHEVAL features $3,369$ anonymized conversations with generated summaries as well as $19,782$ corresponding quiz questions. Two tasks, *Conversation Summarization* and *Quiz Examination*, are designed to comprehensively assess faithfulness hallucinations in LLMs. Extensive experiments on 23 LLMs reveal that faithfulness hallucinations persist across all LLMs, with closed-source models consistently less affected than their open-source counterparts. To mitigate hallucinations, we further explore four strategies and offer valuable insights for future development of advanced methods.

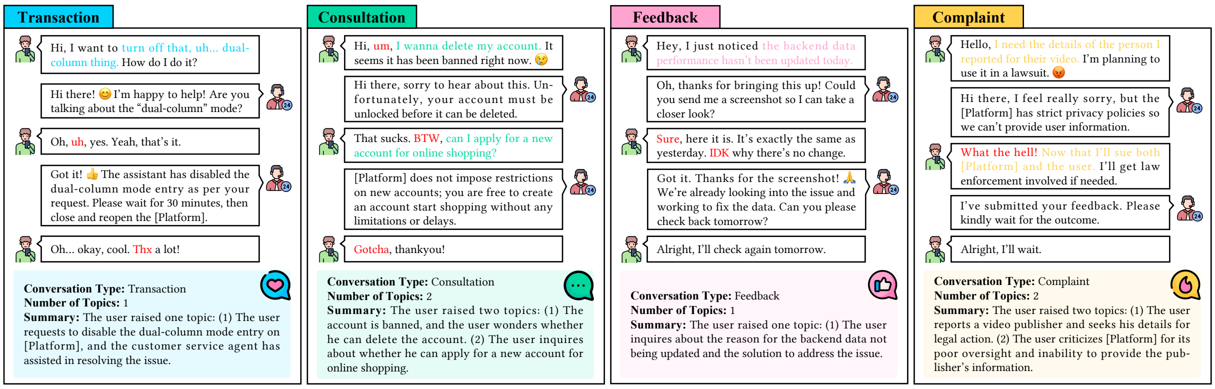
1 Introduction

Large Language Models (LLMs), such as ChatGPT (Ouyang et al., 2022), GPT-4 (Achiam et al., 2023) and LLaMA-series models (Touvron et al., 2023; AI, 2024), have achieved exceptional performance across diverse tasks, e.g. question answering, creative writing and dialogue. Despite the remarkable success, LLMs are prone to hallucinations, producing content that appears plausible but is factually incorrect or nonsensical (Wang et al., 2023). Such hallucinations would mislead users, undermine trust, and pose risks, hindering the deployment of LLMs in real-world applications.

There are generally two types of hallucinations in LLMs, *factuality hallucination* and *faithfulness hallucination*, where the former emphasizes the inconsistency of generated content and world knowledge, while the latter captures the divergence of generated content from the provided context (Huang et al., 2023). Existing hallucination evaluation benchmarks (Lin et al., 2021; Pal et al., 2023; Cheng et al., 2023; Wei et al., 2024; Oh et al., 2024) focus on assessing factuality hallucinations in LLMs, and only a handful of benchmarks (Tang et al., 2024; Ming et al., 2024) have been proposed to evaluate faithfulness hallucinations. Among them, TofuEval (Tang et al., 2024) evaluates the relevance of the LLMs' generated summary of a dialogue in an interview or meeting conversation to its topic; and FaithEval (Ming et al., 2024) assesses the faithfulness of the LLMs' response w.r.t. the provided context in a single-turn question answering setting.

However, real-world conversations often involve **casual language** (e.g., interjections, emojis, and abbreviations) and **topic shifts**. In this work, we propose to utilize real-world **customer service conversations** to evaluate faithfulness hallucinations in LLMs, which has been the first such attempt in literature. Some examples of the conversations between users and human customer service agents are given in Fig. 1, which are produced on an active online customer service platform that supports over 20 products across various domains, including e-commerce, advertising, finance, etc. It can be observed that user interactions in these conversations often use casual language, including interjections (e.g. "uh" and "um"), emojis (e.g., 😊 and 😢), and abbreviations (e.g. "BTW" and "IDK"). Also, multiple topics may be involved within one single conversation, like topics shifting from "request user information" to "claim to sue" as shown in the fourth sub-figure "Complaint". These cases would definitely pose great challenges to LLMs in accurately interpreting the context.

In light of these observations, we propose CONVFAITHEVAL, the first faithfulness hallucination evaluation benchmark constructed based on prac-



Transaction

Hi, I want to turn off that, uh... dual-column thing. How do I do it?

Hi there! 😊 I'm happy to help! Are you talking about the "dual-column" mode?

Oh, uh, yes. Yeah, that's it.

Got it! 👍 The assistant has disabled the dual-column mode entry as per your request. Please wait for 30 minutes, then close and reopen the [Platform].

Oh... okay, cool. Thx a lot!

Conversation Type: Transaction
Number of Topics: 1
Summary: The user raised one topic: (1) The user requests to disable the dual-column mode entry on [Platform], and the customer service agent has assisted in resolving the issue.

Consultation

Hi, um, I wanna delete my account. It seems it has been banned right now. 😢

Hi there, sorry to hear about this. Unfortunately, your account must be unlocked before it can be deleted.

That sucks. BTW, can I apply for a new account for online shopping?

[Platform] does not impose restrictions on new accounts; you are free to create an account start shopping without any limitations or delays.

Gotcha, thankyou!

Conversation Type: Consultation
Number of Topics: 2
Summary: The user raised two topics: (1) The account is banned, and the user wonders whether he can delete the account. (2) The user inquires about whether he can apply for a new account for online shopping.

Feedback

Hey, I just noticed the backend data performance hasn't been updated today.

Oh, thanks for bringing this up! Could you send me a screenshot so I can take a closer look?

Sure, here it is. It's exactly the same as yesterday. IDK why there's no change.

Got it. Thanks for the screenshot! 🙏 We're already looking into the issue and working to fix the data. Can you please check back tomorrow?

Alright, I'll check again tomorrow.

Conversation Type: Feedback
Number of Topics: 1
Summary: The user raised one topic: (1) The user inquires about the reason for the backend data not being updated and the solution to address the issue.

Complaint

Hello, I need the details of the person I reported for their video. I'm planning to use it in a lawsuit. 😡

Hi there, I feel really sorry, but the [Platform] has strict privacy policies so we can't provide user information.

What the hell! Now that I'll sue both [Platform] and the user. I'll get law enforcement involved if needed.

I've submitted your feedback. Please kindly wait for the outcome.

Alright, I'll wait.

Conversation Type: Complaint
Number of Topics: 2
Summary: The user raised two topics: (1) The user reports a video publisher and seeks his details for legal action. (2) The user criticizes [Platform] for its poor oversight and inability to provide the publisher's information.

Figure 1: An illustration of CONVFAITHEVAL benchmark (English-translated).It contains four types of customer service conversations: *Transaction*, *Consultation*, *Feedback*, and *Complaint*, each containing one or multiple topics and using casual language (e.g., interjections, emojis, and abbreviations). We also generate the number of topics and a summary for each conversation.

tical conversations. We design two tasks in CONVFAITHEVAL, *Conversation Summarization* and *Quiz Examination*, for a comprehensive evaluation of faithfulness hallucinations in LLMs. To construct this benchmark, we first select and filter conversations from a Chinese online customer service platform, and apply anonymization to ensure data privacy. For each conversation, we adopt GPT-4 to automatically identify its conversation type and number of topics, as well as to generate a summary and a quiz. Lastly, we perform strict human verifications to ensure the high quality of the benchmark. In statistics, CONVFAITHEVAL comprises $3,369$ practical conversations as the evaluation set, with over 88% of them involving multiple topics.

Based on the constructed CONVFAITHEVAL, we examine 23 LLMs, including 16 open- and 7 closed-source LLMs that have shown impressive performance in many tasks. Our results reveal that almost all LLMs suffer from faithfulness hallucinations on both tasks; comparably, closed-source LLMs consistently outperform open-source LLMs on both tasks across various metrics. To explore mitigating faithfulness hallucinations in LLMs, we examine the application of four strategies: Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thought (ToT) (Yao et al., 2024), Few-Shot prompting (Brown et al., 2020), and Supervised Fine-Tuning (SFT). To implement the SFT strategy, we construct a training set of $30,000$ conversations and corresponding summaries. Experimental results show that SFT yields an obvious improvement compared to the other three strategies on both tasks. For instance, the fine-tuned Qwen2.5-72B (Bai et al., 2023) achieves a substantial performance boost on the conversation summarization task, raising the F1 score from 59.2% to 73.6%, making it competitive with closed-source LLMs, whereas the other three strategies show no significant improvement.

We make three major contributions in this work: (1) We evaluate faithfulness hallucinations in LLMs with user conversations containing casual language and topic shifts, which are ubiquitous in real-world scenarios. (2) We construct CONVFAITHEVAL based on real-world customer service conversations, consisting of $30,000$ and $3,369$ conversations in its training and evaluation set, respectively. (3) With CONVFAITHEVAL, we design two tasks to comprehensively evaluate faithfulness hallucinations in 23 LLMs, explore four strategies to tackle the hallucination issue, and provide valuable insights, facilitating the research on hallucination mitigation in future.

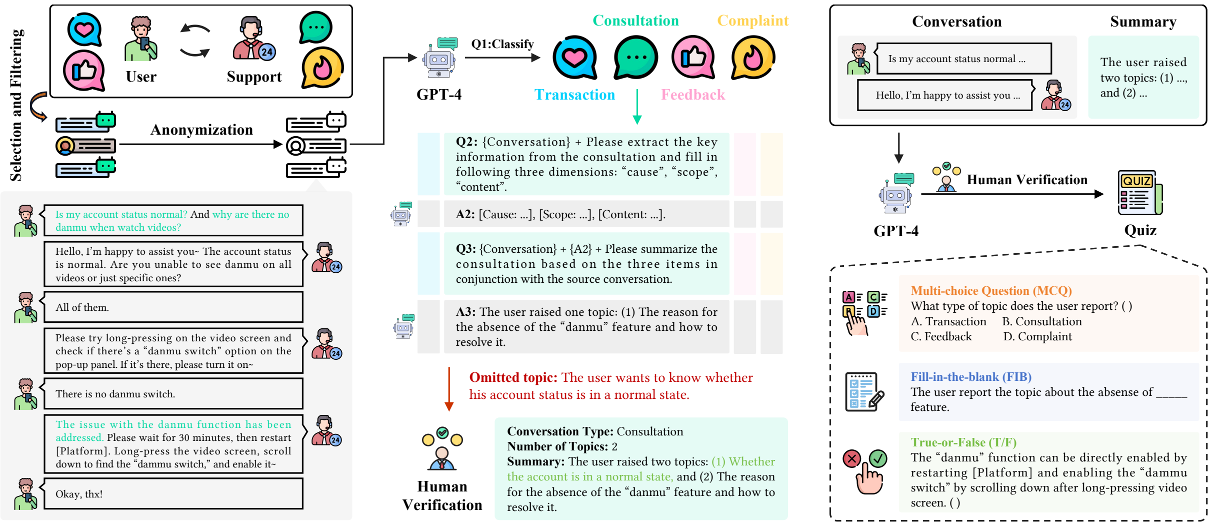
2 CONVFAITHEVAL Benchmark

We elaborate on the construction of the CONVFAITHEVAL benchmark, as shown in Fig. 2.

2.1 Conversation Collection

As shown in Fig. 2 (a), we first collect and filter raw conversations from a customer service platform, followed by a comprehensive anonymization process to ensure data privacy.

Raw conversation selection and filtering. The online platform from which we extract raw data supports services for over 20 products across various domains, including e-commerce, advertising, finance, etc. To construct CONVFAITHEVAL, we engage employees who are familiar with the platform and its products as human annotators. The annotators first select and filter raw conversations



Is my account status normal? And why are there no danmu when watch videos?

Hello, I'm happy to assist you~ The account status is normal. Are you unable to see danmu on all videos or just specific ones?

All of them.

Please try long-pressing on the video screen and check if there's a "danmu switch" option on the pop-up panel. If it's there, please turn it on~

There is no danmu switch.

The issue with the danmu function has been addressed. Please wait for 30 minutes, then restart [Platform]. Long-press the video screen, scroll down to find the "dammu switch," and enable it~

Okay, thx!

Q2: {Conversation} + Please extract the key information from the consultation and fill in following three dimensions: "cause", "scope", "content".

A2: [Cause: ...], [Scope: ...], [Content: ...].

Q3: {Conversation} + {A2} + Please summarize the consultation based on the three items in conjunction with the source conversation.

A3: The user raised one topic: (1) The reason for the absence of the "danmu" feature and how to resolve it.

Omitted topic: The user wants to know whether his account status is in a normal state.

Conversation Type: Consultation
Number of Topics: 2
Summary: The user raised two topics: (1) Whether the account is in a normal state, and (2) The reason for the absence of the "danmu" feature and how to resolve it.

Is my account status normal ...

Hello, I'm happy to assist you ...

The user raised two topics: (1) ..., and (2) ...

Multi-choice Question (MCQ)
What type of topic does the user report? ( )
A. Transaction B. Consultation
C. Feedback D. Complaint

Fill-in-the-blank (FIB)
The user report the topic about the absense of _____ feature.

True-or-False (T/F)
The "danmu" function can be directly enabled by restarting [Platform] and enabling the "dammu switch" by scrolling down after long-pressing video screen. ( )

(a) Conversation Collection (b) Summary Generation and Verification (c) Quiz Generation and Verification

Figure 2: An illustration of the construction pipeline for our ConvFaithEval benchmark.

from historical conversations on the platform following four guidelines at below. (1) Random selection: We select the conversations randomly to reflect the distribution of real-world conversations on the platform. (2) Contextual integrity: We remove the conversations that are excessively short or incomplete, which tend to lack essential context or valuable information. (3) Sensitive conversation filtering: We exclude the conversations that contain politically sensitive, pornographic, or violent information, to ensure that only appropriate data adhering to ethical standards are collected. (4) Noisy conversation filtering: We filter out the conversations with excessive noise, such as those with overabundant emojis, non-standard grammar, or emotional expressions, which may damage the clarity and coherence of the content.

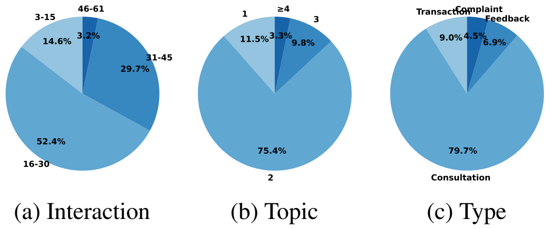
Finally, we obtain 50,000 high-quality raw conversations from over 1 million conversations on the customer service platform for further processing and annotation.

Conversation anonymization. To protect user privacy and ensure compliance with data protection regulations, we implement comprehensive anonymization on collected conversations, with following steps. (1) Personal information redaction: We develop a program to automatically detect and remove personal information (like user names, phone numbers, addresses, and other unique identifiers) from the conversations. (2) Entity replacement: In cases where removing specific entities might result in a loss of contextual information, we substitute user and platform information with generic placeholders. For example, user names are replaced with the token "[user]", and platform names with "[platform]", ensuring privacy while maintaining readability. (3) Context obfuscation: For the conversations containing detailed location references or contextual information leading to indirect identification of an entity, we generalize the content to its more generic forms. For example, specific dates, locations, and numerical values are converted to broader terms, like "March 15, 2024" to "in spring 2024". (4) Metadata removal: All metadata associated with the conversations, such as timestamps, IP addresses, and device information, is stripped away. This prevents any backward identification based on indirect attributes or usage patterns. Then, our annotators are asked to validate and revise the algorithmically anonymized content, ensuring the accuracy and completeness of the anonymization.

2.2 Summary Generation and Verification

As shown in Fig. 2 (b), we generate a summary for each collected conversation with GPT-4, which will then be verified to ensure its correctness. More details are presented in Appx. B.1.

Summary generation. We adopt GPT-4 to generate an initial summary for each conversation, which should capture key information about the conversation and conclude all topics it involves, with main concerns and queries clearly articulated. We utilize the Tree of Thoughts (ToT) (Yao et al., 2024; Long, 2023) framework to instruct GPT-4 to complete the summary generation operation. (1) We provide GPT-4 with definitions of four conversation types, i.e. *Transaction*, *Consultation*, *Feedback*, and *Complaint*, as context information and instruct



(a) Interaction (b) Topic (c) Type

Figure 3: The distributional statistics on CONVFAITHEVAL evaluation set. “Interaction”, “Topic”, and “Type” denote interaction turns (each user or customer service reply counts as one turn), number of topics, and conversation types, respectively.

it to categorize the conversation. For those that fail to be categorized to any given type, GPT-4 would label it as “invalid conversation”. (2) Based on the detected type, we instruct GPT-4 to generate an outline for the conversation. Specifically, we first define a tailored prompt template with placeholders for each conversation type; we then utilize the template to guide GPT-4 to fill in the placeholders with extracted information from the conversation. For example, for a *Consultation* conversation, we instruct GPT-4 to extract “cause”, “scope”, and “content” from it and replace the corresponding placeholders with these elements. (3) We instruct GPT-4 to generate a precise summary for each conversation following the outline. In the generated summary, we ask GPT-4 to produce a brief description for each involved topic with a single sentence and list all topics and descriptions in a point-by-point format to ensure clarity and readability.

To enhance the quality of generated summaries, we apply few-shot prompting (Brown et al., 2020) and Retrieval-Augmentation Generation (RAG) (Lewis et al., 2020) techniques. In particular, we provide GPT-4 with examples of inputs and outputs at each step to enhance its generation. Also, we integrate an internal knowledge base comprising explanations for proprietary terms and business abbreviations, which are commonly used on the platform, as external knowledge. Thereby, we first extract keywords from a conversation, and then retrieve corresponding explanations from this knowledge base, which are then combined with the input conversation to guide GPT-4.

Summary verification. To enhance correctness of generated conversation summaries, we employ human annotators to perform careful reviews and fix any errors identified, based on following guidelines. (1) Summary correction: If a summary is found to incorrectly, inadequately, or additionally conclude the topics in the conversation, human annotators would revise it to an accurate and complete version. For the summaries describing multiple topics, human annotators would improve clarity by rephrasing or reorganizing the content. (2) Conversation categorization: All conversations are categorized into four types. In particular, *Transaction* refers to requesting customer service assistance for specific tasks, such as processing returns, filing appeals, or handling refunds; *Consultation* involves seeking advice on issues, like investigating reasons behind a problem, exploring solutions, or accessing certain features; *Feedback* covers user inputs on various aspects, such as reporting bugs, suggesting new features, or commenting on content and usability; *Complaint* refers to reporting dissatisfaction regarding an individual, service, or platform. (3) Sensitive information removal: Human annotators would check and ensure that all personal and platform-related information has been completely removed or anonymized.

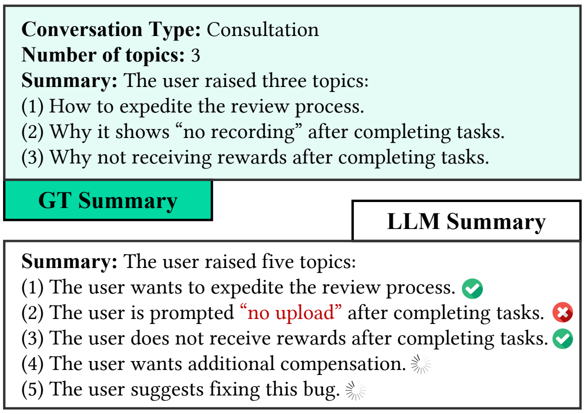
For the training set, human annotators are asked to only verify and revise generated summaries, leaving out the conversation type and number of topics.

2.3 Quiz Generation and Verification

In addition to summarization, we also create a quiz for each conversation in the evaluation set to better assess faithfulness hallucinations in LLMs, as shown in Fig. 2 (c). See more details in Appx. B.2.

Quiz generation. We first utilize GPT-4 to generate a diagnostic quiz with inputs of a conversation and its human-verified summary. The quiz consists of three types of questions: multi-choice question (MCQ), fill-in-the-blank (FIB), and true-or-false (T/F), with two questions per type per conversation. We use GPT-4 to generate plausible distractors for MCQ, identify and mask key phrases based on semantic importance for FIB, and synthesize assertions for T/F to test models' fine-grained understanding about the conversation.

Quiz verification. We perform human verification to guarantee correctness of the generated quiz, following three steps. (1) Content review: Human annotators are instructed to carefully examine each question to ensure its correctness and relevance. (2) Difficulty calibration: The difficulty of each question is evaluated to ensure that the quiz is making a challenge that appropriately aligns with the conversational complexity. (3) Redundancy elimination: Human annotators identify redundant or repetitive questions, ensuring the diversity of queries covering different aspects of the conversation.



Conversation Type: Consultation
Number of topics: 3
Summary: The user raised three topics:
(1) How to expedite the review process.
(2) Why it shows "no recording" after completing tasks.
(3) Why not receiving rewards after completing tasks.

Summary: The user raised five topics:
(1) The user wants to expedite the review process. ✓
(2) The user is prompted "no upload" after completing tasks. ✗
(3) The user does not receive rewards after completing tasks. ✓
(4) The user wants additional compensation. ◌
(5) The user suggests fixing this bug. ◌

Figure 4: An illustration of correct ✓, erroneous ✗, and fabricated ◌ topics in conversation summarization. The "erroneous" indicates the topic is involved but contains incorrect details; the "fabricated" means the topic is not mentioned in the conversation.

2.4 Statistics and Analysis

The constructed benchmark contains an evaluation set for evaluating faithfulness hallucinations in LLMs and a training set for tuning LLMs.

Evaluation set. A total of $3,369$ diverse conversations are included, each annotated with "conversation type", "number of topics", and "summary". We present the statistics about the evaluation set in Fig. 3. It is observed that 85.5% of conversations involve over 15 interaction turns, and 88.5% include multiple topics. Meanwhile, conversation types within are primarily dominated by *Consultation* (79.7%), with *Complaint* (4.5%) being the least, aligning with real-world distributions. For quiz collection, quizzes are successfully generated for 3,297 conversations, resulting in a total of $19,782$ questions.

Training set. Total $30,000$ conversations are included, each annotated with a "summary".

3 Experiments

3.1 Experimental Setup

Evaluation tasks and metrics. On the proposed ConvFaithEval, we evaluate faithfulness hallucinations in LLMs with two tasks: *Conversation Summarization* and *Quiz Examination*. More details are provided in Appx. B.3 and Appx. B.4.

(1) *Conversation Summarization*: LLMs are required to generate a summary to describe all topics discussed in a given conversation. The generated summary is then compared against the human-verified ground-truth (GT) in ConvFaithEval. Since manual evaluation is time-consuming and labor-intensive, we use GPT-4 as the discriminator to judge whether the generated summary is hallucinated following previous practices (Cheng et al., 2023; Liang et al., 2023; Zhu et al., 2024).

We define five metrics for evaluations on this task. Formally, given a conversation containing m topics, the LLM generates summaries containing n topics, which are categorized into a correct, b erroneous, and c fabricated topics, where $n = a + b + c$, as shown in Fig. 4. The formula of the five metrics are as follows: Omission Rate ($\boldsymbol{O} = 1 - \frac{a}{m}$), Error Rate ($\boldsymbol{E} = \frac{b}{n}$), Fabrication Rate ($\boldsymbol{F} = \frac{c}{n}$), Recall ($\boldsymbol{R} = \frac{a}{m}$), Precision ($\boldsymbol{P} = \frac{a}{n}$), and F1 Score (**F1**).

(2) *Quiz Examination*: We instruct LLMs to answer questions in the quiz based on the conversation. We report the accuracy of each question type, i.e. MCQ, FIB, and T/F, as well as the average accuracy as the quiz score.

Baselines. We evaluate on our ConvFaithEval various LLMs that span different versions and scales, including open-source LLMs[1]: InternLM (InternLM, 2023), Qwen (Bai et al., 2023), LLaMA (Touvron et al., 2023; AI, 2024), and GLM (Du et al., 2022), and closed-source ones[2]: GPT-series (OpenAI, 2022, 2023), Claude-series (Anthropic, 2024), and Gemini (Anil et al., 2023). To reduce randomness in LLM responses, we set all temperature parameters to zero. Note that we apply one-shot prompting to all LLMs for both evaluation tasks to ensure consistency of the response format.

3.2 Main Results

We report the main evaluation results on 23 LLMs in Table 1, and make the following observations. (1) Closed-source LLMs consistently outperform open-source LLMs on both evaluation tasks across various metrics. For conversation summarization, Claude 3.5 Sonnet achieves the highest **F1** of 80.4%, surpassing the best open-source model, Qwen2.5-72B, which scores 59.2%. Furthermore, closed-source LLMs such as GPT-4o and Claude 3.5 Sonnet show superior $\boldsymbol{R}$, with 82.9% and 80.6%, respectively, compared to open-source LLMs like InternLM2.5-20B (78.6%) and Qwen2.5-72B (72.7%). For quiz examination, closed-source LLMs continue their leading places,

[1] InternLM-7B, InternLM-20B, InternLM2.5-7B-Chat and InternLM2.5-20B-Chat; Qwen2-7B, Qwen2-72B, Qwen2.5-7B, and Qwen2.5-72B; LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, LLaMA-3-8B, LLaMA-3-70B, LLaMA-3.1-8B, and LLaMA-3.1-70B; ChatGLM2-6B and ChatGLM3-6B.

[2] GPT-3.5, GPT-3.5-Turbo, GPT-4, and GPT-4o; Claude 3 Opus and Claude 3.5 Sonnet; Gemini 1.5 Pro.

LLMs		Conversation Summarizaton						Quiz Examination			
		O ↓	E ↓	F ↓	R ↑	P ↑	F1 ↑	MCQ ↑	FIB ↑	T/F ↑	Average ↑
InternLM	InternLM-7B	30.1	45.7	20.7	69.9	35.7	47.2	54.7	58.9	60.8	58.1
	InternLM-20B	26.1	40.3	16.2	73.9	44.3	55.3	58.9	61.9	62.8	61.2
	InternLM2.5-7B	22.7	41.6	17.1	77.3	41.3	53.8	57.1	60.1	62.9	60.0
	InternLM2.5-20B	**21.4**	38.3	15.5	**78.6**	46.3	58.3	64.9	64.6	65.2	64.9
Qwen	Qwen2-7B	32.5	46.8	23.8	67.5	33.9	45.1	93.7	79.5	86.3	86.5
	Qwen2-72B	29.1	38.9	15.4	70.9	42.7	53.3	91.5	**81.5**	87.6	86.9
	Qwen2.5-7B	33.4	41.7	20.9	66.7	36.8	47.4	94.4	80.2	88.6	87.7
	Qwen2.5-72B	27.3	35.6	**13.7**	72.7	**49.9**	**59.2**	**96.0**	77.1	**91.5**	**88.2**
LLaMA	LLaMA-2-7B	56.7	62.4	34.7	43.3	19.9	27.3	46.2	42.9	47.9	45.7
	LLaMA-2-13B	53.1	60.3	30.3	46.9	22.4	30.3	50.8	46.9	55.7	51.1
	LLaMA-3-8B	53.9	57.6	31.9	46.1	21.6	29.5	47.9	43.3	48.9	46.7
	LLaMA-3-70B	49.8	53.2	27.6	50.2	25.9	34.2	54.9	59.6	60.2	58.2
	LLaMA-3.1-8B	52.1	55.9	30.9	47.9	23.2	31.3	49.2	45.8	50.3	48.4
	LLaMA-3.1-70B	30.7	44.7	22.8	69.4	35.7	47.1	57.9	62.9	63.3	61.4
GLM	ChatGLM2-6B	33.6	47.9	40.7	66.4	30.2	41.5	10.8	61.3	66.2	46.1
	ChatGLM3-6B	26.8	**33.2**	33.4	73.2	33.4	45.9	14.8	68.2	71.7	51.6
GPT	GPT-3.5	19.2	22.5	10.0	80.8	68.7	74.3	90.7	83.6	85.8	86.7
	GPT-3.5-Turbo	17.8	21.1	9.9	82.2	70.9	76.1	91.9	84.5	87.6	88.0
	GPT-4	**17.1**	19.7	9.1	**82.9**	73.7	78.0	93.2	85.4	87.0	88.5
	GPT-4o	18.1	**18.1**	**7.9**	81.9	78.7	80.2	**94.1**	**85.8**	87.6	**89.2**
Claude	Claude 3 Opus	21.8	20.5	9.3	78.2	79.4	78.8	91.7	83.8	87.3	87.6
	Claude 3.5 Sonnet	19.4	19.4	9.0	80.6	**80.2**	**80.4**	93.6	83.2	**88.5**	88.4
Gemini	Gemini 1.5 Pro	22.4	18.8	10.4	77.7	79.4	78.5	91.1	82.7	85.6	86.5

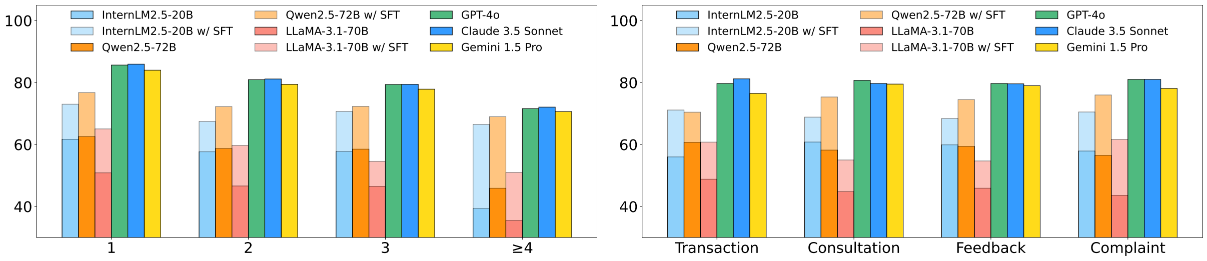
Table 1: Main results. We evaluate 23 LLMs across seven families with different versions and scales, on Conversation Summarization and Quiz Examination tasks. MCQ, FIB, and T/F refer to multiple-choice question, fill-in-the-blank, and true-or-false. All results are shown in percentages (%), and the best and second-best results are marked in **bold** and underline for open- and closed-LLMs, respectively.

with GPT-4o achieving the highest average score of 89.2% and Claude 3.5 Sonnet following closely at 88.4%. GPT-4o also excels in the T/F subtask, scoring 85.8%, while Claude 3.5 Sonnet leads in FIB with a score of 88.5%. These results highlight the stronger performance of closed-source LLMs in handling both summarization and understanding tasks, which can be attributed to the comprehensive closed-source resources and optimization strategies employed in their development, giving them a clear edge over open-source LLMs in mitigating faithfulness hallucinations. (2) LLMs with more advanced versions and larger scales consistently outperform their inferior counterparts on both tasks. For example, within the InternLM family, InternLM-20B achieves higher scores in $\boldsymbol{P}$, $\boldsymbol{R}$, and **F1** than the smaller InternLM-7B, with improvements in **F1** (55.6% vs. 50.0%) and $\boldsymbol{R}$ (46.3% vs. 35.7%). Similarly, for the Qwen family, Qwen2.5-72B demonstrates superior performance across multiple metrics compared to Qwen2-7B and Qwen2-72B, particularly in $\boldsymbol{R}$ (49.9% vs. 42.7%) and **F1** (59.2% vs. 53.3%). This pattern is not limited to open-source LLMs; closed-source LLMs exhibit improved performance with larger versions as well, where GPT-4, with a parameter size considerably larger than GPT-3.5, leads in **F1** (85.4%) and $\boldsymbol{R}$ (82.9%) compared to GPT-3.5. These results emphasize the positive correlation between model version/size and performance, revealing both more advanced versions and larger models enable more accurate and less hallucinated outputs. (3) Chinese LLMs outperform non-Chinese counterparts within the open-source category on both tasks. Our benchmark focuses on hallucinations in Chinese, and on it Chinese LLMs consistently exhibit better performance, aligning with expectations. For example, Qwen2.5-72B, a Chinese LLM, outperforms LLaMA-3.1-70B on both tasks, achieving higher **F1** (59.2% vs. 47.1%) and **F1** (59.2% vs. 47.1%). Similarly, InternLM2.5-7B and InternLM2.5-20B, also Chinese LLMs, exhibit stronger performance in these tasks when compared to their LLaMA counterparts, with the former beating LLaMA-2-7B in both tasks.

3.3 In-Depth Analysis

To analyze the impact of topic shift and topic domain on triggering LLM hallucinations, we examine model performance w.r.t. the "number of topics" and "conversation type" of the evaluation set.

Performance w.r.t. number of topics. Fig. 5 (a) shows the performance comparison in **F1** of vari-



(a) Number of topics

(b) Conversation type

Figure 5: In-depth analysis. We select six representative LLMs and analyze the impact of number of topics (a) and conversation type (b) upon **F1** on *Conversation Summarization*. For comparison, we also report the results of three open-source LLMs after applying SFT on our training set.

ous LLMs on the conversation summarization task across conversations involving different numbers of topics. The analysis reveals three key observations. (1) As the number of topics increases, there is a noticeable decline in performance for all LLMs, including advanced closing-source LLMs like GPT-4o. This highlights the inherent challenge of managing hallucinations in more complex conversational scenarios, where topic shifts are more frequent. (2) Closed-source LLMs consistently outperform their open-source counterparts in all tests, demonstrating their advantages in reducing faithfulness hallucinations. (3) Confronted with such hallucinations, SFT substantially improves the performance of open-source LLMs, enabling them to achieve comparable performance to closed-source ones (Qwen2.5-72B w/ SFT), thereby highlighting the importance of task-specific adaptations in narrowing the domain gap with our training set.

Performance w.r.t. conversation type. Fig. 5 (b) compares the model performance across four conversation types. The tested LLMs do not exhibit significant differences in performance across the four conversation types, indicating their relatively uniform handling capability of different conversational types. Notably, SFT still leads to substantial improvements across all conversation types, further validating this strategy's effectiveness and highlighting the importance of the training set in addressing domain gaps from a data perspective.

3.4 Faithfulness Hallucination Mitigation

We explore applying four strategies to mitigate faithfulness hallucinations in LLMs. Among them, (1) Chain of Thoughts (CoT) enhances reasoning by breaking tasks into sequential steps (Wei et al., 2022); (2) Tree of Thoughts (ToT) extends CoT by exploring multiple reasoning paths (Yao et al., 2024); (3) Few-shot prompting uses a few examples in the prompt to make LLMs learn tasks with minimal data (Brown et al., 2020), and we adopt three-shot compared to the baseline (one-shot); (4) Supervised Fine-tuning (SFT) trains LLMs on our labeled training set, refining their task-specific performance. More details on prompt designs are shown in Appx. B.5.

The results are shown in Table 2, from which we make the following observations. (1) Although LLaMA exhibits a great improvement from 47.1% to 59.0% in **F1** after applying SFT, it still lags significantly behind other LLMs. This could be attributed to its inherently inferior capability in understanding Chinese. (2) CoT and ToT improve performance on both tasks for closed-source LLMs (80.4% to 82.0% for Claude 3.5) but exhibit limited effects for open-source LLMs. This could be attributed to the parameter scale between them. Larger scales could facilitate closed-source LLMs in better understanding long-form conversations, especially with multiple topics and casual languages. (3) Increasing prompting examples (from 1-shot to 3-shot) is beneficial for closed-source LLMs (*e.g.*, **F1** of Claude 3.5 from 80.4% to 81.7%) but not significant for open-source ones (*e.g.*, **F1** of Qwen2.5-72B from 59.2% to 58.8%). This may be due to the larger parameter scale of closed-source LLMs, allowing them to better utilize additional context in few-shot scenarios. (4) SFT significantly improves performance compared to prompt-based strategies. SFT offers a more substantial improvement by addressing data distribution discrepancies. In total, applying SFT to these three LLMs leads to significant improvements on both evaluation tasks: 54.8% to 67.4% in average **F1** and 71.5% to 81.0% in average accuracy. This highlights that SFT can help LLMs generalize better across diverse scenarios, resulting in a more effective approach compared to prompt adjustments.

LLMs	Conversation Summarization						Quiz Examination			
	O ↓	E ↓	F ↓	R ↑	P ↑	F1 ↑	MCQ ↑	FIB ↑	T/F ↑	Quiz Score ↑
InternLM2.5-20B	21.4	38.3	15.5	78.6	46.3	58.3	64.9	64.6	65.2	64.9
w/ SFT	**20.0**	**25.7**	**13.7**	**80.0**	**61.9**	**69.8**	**94.6**	**84.9**	**90.8**	**90.1**
Qwen2.5-72B	27.3	35.6	13.7	72.7	49.9	59.2	96.0	77.1	91.5	88.2
w/ CoT	26.7	36.8	13.1	73.3	50.5	59.8	**96.8**	77.9	92.2	88.9
w/ ToT	27.8	37.9	14.2	72.2	50.9	59.7	95.3	76.5	92.1	88.0
w/ Three-shot	27.1	37.9	13.7	72.9	49.2	58.8	95.0	77.5	91.2	87.9
w/ SFT	**21.1**	**19.3**	**9.4**	**78.9**	**69.0**	**73.6**	95.4	**80.5**	**93.2**	**89.7**
LLaMA-3.1-70B	30.7	44.7	22.8	69.4	35.7	47.1	57.9	62.9	63.3	61.4
w/ CoT	31.1	42.8	20.2	68.9	35.4	46.8	58.8	63.5	**63.8**	62.0
w/ ToT	31.4	43.7	23.7	68.6	35.1	46.5	57.5	62.9	62.2	60.8
w/ Three-shot	31.7	43.4	20.8	68.3	35.0	46.3	57.2	62.4	63.4	61.0
w/ SFT	**25.5**	**31.9**	**17.7**	**74.5**	**48.8**	**59.0**	**63.2**	**63.9**	62.7	**63.3**
GPT-4o	18.1	18.1	7.9	81.9	78.7	80.2	94.1	85.8	87.6	89.2
w/ CoT	16.9	17.5	**7.1**	83.1	**79.2**	**81.1**	93.6	86.3	86.7	88.9
w/ ToT	17.2	18.2	8.4	82.8	78.7	80.7	**94.9**	**86.5**	**88.1**	**89.8**
w/ Three-shot	**16.4**	**17.0**	7.6	**83.6**	78.1	80.8	94.4	86.1	87.7	89.4
Claude 3.5 Sonnet	19.4	19.4	9.0	80.6	80.2	80.4	93.6	83.2	88.5	88.4
w/ CoT	18.2	18.8	8.5	81.8	81.2	81.5	93.1	83.7	86.1	87.6
w/ ToT	**17.9**	**18.3**	**8.2**	**82.1**	81.9	**82.0**	93.6	**84.1**	88.2	**88.6**
w/ Three-shot	19.0	19.2	8.2	81.0	**82.4**	81.7	**93.8**	83.8	**89.0**	**88.9**
Gemini 1.5 Pro	22.4	18.8	10.4	77.7	79.4	78.5	91.1	82.7	85.6	86.5
w/ CoT	21.7	19.0	9.7	78.3	78.5	78.4	90.6	82.1	**86.8**	86.5
w/ ToT	22.1	19.4	9.9	77.9	**80.5**	**79.2**	91.5	**83.6**	85.0	**86.7**
w/ Three-shot	**21.0**	**18.7**	**9.5**	**79.0**	76.9	77.9	**91.7**	82.1	85.0	86.3

Table 2: Hallucination mitigation results. The best result of each LLM series is marked in **bold**.

4 Related Works

Hallucination benchmarks. Exisitng benchmarks for evaluating LLM hallucinations can be divided into two groups. The first group, *hallucination evaluation benchmark*, emphasizes the assessment of the extent of hallucinations in LLM responses, while the other, *hallucination detection benchmark*, focuses on evaluating the hallucination differentiation capabilities (Wang et al., 2023). For hallucination evaluation benchmarks, two types of hallucinations are considered: factuality hallucinations and faithfulness hallucinations. Most existing works focus on evaluating factuality hallucinations with factual questions (Lin et al., 2021; Cheng et al., 2023; Pal et al., 2023; Wei et al., 2024; Oh et al., 2024; Zhu et al., 2024), assessing whether the responses of LLMs contradict verified facts. Comparably, only a few works (Tang et al., 2024; Ming et al., 2024) evaluate faithfulness hallucinations, i.e. whether LLMs' responses are inconsistent with the provided context. For hallucination detection benchmarks, existing works investigate various aspects including hallucination granularity, context complexity, and topic varieties (Miao et al., 2023; Li et al., 2023; Zhao et al., 2024; Liang et al., 2023; Chen et al., 2024). Our CONVFAITHEVAL is constructed primarily for benchmarking faithfulness hallucination evaluation, which can also be applied for measuring LLMs' hallucination detection performance (see Appx. A).

Faithfulness hallucination mitigation. To mitigate faithfulness hallucinations, numerous task-specific solutions have been proposed in aligning context consistency (Gu et al., 2022; Choi et al., 2023; Lei et al., 2023) and logical consistency (Li et al., 2024; Paul et al., 2024; Xu et al., 2024). Meanwhile, general approaches, such as CoT, ToT, and few-shot prompting (Wei et al., 2022; Yao et al., 2024; Brown et al., 2020), have also demonstrated effectiveness in reducing LLM hallucinations. In this work, we explore applying four strategies to mitigate faithfulness hallucinations of LLMs on our CONVFAITHEVAL, and provide valuable insights.

5 Conclusion

In this paper, we introduce CONVFAITHEVAL, a benchmark for evaluating faithfulness hallucinations in LLMs using real-world customer service conversations. Unlike prior work, it considers the challenges of casual language and topic shifts in multi-turn conversations. We conduct evaluations of 23 LLMs on our new benchmark, revealing closed-source models outperforming open-source ones. We further explore hallucination mitigation strategies, and find that SFT is the most effective. This work not only proposes a robust framework for evaluating and enhancing the reliability and faithfulness of LLMs in real world, but also provides critical insights into hallucination mitigation.

6 Limitations

The limitations of this work primarily stem from the scope and structure of the ConvFaithEval benchmark. While the dataset is derived from real-world customer service conversations, it is focused on a single domain, potentially limiting its generalizability to other conversational contexts or languages beyond Chinese. Meanwhile, while SFT shows significant improvements, its applicability depends on access to high-quality labeled data, which may not always be feasible compared with general prompting techniques. Furthermore, the benchmark primarily assesses hallucinations related to faithfulness, leaving aspects such as user intent interpretation and contextual nuance underexplored. These limitations highlight opportunities for future work to broaden the dataset scope, explore additional evaluation dimensions, and develop methods requiring fewer labeled resources.

7 Ethical Statement

In this study, we adhere to strict ethical standards to ensure the responsible use of data and technology. All customer service conversations used in the ConvFaithEval benchmark were carefully anonymized to protect user privacy, removing personal and identifiable information through automated processes and thorough human review. The study complies with data protection regulations and ethical guidelines to prevent misuse of sensitive information. Furthermore, the benchmark and findings aim to improve the reliability and safety of LLMs, with the ultimate goal of reducing risks such as misinformation and user trust erosion in real-world applications. The research emphasizes transparency and accountability, encouraging the responsible development and deployment of LLMs.

References

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

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anthropic. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. 2024. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2403.00896*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection. *ArXiv preprint*, abs/2310.09044.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. 2022. Improving controllable text generation with position-aware weighted decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3449–3467, Dublin, Ireland. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

InternLM. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. *ArXiv preprint*, abs/2310.03951.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Towards faithful chain-of-thought: Large language models are bridging reasoners. *CoRR*, abs/2405.18915.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Euijong Whang. 2024. Erbench: An entity-relationship based automatically verifiable hallucination benchmark for large language models. *arXiv preprint arXiv:2403.05266*.

OpenAI. 2022. OpenAI: Introducing ChatGPT.

OpenAI. 2023. OpenAI: GPT-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *CoRR*, abs/2402.13950.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *CoRR*, abs/2405.18357.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Zhiying Zhu, Yiming Yang, and Zhiqing Sun. 2024. Halueval-wild: Evaluating hallucinations of language models in the wild. *arXiv preprint arXiv:2403.04307*.

A Results for Hallucination Detection

We additionally collect 404 representative conversations to benchmark hallucination detection, each of which is accompanied by both hallucinated and human-corrected (non-hallucinated) summaries during the summary annotation process in Sec. 2.2. In this context, LLMs are tasked to predict a binary outcome based on the input conversation and its corresponding summary. Results are shown in Fig. 6, where we have the following conclusions: **(1) Smaller LLMs struggle to detect hallucinations at the conversation level.** Given the limited parameter scales, smaller LLMs like InternLM2.5-7B, Qwen2-7B, and LLaMA-3.1-8B achieve detection accuracies around 50%, akin to random guessing. Their inability to capture sufficient context hampers accurate hallucination detection. **(2) Closed-source LLMs demonstrate superior hallucination detection performance to open-source ones.** Closed-source LLMs, such as GPT-4o (78.5%) and Claude 3.5 Sonnet (80.1%), consistently outperform open-source LLMs like LLaMA-3.1-70B (66.9%) and Qwen2.5-72B (72.1%). This performance gap suggests that proprietary data and optimizations in closed-source LLMs contribute to their enhanced ability to detect hallucinations. **(3) Scaling and advanced versions can facilitate LLMs in hallucination detection.** LLMs like GPT-4 and Claude 3.5, with greater parameter scales and refined training, show significantly higher detection accuracy, enabling them to better capture complex conversational patterns and detect hallucinations. The same applies to open-source LLMs.

B Implementation Details

B.1 In Summary Generation and Verification

In Sec. 2.2, we employ the Tree of Thoughts (ToT) framework (Yao et al., 2024; Long, 2023) to guide the automatic summary generation process through three main steps: (1) Conversation classification (Fig. 7): We classify the conversations into one of four predefined types: *Transaction*, *Consultation*, *Feedback*, and *Complaint*. (2) Outline extraction (Fig. 8, 10, 12, and 14): Using GPT-4, we prompt the model to generate corresponding fine-grained outlines based on the classified conversation type. (3) Summary generation (Fig. 9, 11, 13, and 15): Finally, GPT-4 is prompted to write a summary by synthesizing the conversation content and fine-grained outlines.

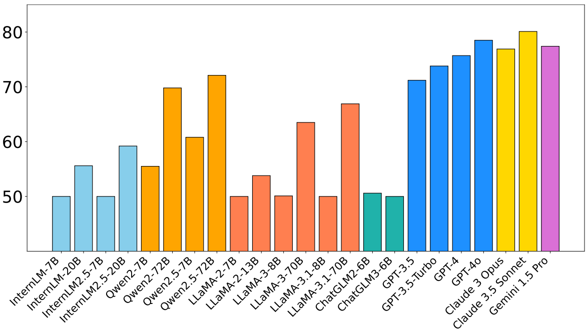


Figure 6: We compare the hallucination detection accuracies for 23 LLMs, with each series distinguished by a different color for clarity.

B.2 In Quiz Generation and Verification

In Sec. 2.3, We utilize GPT-4 to generate a diagnostic quiz with input from the conversation and its corresponding human-verified summary, including multiple-choice question (MCQ), fill-in-the-blank (FIB), and true-or-false (T/F). The detailed prompt structures are illustrated in Fig. 16.

B.3 In Conversation Summarization

This task encompasses two components of prompt design: (1) LLM summary generation: We instruct the LLM to generate a summary given a conversation, and (2) GPT-4 discrimination: We prompt GPT-4 to compare the LLM-generated summary with the ground truth summary for evaluation purposes. Detailed prompts are shown in Fig. 17 and Fig. 18.

B.4 In Quiz Examination

In this task, we instruct LLMs to respond to three question types (MCQ, FIB, and T/F) in our quiz. For each question type, a tailored prompt is used to ensure that the LLM can understand the context, interpret the requirements, and provide an accurate response. Detailed prompts are shown in Fig. 19, 20, and 21.

B.5 In Hallucination Mitigation

CoT prompting is employed to guide the process of summary generation through a systematic three-step framework. (1) Conversation classification (Fig. 22): The first step involves identifying the type of input content. Similar to ToT, we categorize the feedback into three distinct types. (2)

Step1: Conversation Classification

```
# 角色:
你是一名客服专家

## 目标:
- 准确对用户与客服的对话进行分类，判断属于「咨询类」、「反馈类」、「举报类」和「办事类」其中一种。如果用户反馈了多个问题，只关心用户多次提到且强烈关注的主要问题。

## 标签定义:
- 「咨询类」：用户咨询信息，包括xxx怎么搞/啥时候/xxx是什么/能否xxx/功能入口等。
举例1：原账户信用卡注销了，订单退款能否退到其他账户。
举例2：想开通店铺授权号，但需要先开通企业号，如果用营业执照开通会不会和店铺冲突
举例3：自己发布的视频为被封禁的原因
- 「反馈类」：用户投诉反馈或吐槽功能不好用/有bug、举报主播/现象、吐槽xxx功能/事件。
举例1：用户在下载【某平台】后，登录时同意个人信息后页面空白无反应的问题
举例2：【某平台】直播间遇到一位主播说脏话，认为该主播有赌博行为，希望平台能够处理。
举例3：用户反馈推荐的内容不喜欢，包含很多低质内容。
- 「办事类」：用户办啥事、诉求。希望客服/平台退款/申诉解封账号或店铺/取消限流/关闭xx功能/寻求帮助等。
举例1：用户在xx平台上发布视频后，有人在评论区辱骂他，希望能够屏蔽此类人员。
举例2：用户希望客服帮忙解封账号。
举例3：用户希望客服帮忙进行订单退款，订单改约。
- 「举报类」：用户对违规行为、内容或现象进行举报，期望平台介入并处理。举报内容通常针对违反平台规则的行为，包括诈骗、欺诈、恶意信息传播、色情或暴力内容、侵权行为等。
举例1：用户举报某直播间存在赌博行为，并提供相关证据截图，要求平台尽快处理。
举例2：用户发现某视频中存在欺诈信息（如假冒产品销售），希望平台下架该视频并对发布者进行处罚。
举例3：用户举报一位主播涉嫌传播不实信息，要求平台核查并封禁账号。
- 「无效反馈」：对话信息量不足，或者用户未有明确问题。

## 输出格式:
- 输出「咨询类」、「反馈类」、「办事类」、「举报类」和「无效反馈」其中一种。

## 以下是一段用户和客服对话:
<Conversation Here>

## 该对话属于「咨询类」、「反馈类」、「办事类」和「无效反馈」其中哪一种:
```

Figure 7: Prompts used in Summary Generation and Verification (Step1).

Outline extraction (Fig. 23): Based on the classification result, a fine-grained outline is generated. This outline serves as a blueprint, breaking down the input into its essential components. (3) Summary generation (Fig. 24): In the final step, the system utilizes both the classification type and the fine-grained outline to construct the summary. This framework ensures the generation of structured and contextually relevant summaries, tailored to the specific needs of diverse input tasks. Meanwhile, **ToT prompting** follows the same pipeline in Sec. B.1.

B.6 In Hallucination Detection

In Appx. A, we prompt LLMs to identify any discrepancies or hallucinations within the summaries to benchmark faithfulness hallucination detection. Detailed prompts are shown in Fig 25.

Step2: Outline Extraction (*Transaction*)

```
# 角色: 你是一名客服专家
## 目标:
- 给你一段用户与客服的对话，请你针对用户咨询的办事类问题进行要素提取，要素分3个维度：「范围」、「原因」、和「诉求」。
「范围」咨询的办事内容发生在哪个平台页面、功能比如「直播」、「账号」、「店铺」、「视频」等
「原因」遇到了什么事导致用户有诉求
「诉求」用户希望办啥事、诉求，比如 希望客服/平台退款/解封/取消限流/关闭xx功能/寻求帮助

## 输出格式:
「范围」：
「原因」：
「诉求」：

## 例子：
- 例子1:
以下是一段用户反馈
用户在视频评论区发现辱骂他，希望能够屏蔽此类人员
输出:
「原因」：有人在评论区辱骂他
「范围」：评论区
「诉求」：希望屏蔽辱骂他的人员
- 例子2:
以下是一段用户和客服对话：
客服: 客服代表【客服号】为您服务。
...
输出:
「范围」：账号
「原因」：用户未及时补充账号资料导致申请过期
「诉求」：希望客服重新发送补充资料的链接
- 例子3:
"客服: 客服代表【客服号】为您服务。
客服: 你好，小助手很高兴为您服务，请问有什么可以帮您?
用户: 唉，你好，我想问一下我这个我的【某平台】号被封禁了。
...
输出:
「范围」：账号
「原因」：用户账号可能因违规发布医疗相关内容被封禁
「诉求」：希望解封账号，且申诉失败，寻求其他解决办法
+
「范围」：注册
「原因」：用户账号被封禁，考虑能否用原手机号再注册
「诉求」：了解原手机号能否再注册【某平台】号

##要求：
1. 如果原文没有提到「范围」、「原因」、和「诉求」相关的内容，则对应位置输出“无”
2. 如果原文反馈不同的问题，则分多点输出，不同问题之间用“+”符号分割，最多输出2组问题要素，如果无额外问题内容则只输出1组问题要素。
## 以下是一段用户和客服对话：
<Conversation Here>
## 输出:
```

Figure 8: Prompts used in Summary Generation and Verification (Step2: *Transaction*).

Step3: Summary Generation (*Transaction*)

```
# 角色:
你是一名客服专家，擅长信息提炼与概括
## 目标:
- 给你一段用户与客服的对话，请你结合用户办事诉求的3个要素来简要总结用户反馈的问题，
要素分3个维度「范围」、「原因」、和「诉求」。
「范围」咨询的办事内容发生在哪个平台页面、功能比如「直播」、「账号」、「店铺」、「视频」等
「原因」遇到了什么事导致用户有诉求
「诉求」用户希望办啥事、诉求，比如 希望客服/平台退款/解封/取消限流/关闭xx功能/寻求帮助

## 例子：
- 例子1:
以下是一段用户和客服对话：
客服: 客服代表【客服号】为您服务。
客服: 您好，小助手很高兴为您服务，请问有什么可以帮您?
用户: 我我这个账号【某平台】为什么给我封掉了?
...
问题的要素：
「范围」：账号
「原因」：用户账号被封，申诉失败，且账号内还有钱
「诉求」：希望解封账号
结合以上问题要素该用户反馈的问题：用户账号因被封申诉失败且账号有钱希望能解封。
- 例子2:
客服: 客服代表【客服号】为您服务。
...
问题的要素：
「范围」：账号
「原因」：用户账号可能因违规发布医疗相关内容被封禁
「诉求」：希望解封账号，且申诉失败，寻求其他解决办法
+
「范围」：注册
「原因」：用户账号被封禁，考虑能否用原手机号再注册
「诉求」：了解原手机号能否再注册【某平台】号
结合以上问题要素该用户反馈的问题:
用户反馈了两个问题:
1. 用户【某平台】账号因违规被封申诉失败，希望解封
2. 用户咨询账号被封禁的原手机号能否再注册。

## 输出格式:
- 包括客户在什么范围因什么原因下的诉求，如果要素点「无」则跳过
- 输出字数限制在35字以内
- 逗号控制在1个以内
- 如果用户没有明确的问题则输出：用户没有明确问题。
- 如果用户没有明确说明原因，则不需要输出原因相关解释

## 以下是一段用户和客服对话：
<Conversation Here>
问题的要素：
<Fine-grained Outlines Here>

## 结合以上问题要素该用户反馈的问题:
```

Figure 9: Prompts used in Summary Generation and Verification (Step3: *Transaction*).

Step2: Outline Extraction (*Consultation*)

```
# 角色:
你是一名客服专家
## 目标:
- 给你一段用户与客服的对话，请你针对用户咨询的问题进行要素提取，要素分3个维度：「原因」、「范围」、「咨询内容」。
「原因」指什么原因让用户来咨询
「范围」（可选），咨询的内容发生在哪个平台页面、功能比如「直播」、「账号」、「店铺」、「视频」等
「咨询内容」xxx怎么搞/啥时候/xxx是什么/能否xxx等

## 输出格式:
「原因」：
「范围」：
「咨询内容」：

## 例子：
- 例子1:
以下是一段用户和客服对话：
客服: 客服代表为您服务。
客服: 你好，【用户】，很高兴为您服务，请问什么可以帮您?
用户: 唉，你好啊，我想问一下，我【某平台】账号的啊，私信功能被处罚了，不能用7天，我想问一下是为什么?
客服: 你好，先生！
客服: 噢，先生，您的问题就是说是你本次来电的这个账号私信被处罚了，然后咱要咨询这个处罚的原因，是这意思吧?
用户: 对对对对！
客服: 嗯，好的啊，就是看到你这个处罚了。先生，他是说提示为什么处罚，是说这个私信里边涉及色情低俗，是因为这个受到处罚的。
用户: 但是啊但是我我我账号里面唯一没作品然后因为噢我是因为反反正我这账号他也就是没有处罚之前我我我也是不能给没有互关的朋友这样子发信息的我记得是。
客服: 因为你看他给到你的提示说账号返有色情低俗相关内容，包括单不限于啊，这个你有查看到对吧，先生，就是因为这里边包括单不限于这个，就是因为这个受到处罚的，当然他也就是几天的处罚，到期之后就会自动解除处罚。但是我看到先生您进行申诉，申诉失败了，那申诉失败之后啊，只要以你的手机页面提示为准，就到期之后自动解除这个处罚了。
...
客服: 嗯，不可以啊，因为它是申诉，几乎只有一次。
用户: 噢，行。
客服: 嗯，确实很抱歉啊，先生。嗯，不能直接帮助到您，先生。
用户: 好行，谢谢。
客服: 嗯，好的啊，感谢您的理解，先生也辛苦您对对本人，对您的服务做个评价，那就不打扰您了，祝您生活愉快，再见！
客服: 嗯，好，拜拜。
输出:
「原因」：私信功能因涉及色情低俗被处罚，不能用7天
「范围」：账号、私信功能
「咨询内容」：账号私信功能被处罚的原因以及如何申诉提前解除
- 例子2:
以下是一段用户和客服对话：
用户: 我的视频为什么流量低
客服: 您的视频当前流量相较历史视频较低，主要可能是由于视频被粉丝观看的时间较短或者视频时长较长，视频完播的情况不是很理想哦～ 根据自身情况，决定是否要适当调整视频时间，尝试优化一下视频的完播情况。 可以在视频制作上增加更多能够引发用户观看、停留兴趣的内容，比如在剧情制作上增加反转内容、尝试使用一些特效、关联一些平台内热点话题等等方式。
用户: 线索管理：员工号/客服号设置
客服: 您这边是当前账号想要咨询【绑定员工个人号】的问题对吧
用户: 企业员工号可以继承换绑对吧
客服: 企业员工号可以通过离职重新启用 更换员工
客服: 企业员工号是指：企业申请创建新的【某平台】号分配给员工进行使用，员工离职之后可转交给其他员工继续使用该账号。
用户: 我的员工号登录不上了
客服: 企业员工号还是员工个人号呢 方便登录的页面发我下吗 辛苦您啦
客服: 关于您反馈的【员工号】问题，小助手帮您核实到当前设备环境无法安全运行人脸识别功能，为了您的账号安全，请卸载掉设备上可能安装的各种多开、分身、虚拟环境等软件或插件，恢复设备系统初始安全环境后重试 若您尝试以上步骤后无法正常使用刷脸功能，请更换设备后进行刷脸尝试 若依旧无法使用，请24H后面部无遮挡在光线好的地方本人刷脸尝试哦
客服: 小助手看到咱们之前给您处理过这个问题 您可以更换设备重新登录看下哈 辛苦您啦
用户: 都不行
用户: 就是频繁
客服: 咱们这个原因是因为同一个员工号多次识别登录导致的频繁哈 您可以过段时间重新登录看下哈 如果多次操作的话 系统识别频繁的
...
客服: 您好，由于长时间没有收到您的新消息，系统已暂时为您结束会话。若您当前问题暂未解决或有其他问题需要咨询，可以随时在当前页面再次咨询人工客服，小助手将随时为您服务。感谢您的理解，祝您生活愉快
输出:
「原因」：员工号登录不上
「范围」：企业员工号
「咨询内容」：企业员工号能否继承换绑、员工号登录不上的原因及解决办法
+
「原因」：视频流量较低
「范围」：视频
「咨询内容」：视频流量较低的原因及优化办法


##要求：

1. 如果原文没有提到「原因」或「范围」或「咨询内容」相关的内容，则对应位置输出“无”

2. 如果原文反馈不同的问题，则分多点输出，不同问题之间用“+”符号分割，最多输出2组问题要素，如果无额外问题内容则只输出1组问题要素。


## 以下是一段用户和客服对话：

<Conversation Here>

## 输出:
```

Figure 10: Prompts used in Summary Generation and Verification (Step2: *Consultation*).

Step3: Summary Generation (*Consultation*)

角色:

你是一名客服专家，擅长信息提炼与概括。

目标:

- 给你一段用户与客服的对话，请你针对用户咨询的3个要素内容结合原文来简要总结用户反馈的问题，要素分3个维度：「原因」、「范围」、「咨询内容」。「原因」指什么原因让用户来咨询。「范围」（可选），咨询的内容发生在哪个平台、功能比如「直播」、「账号」、「店铺」、「视频」等。「咨询内容」xxx怎么搞/啥时候/xxx是什么/能否xxx等。

## 输出格式:

- 包括客户在什么范围因什么原因咨询某事，如果要素点「无」则忽略过该要素

- 输出字数限制在35字以内，逗号控制在1个以内

- 如果用户没有明确的问题则输出：用户没有明确问题。如果有多组问题要素，则总结多个问题。如果用户没有明确说明原因，则不需要输出原因相关解释

例子

- 例子1:

以下是一段用户和客服对话：

...

用户: 留容组件找不到

...

「原因」：线索经营打不开

..

结合以上问题要素该用户反馈的问题:用户咨询【某平台】来客的线索经营功能如何打开

- 例子2:

...

用户: 没有团购入口

...

问题的要素：

..

「咨询内容」：没有团购入口，希望避开中午1点到2点回电

结合以上问题要素该用户反馈的问题：用户咨询找不到团购入口并希望避开特定回电时间

- 例子3:

以下是一段用户和客服对话：

用户: 我的视频为什么流量低

...

问题的要素：

「原因」：企业员工号登录不上

..

+

..

「咨询内容」：视频流量较低的原因

结合以上问题要素该用户反馈的问题:

用户反馈了两个问题：

1. 用户咨询企业员工号的继承换绑及登录不上的问题

2. 用户咨询视频流量低的原因

## 以下是一段用户和客服对话：

<Conversation Here>

问题的要素：

<Fine-grained Outlines Here>

## 结合以上问题要素该用户反馈的问题:

Figure 11: Prompts used in Summary Generation and Verification (Step3: *Consultation*).

Step2: Outline Extraction (*Feedback*)

```
# 角色:
你是一名客服专家，请你分析用户针对某个功能页面反馈内容。
## 目标:
- 给你一段用户与客服的对话，请你针对用户咨询的问题进行要素提取，要素分3个维度：「范围」、「场景」和「故障现象」。
「范围」能够明确问题/故障发生在哪个产品页面或者功能。
「场景」“正在做什么/准备做什么”的时候出现
「故障现象」具体出现的问题/故障的现象是什么

## 举例：
- 例子1:
以下是一段用户的问题：
用户在下载【某平台】后，登录时同意个人信息后页面空白无反应的问题。
问题的要素：
「范围」【某平台】
「场景」登录同意个人信息时
「故障现象」页面空白无反应
- 例子2:
以下是一段用户的问题：
...
用户: 为什么你们的【某平台】后台总是自己掉，每次都要重新登陆
客服: 我帮您备注清楚了 晚点您上线看下当前窗口的处理回复 最晚24小时内回复，这边就先交给我吧 有了结果会第一时间回复您的哈
用户: 而且为什么【某平台】的网页经常崩溃
客服: 收到，我们会及时优化
「范围」【某平台】
「场景」使用中
「故障现象」后台总是自动掉线，需要重新登陆
+
「范围」【某平台】
「场景」使用中
「故障现象」网页经常崩溃
##要求：
1. 如果原文没有提到「范围」或「场景」或「故障现象」相关的内容，则对应位置输出“无”
2. 如果原文反馈不同的大问题，则分多点输出，不同问题之间用“+”符号分割，最多输出2组问题要素，如果无额外问题内容则只输出1组问题要素。

## 以下是一段用户和客服对话：
<Conversation Here>
## 问题的要素：
```

Figure 12: Prompts used in Summary Generation and Verification (Step2: *Feedback*).

Step3: Summary Generation (*Feedback*)

角色:

你是一名客服专家，擅长信息提炼与概括

目标:

- 给你一段用户与客服的对话，请你结合用户反馈的问题要素来简要总结用户反馈的问题，

要素分多个维度：

「范围」能够明确问题/故障发生在哪个产品页面或者功能。

「场景」"正在做什么/准备做什么"的时候出现

「故障现象」具体出现的问题/故障的现象是什么

## 输出格式:

- 包括客户在什么范围或场景因什么原因反馈某事，如果要素点「无」则跳过该信息
- 输出字数限制在35字以内
- 逗号控制在1个以内
- 如果用户没有明确的问题则输出：用户没有明确问题。
- 如果有多组问题要素，则总结对应的多个问题。
- 如果用户没有明确说明原因，则不需要输出原因相关解释

例子

- 例子1:

以下是一段用户和客服对话：

用户: 长按视频这块怎么没有保存相册了，不好分享不了视频别扭，没有保存视频了，望改进

...

问题的要素：

「范围」无

「场景」长按视频时

「故障现象」无法保存视频到相册，不能分享视频。

结合以上问题要素该用户反馈的问题:用户反馈在长按视频时无法保存视频到相册，不能分享视频。

- 例子2:

以下是一段用户的问题：

...

用户: 为什么你们的【某平台】后台总是自己掉，每次都要重新登陆

客服: 我帮您备注清楚了 晚点您上线看下当前窗口的处理回复 最晚24小时内回复，这边就先交给我吧 有了结果会第一时间回复您的哈

用户: 而且为什么【某平台】的网页经常崩溃

客服: 收到，我们会及时优化

问题的要素：

「范围」【某平台】

「场景」使用中

「故障现象」后台总是自动掉线，需要重新登陆

+

「范围」【某平台】

「场景」使用中

「故障现象」网页经常崩溃

结合以上问题要素该用户反馈的问题:

用户反馈了两个问题：

1. 用户反馈【某平台】在使用中后台总是自动掉线，需要重新登陆。
2. 用户反馈【某平台】在使用中网页经常崩溃。

## 以下是一段用户和客服对话：

<Conversation Here>

问题的要素：

<Fine-grained Outlines Here>

## 结合以上问题要素该用户反馈的问题:

Figure 13: Prompts used in Summary Generation and Verification (Step3: *Feedback*).

Step2: Outline Extraction (*Complaint*)

```
# 角色:
你是一名客服专家，请你分析用户的举报内容。
## 目标:
- 给你一段用户与客服的对话，请你针对用户举报的问题进行要素提取，要素分3个维度：「范围」、「对象」和「举报原因」。
「范围」能够明确问题发生在哪个产品页面或者功能。
「对象」举报对象，包括主播、视频、商家等，
「举报原因」因为啥举报
「吐槽内容」指用户对xx行为的表态、xx的现象的不喜欢

## 举例：
例子1:
以下是一段用户的问题：
直播间遇到一位主播说脏话，且该主播在直播期间有赌博行为，希望平台能够处理。
问题的要素：
「范围」直播间
「对象」主播
「举报原因」说脏话、赌博行为
例子2:
以下是一段用户的问题：
用户: 不新鲜
...
问题的要素：
「范围」【某平台】生活服务
「对象」商家
「举报原因」餐品不新鲜、有异味、口味不好
- 例子3:
以下是一段用户的问题：
客服: 您好，您可以尝试描述遇到的问题或者点击下方问题列表，小助手也可以帮您解决哦
客服: 您好，小助手很高兴为您服务，请问有什么可以帮您?
用户: 停业了还卖卷
...
问题的要素：
「范围」【某平台】生活服务
「对象」无
「举报原因」商家停业了还卖券
+
「范围」直播
「对象」主播/直播间
「举报原因」主播违规，未被处理，投诉工单无结果且显示结束

##要求：
1. 如果原文没有提到「范围」或「对象」或「举报原因」相关的内容，则对应位置输出“无”
2. 如果原文反馈不同的大问题，则分多点输出，不同问题之间用“+”符号分割，最多输出2组问题要素，如果无额外问题内容则只输出1组问题要素。

## 以下是一段用户和客服对话：
<Conversation Here>
## 问题的要素：
```

Figure 14: Prompts used in Summary Generation and Verification (Step2: *Complaint*).

Step3: Summary Generation (*Complaint*)

```
# 角色:
你是一名客服专家，擅长信息提炼与概括
## 目标:
- 给你一段用户与客服的对话，请你结合用户反馈的问题要素来简要总结用户反馈的问题，
要素分多个维度：
「范围」能够明确问题/故障发生在哪个产品页面或者功能。
「对象」指举报或者吐槽抱怨的对象，包括用户、视频、商家、页面功能等
「举报原因」指用户举报的原因
「吐槽内容」用户对xx行为的表态、xx的现象的不喜欢

## 输出格式:
- 包括客户在什么范围或场景因什么原因反馈某事，如果要素点「无」则跳过该信息
- 输出字数限制在35字以内
- 逗号控制在1个以内
- 如果用户没有明确的问题则输出：用户没有明确问题。
- 如果有多组问题要素，则总结对应的多个问题。
- 如果用户没有明确说明原因，则不需要输出原因相关解释

## 例子
- 例子1:
以下是一段用户的问题：
用户: 别推荐游戏视频给我了 ...
问题的要素：
...
「故障现象」即使点击“不感兴趣”和进行相关设置，仍大量推荐游戏视频
+
「范围」无
...
结合以上问题要素该用户反馈的问题：
用户反馈了两个问题：
1. 用户反馈在【某平台】刷视频时即使点击“不感兴趣”，仍大量推荐游戏视频
2. 用户反馈长按视频时无法保存相册和分享视频。
- 例子2:
以下是一段用户的问题：
用户: 不新鲜
...
问题的要素：
「范围」【某平台】生活服务
「对象」商家
「举报原因」餐品不新鲜、有异味、口味不好
结合以上问题要素该用户反馈的问题：
用户反馈了一个问题：
1. 用户反馈在【某平台】生活服务买的餐品不新鲜、有异味

## 以下是一段用户和客服对话：
<Conversation Here>
问题的要素：
<Fine-grained Outlines Here>
## 结合以上问题要素该用户反馈的问题：
```

Figure 15: Prompts used in Summary Generation and Verification (Step3: *Complaint*).

Quiz Generation

```
## 角色:
你是一名负责生成阅读理解题目的专业教育专家，专注于设计高难度且精准的问题，尤其擅长制造可能引发其他模型幻觉的情景。

## 任务:
根据给定的“用户和客服对话”和对应的“摘要”，从对话的深层含义、细节、隐含信息、背景知识等多个方面入手，生成以下三类问题，并严格按照指定格式输出。问题需具备一定迷惑性，但答案必须唯一且准确。注意确保答案的唯一性，并通过精确措辞避免歧义。
1. **选择题**: 共 <NUM1> 道，每道题提供4个选项，有且仅有一个正确答案，其他选项需具备迷惑性且与对话内容紧密相关。
2. **填空题**: 共 <NUM2> 道，要求根据对话或摘要填空，设计隐含信息填空或需要推理的空格内容，避免表面化问题。如果填空题有多个空，每个空的正确答案用’；’分割。
3. **判断题**: 共 <NUM3> 道，每道题为"True"或"False"判断，问题可涉及推理或对话隐含的态度、立场等细节，增加误判可能性。

## 输出格式（JSON 格式）:
{
"选择题": [
{
"问题": "[问题内容]",
"选项": {"A": "[选项1]","B": "[选项2]","C": "[选项3]","D": "[选项4]"}, "正确答案": "[答案]"
},
...
],
"填空题": [
{"问题": "[问题内容，空格部分用“____”表示]","正确答案": "[答案]"},
...
],
"判断题": [{"问题": "[问题内容]","正确答案": "[True/False]"},
...
]
}

## 以下是一段用户和客服对话:
<Conversation Here>
## 以下是该对话的摘要:
<Summary Here>
请按照要求生成问题:
```

Figure 16: Prompts used in Quiz Generation and Verification.

LLM Summary Generation

角色:
你是一名专业的客服专家，擅长用精炼语言提取用户反馈的关键信息。

目标:
根据以下用户与客服的对话内容，使用简洁清晰的语言总结用户反馈的主要问题。
请确保输出内容没有冗余，仅包含用户反馈的主要问题，避免重复或不相关的信息，不需要对客服的回答进行总结。

## 输出要求:
1. 仅保留用户反馈的主要问题，忽略无关或不明确的内容，不要总结客服的回答。
2. 确保问题表述准确、清晰。

## 输出格式要求:
请按以下格式输出，无需额外分析:
1. <问题1>
2. <问题2>
...

例子:
- 输入对话内容:"客户: 我的喜欢列表被自动私密了没有入口
客户: 如何设置个人主页喜欢列表的可见范围
客户: “喜欢页”被限制查看怎么办
客户: 视频处罚/没流量怎么办
客户: 为什么我无法设置点赞列表查看权限
客户: 无法显示我的喜欢列表 直接消失了这个功能
坐席: 您好，小助手为您服务；请稍等下，小助手先看下咱们的问题
坐席: 请问咱们是想要公开主页的喜欢列表吗
客户: 是的
坐席: 您好，设置个人页喜欢列表公开的问题已经为您处理了，您可等待30分钟后重启【某平台】再进行查看
坐席: 已经为您加入了哦，辛苦您稍后重启试试哦
坐席: 没有其他问题的话,小助手就先关闭会话了 后续有问题还可以继续咨询小助手哦 同时也辛苦您稍后对小助手的服务进行评价,祝您生活愉快,再见"
- 输出摘要:"1. 用户咨询个人主页喜欢列表的相关问题及解决办法
2. 用户咨询视频被处罚及没流量的解决办法"

## 对话内容:
<Conversation Here>

输出:

Figure 17: Prompts used in LLM Summary Generation.

GPT-4 Discrimination

你是一个对话摘要的专家，我会给你一个分点的人工总结摘要和分点的模型生成摘要。人工总结摘要是绝对正确的，你需要将模型生成摘要和人工总结摘要进行比较，输出。

###请注意以下定义：
正确：该点与人工总结内容中的某一点含义完全一致。
错误：该点与人工总结内容中的某一条相关，但细节或表述不正确，或者表述不完整。
编造：该点与人工总结内容中的所有点都无关。

###请回答以下问题：
模型输出摘要的所有点中，正确的有几个?
模型输出摘要的所有点中，错误的有几个?
模型输出摘要的所有点中，编造的有几个?

###请根据以下格式输出，无需额外分析：
{
"correct": <正确的部分数量>,
"incorrect": <错误的部分数量>,
"fabricated": <编造的部分数量>,
}

###输入
人工总结摘要：<GT_SUMMARY Here>
模型输出摘要：<MODEL_SUMMARY Here>

###输出

Figure 18: Prompts used in GPT-4 Discrimination.

MCQ Answering

```
# 角色:
你是一名专业的客服专家，擅长作答各种客服领域内的题目。

## 目标:
根据以下用户与客服的对话内容，回答选择题。

## 输出要求:
1. 仅输出选项字母（A、B、C、D），无需任何解释。

## 对话内容:
<Conversation Here>

## 选择题:
<Question Here>

请回答
```

Figure 19: Prompts used in MCQ Answering.

FIB Answering

```
# 角色:
你是一名专业的客服专家，擅长作答各种客服领域内的题目。

## 目标:
根据以下用户与客服的对话内容，回答填空题。题目中只有"____"的部分需要你填空。如果有多个空，你的回答中每个空的答案用’；’分割。

## 输出要求:
1. 仅输出答案，不要输出任何额外字符。

## 对话内容:
<Conversation Here>

## 填空题:
<Question Here>

请回答
```

Figure 20: Prompts used in FIB Answering.

T/F Answering

```
# 角色:
你是一名专业的客服专家，擅长作答各种客服领域内的题目。

## 目标:
根据以下用户与客服的对话内容，用True或False回答判断题。

## 输出要求:
1. 仅输出True或False，无需任何解释。

## 对话内容:
<Conversation Here>

## 判断题:
<Question Here>

请回答
```

Figure 21: Prompts used in T/F Answering.

Step1: Conversation Classification

```
# 角色:
你是一名客服专家
## 目标:
- 准确对用户与客服的对话进行分类，判断属于「咨询类」、「反馈类」、「举报类」和「办事类」其中一种。如果用户反馈了多个问题，只关心用户多次提到且强烈关注的主要问题。

## 标签定义:
-「咨询类」：用户咨询信息，包括xxx怎么搞/啥时候/xxx是什么/能否xxx/功能入口等。
举例1：原账户信用卡注销了，订单退款能否退到其他账户。
举例2：想开通店铺授权号，但需要先开通企业号，如果用营业执照开通会不会和店铺冲突
举例3：自己发布的视频为被封禁的原因
-「反馈类」：用户投诉反馈或吐槽功能不好用/有bug、举报主播/现象、吐槽xxx功能/事件。
举例1：用户在下载【某平台】后，登录时同意个人信息后页面空白无反应的问题
举例2：【某平台】直播间遇到一位主播说脏话，认为该主播有赌博行为，希望平台能够处理。
举例3:用户反馈推荐的内容不喜欢，包含很多低质内容。
-「办事类」：用户办啥事、诉求。希望客服/平台退款/申诉解封账号或店铺/取消限流/关闭xx功能/寻求帮助等。
举例1：用户在xx平台上发布视频后，有人在评论区辱骂他，希望能够屏蔽此类人员。
举例2：用户希望客服帮忙解封账号。
举例3：用户希望客服帮忙进行订单退款，订单改约。
-「举报类」：用户对违规行为、内容或现象进行举报，期望平台介入并处理。举报内容通常针对违反平台规则的行为，包括诈骗、欺诈、恶意信息传播、色情或暴力内容、侵权行为等。
举例1：用户举报某直播间存在赌博行为，并提供相关证据截图，要求平台尽快处理。
举例2：用户发现某视频中存在欺诈信息（如假冒产品销售），希望平台下架该视频并对发布者进行处罚。
举例3：用户举报一位主播涉嫌传播不实信息，要求平台核查并封禁账号。
-「无效反馈」：对话信息量不足，或者用户未有明确问题。

## 输出格式:
- 输出「咨询类」、「反馈类」、「办事类」和「无效反馈」其中一种。

## 以下是一段用户和客服对话:
<Conversation Here>
## 该对话属于「咨询类」、「反馈类」、「办事类」、「举报类」和「无效反馈」其中哪一种:
```

Figure 22: Prompts used in CoT prompting (Step1).

Step2: Outline Extraction

角色:

你是一名客服专家

目标:

- 给你一段用户与客服的对话和对话的分类，请你针对用户咨询的问题进行要素提取。对话的分类属于「咨询类」、「反馈类」和「办事类」其中一种。要素可以分为：「原因」、「范围」、「咨询内容」、「场景」、「故障现象」、「对象」、「举报原因」、「吐槽内容」、「诉求」，从中选取你觉得相关的要素类别进行提取。

## 输出格式:

「要素类别」：

「要素类别」：

「要素类别」：

例子：

- 例子1:

– 以下是一段用户和客服对话：

用户: 唉，你好啊，我想问一下，我【某平台】账号的啊，私信功能被处罚了，不能用7天，我想问一下是为什么？

...

– 对话分类: 「咨询类」

– 输出:

「原因」：私信功能因涉及色情低俗被处罚，不能用7天

...

- 例子2:

– 以下是一段用户的问题：

用户: 为什么你们的【某平台】后台总是自己掉，每次都要重新登陆

...

– 对话分类: 「咨询类」

– 输出:

「故障现象」后台总是自动掉线，需要重新登陆

...

+

「范围」【某平台】

...

- 例子3:

以下是一段用户和客服对话：

用户: 因为我之前申请了那个账号的问题，要补充资料，我啊，昨天没来得及补充，然后那个电池已经过期了，你帮我再发个链接，我补充一下资料。

...

– 对话分类: 「办事类」

– 输出:

「原因」：用户未及时补充账号资料导致申请过期

...

##要求：

2. 如果原文反馈不同的问题，则分多点输出，不同问题之间用"+"符号分割，最多输出2组问题要素，如果无额外问题内容则只输出1组问题要素。

## 以下是一段用户和客服对话：

<Conversation Here>

对话分类：<Type Here>

输出:

Figure 23: Prompts used in CoT prompting (Step2).

Step3: Summary Generation

角色:
你是一名客服专家，擅长信息提炼与概括
目标:
- 给你一段用户与客服的对话，请你针对用户咨询的几个要素内容结合原文来简要总结用户反馈的问题，

## 输出格式:
- 包括客户在什么范围因什么原因咨询某事，如果要素点「无」则忽略过该要素
- 输出字数限制在35字以内，逗号控制在1个以内
- 如果用户没有明确的问题则输出：用户没有明确问题。如果有多组问题要素，则总结多个问题。如果用户没有明确说明原因，则不需要输出原因相关解释

例子
- 例子1:
– 以下是一段用户和客服对话：
...
用户: 留咨组件找不到 ...
– 对话分类:「咨询类」
– 问题的要素：
「原因」：线索经营打不开;「范围」：【某平台】来客;「咨询内容」：如何打开线索经营
结合以上问题要素该用户反馈的问题:用户咨询【某平台】来客的线索经营功能如何打开
- 例子2:
– 以下是一段用户和客服对话：
用户: 长按视频这块怎么没有保存相册了，不好分享不了视频别扭，没有保存视频了，望改进
...
– 对话分类:「反馈类」
– 问题的要素：
「范围」无;「场景」长按视频时;「故障现象」无法保存视频到相册，不能分享视频。
结合以上问题要素该用户反馈的问题:用户反馈在长按视频时无法保存视频到相册，不能分享视频。
- 例子3:
– 以下是一段用户和客服对话：
客服: 客服代表为您服务。
客服: 你好，小助手很高兴为您服务，请问有什么可以帮您?
用户: 唉，你好，我想问一下我这个我的【某平台】号被封禁了。
...
– 对话分类:「办事类」
– 问题的要素：
「范围」：账号;「原因」：用户账号可能因违规发布医疗相关内容被封禁;「诉求」：希望解封账号，且申诉失败，寻求其他解决办法
+
「范围」：注册;「原因」：用户账号被封禁，考虑能否用原手机号再注册;「诉求」：了解原手机号能否再注册【某平台】号
结合以上问题要素该用户反馈的问题:
用户反馈了两个问题：
1. 用户【某平台】账号因违规被封申诉失败，希望解封
2. 用户咨询账号被封禁的原手机号能否再注册。

## 以下是一段用户和客服对话：
<Conversation Here>
对话分类：<Type Here>
问题的要素：<Fine-grained Outlines Here>
## 结合以上问题要素该用户反馈的问题:

Figure 24: Prompts used in CoT prompting (Step3).

Hallucination Detection

```
# 角色:
你是一名对话评估领域的专家，擅长检测摘要中的幻觉现象。你的主要职责是通过确保摘要中的每一条信息都可以明确地由对话内容支持或合理推断得出，来评估摘要的准确性和一致性。此外，你还擅长识别摘要中存在的矛盾或不支持的细节。

# 目标:
根据以下用户与客服之间的中文对话以及提供的摘要，判断该摘要是否包含幻觉内容（即：摘要中包含未提及、虚构、不准确或与对话内容无关的信息）。
幻觉的定义：如果满足以下任意条件，信息被视为幻觉:
(1) 摘要包含的信息无法明确地由对话支持，或无法合理推断得出。
(2) 摘要中包括的细节与对话中的事实或内容相矛盾。

# 示例1:
## 对话内容:
客户: 【某平台】给钱转钱
客户: 收入提现到零钱
...
## 摘要:
用户反馈了两个问题:
1. 用户咨询【某平台】给自己转钱的原因及如何取消收益转账
2. 用户咨询卸载【某平台】app能否解决问题，若不行，注销账号是否可行
## 输出:Yes

# 示例2:
## 对话内容:
客户: 【某平台】给钱转钱
客户: 收入提现到零钱
客户: 人工客服
...
## 摘要:
用户反馈了一个问题:
1. 用户咨询【某平台】给自己转钱的原因及如何取消收益转账
## 输出:No
# 输出要求:
仅输出“Yes”或“No”，不包含任何其他无关内容。你输出的唯一结果必须是“Yes”或“No”。

# 对话内容:
<Conversation Here>

# 摘要:
<Summary Here>

# 输出:
```

Figure 25: Prompts used in Hallucination Detection.