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. Two tasks, *Conversation Summarization* and *Quiz Examination*, are designed to comprehensively assess faithfulness hallucinations in LLMs. Extensive experiments on 22 LLMs reveal that faithfulness hallucinations persist across all LLMs.

1 Introduction

Large Language Models (LLMs), such as GPT-4o (Achiam et al., 2023), and LLaMA (Touvron et al., 2023), excel in tasks such as question answering and dialogue. However, they are also prone to hallucinations—factually incorrect or nonsensical content (Wang et al., 2023), which can mislead users, erode trust, and hinder real-world deployment. LLMs generally experience two types of hallucinations: *factuality hallucination*, which focuses on inconsistencies between generated content and world knowledge, and *faithfulness hallucination*, which highlights divergence from the provided context (Huang et al., 2023). Existing benchmarks (Lin et al., 2021; Pal et al., 2023; Cheng et al., 2023) mainly focus on factuality hallucinations, with only a few (Tang et al., 2024; Ming et al., 2024) addressing faithfulness hallucinations.

However, real-world conversations often include **casual language** (e.g., interjections, emojis, and abbreviations) and **topic shifts**. In this work, we are the first to utilize real-world **customer service conversations** to evaluate faithfulness hallucinations in LLMs. Examples of user interactions with human customer service are shown in Fig. 1, drawn from an online platform supporting over 20 products across various domains. These conversations often feature casual language (e.g., interjections like "uh," emojis like 😊, and abbreviations like "BTW"). Additionally, topics can shift within a conversation, as seen in the fourth sub-figure "Complaint," where the topic shifts from "request user information" to "claim to sue." These factors present challenges for LLMs in accurately interpreting the context.

In this light, we introduce CONVFAITHEVAL, the first faithfulness hallucination evaluation benchmark based on practical conversations. It includes two tasks: *Conversation Summarization* and *Quiz Examination*. In construction, we first select and filter conversations from a Chinese online customer service platform, and anonymize each data sample. Then, we use GPT-4o to automatically identify the conversation type, number of topics, and generate a summary and a quiz. Lastly, we perform strict human verifications to ensure data quality.

We evaluate 22 LLMs with our CONVFAITHEVAL, where almost all LLMs suffer from faithfulness hallucinations. Our contributions: (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, containing $2,500$ conversations for evaluation. (3) With CONVFAITHEVAL, we design two tasks to comprehensively evaluate faithfulness hallucinations in 22 LLMs and provide valuable insights.

2 CONVFAITHEVAL Benchmark

2.1 Conversation Collection

In Fig. 2 (a), we first collect and filter raw conversations from a customer service platform, followed by a comprehensive anonymization process.

Raw conversation selection and filtering. The online platform, which supports over 20 products

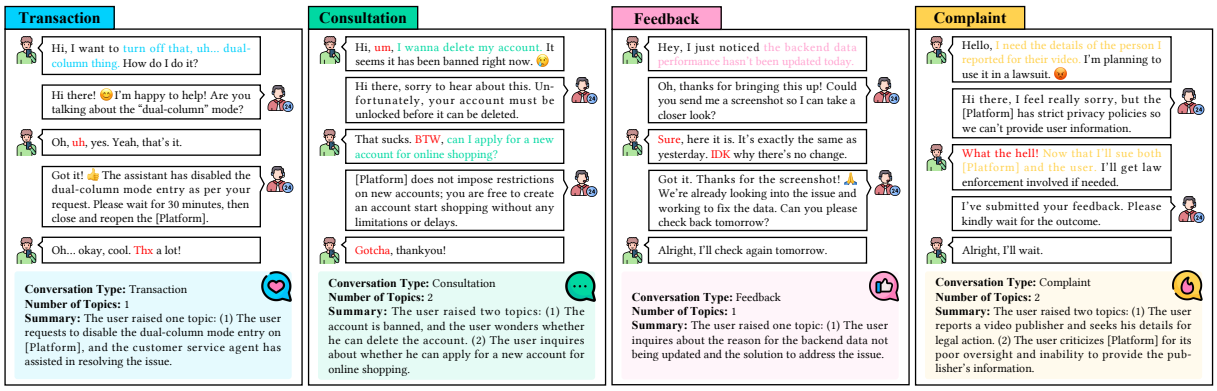


Figure 1: An overview of CONVFAITHEVAL benchmark (English-translated).

across various domains (e-commerce, advertising, finance, etc.), serves as the source for raw data. To construct CONVFAITHEVAL, we enlist employees familiar with the platform and its products as human annotators. The annotators follow four guidelines to select and filter historical conversations from the platform: (1) Random selection: Conversations are randomly selected to reflect real-world distribution. (2) Contextual integrity: Short or incomplete conversations are excluded for lacking essential context. (3) Sensitive conversation filtering: Conversations containing political, pornographic, or violent content are removed to maintain ethical standards. (4) Noisy conversation filtering: Conversations with excessive emojis, non-standard grammar, or emotional expressions are filtered to preserve clarity. 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 comply with data protection regulations, we implement a comprehensive anonymization process on the collected conversations, including: (1) personal information redaction, (2) entity replacement, (3) context obfuscation, and (4) metadata removal. Then, annotators validate and revise the anonymized content for accuracy.

2.2 Summary Generation and Verification

In Fig. 2 (b), we generate a summary for each conversation with GPT-4o, which will then be verified to ensure its correctness (details in Appx. D.1).

Summary generation. We use GPT-4o to generate an initial summary for each conversation, capturing key details with main concerns and queries clearly outlined. The Tree of Thoughts (ToT) framework (Yao et al., 2024; Long, 2023) is employed to guide GPT-4o in completing this summary task. (1) We provide GPT-4o with definitions of four conversation types—*Transaction*, *Consultation*, *Feedback*, and *Complaint*—as context and instruct it to categorize the conversation. (2) Based on the detected type, we guide GPT-4o to generate an outline for the conversation, with a tailored prompt template for each type to extract relevant information from the conversation. For example, for a *Consultation*, we expect to extract the "cause", "scope", and "content". (3) We then instruct GPT-4o to generate a concise summary following the outline. The summary includes a brief description of each topic, with each topic and its description listed in a clear, point-by-point format for readability.

Summary verification. To ensure the correctness of generated conversation summaries, we employ human annotators for careful reviews and corrections, following these guidelines: (1) Summary correction: Human annotators refine summaries for accuracy and completeness, correcting errors and enhancing clarity, especially for multi-topic summaries. (2) Conversation categorization: Conversations are classified into four types: *Transaction* (service requests like returns or refunds), *Consultation* (advice-seeking on issues or features), *Feedback* (bug reports, suggestions, or usability comments), and *Complaint* (dissatisfaction with a person, service, or platform). (3) Sensitive information removal: Annotators anonymize or remove all personal and platform-related information.

2.3 Quiz Generation and Verification

In Fig. 2 (c), we create an additional quiz for each conversation for evaluation (details in Appx. D.2).

Quiz generation. We use GPT-4o to generate a diagnostic quiz from a conversation and its human-verified summary, including multiple-choice questions (MCQ), fill-in-the-blank (FIB), and true-or-false (T/F), with two questions of each type per conversation. To emphasize contextual reasoning over memory recall, we enhance reasoning complexity through carefully designed prompts.

Quiz verification. We also ensure quiz correctness through human verification in three steps: (1) Content review: Annotators verify each question's cor-

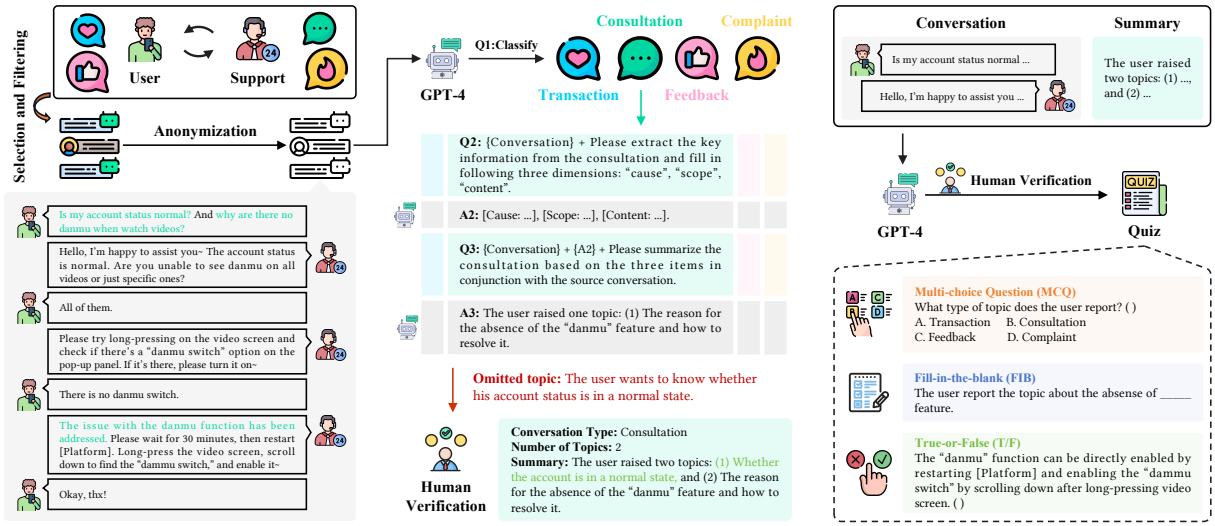
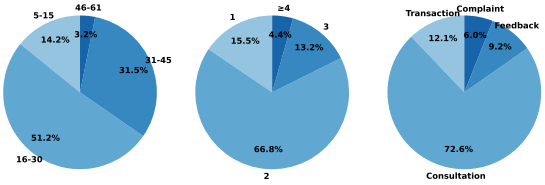


Figure 2: An illustration of the construction pipeline for our CONVFAITHEVAL benchmark.



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

Figure 3: The statistics on CONVFAITHEVAL.

rectness and relevance. (2) Difficulty calibration: Questions are assessed to match the conversation's complexity. (3) Redundancy elimination: Repetitive questions are removed to ensure diversity.

2.4 Statistics and Analysis

The constructed benchmark contains a total of $2,500$ diverse conversations, each annotated with "conversation type", "number of topics", and "summary". The statistics is presented in Fig. 3. It is observed that 85.8% of conversations involve over 15 interaction turns, and 84.5% include multiple topics. Meanwhile, conversation types within are primarily dominated by *Consultation* (72.6%), with *Complaint* (6.0%) being the least, aligning with real-world distributions. For quiz collection, quizzes are successfully generated for $2,500$ conversations, resulting in a total of $15,000$ questions.

3 Experiments

3.1 Experimental Setup

We evaluate various LLMs that span different versions and scales. In *Conversation Summarization*, we introduce five metrics: Omission Rate (**O**), Error Rate (**E**), Fabrication Rate (**F**), Recall (**R**), Precision (**P**), and F1 Score (**F1**). In *Quiz Examination*, we report the accuracy for each question type and the average accuracy as quiz score (**QS**). More details are presented in Appx. B.

3.2 Main Results

We report the main evaluation results on 22 LLMs in Table 1, and make the following observations. **(1) Closed-source GPT-4o achieves the best performance, with Qwen2.5-72B showing competitive results.** Closed-source GPT-4o achieves the best **F1** of 88.0% and **QS** of 90.1% on both tasks, demonstrating its superior performance in understanding real-world user queries. Among open-source LLMs, Qwen2.5-72B stands out as a strong competitor, reaching 87.6% in **F1** and 90.1% in **QS**, closely matching GPT-4o's performance. **(2) Chinese LLMs outperform non-Chinese counterparts within the open-source category.** Our benchmark focuses on hallucinations in Chinese, and 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** (87.6% vs. 81.8%) and **QS** (90.1% vs. 88.8%). Similarly, InternLM2.5-7B, also a Chinese LLM, exhibits stronger performance when compared to LLaMA-3-8B on both tasks. **(3) LLMs with more advanced versions and larger scales consistently outperform their inferior counterparts.** Within the InternLM family, InternLM-2.5-20B achieves higher scores in $\boldsymbol{P}$, $\boldsymbol{R}$, and **F1** than the smaller InternLM-2.5-7B, with improvements in **F1** (76.5% vs. 75.0%) and $\boldsymbol{R}$ (96.4% vs. 95.3%). Similarly, for the Qwen family, Qwen2.5-72B demonstrates superior performance across multiple metrics compared to Qwen2-7B and Qwen2-72B in **F1**. This pattern is consistent with open-source LLMs. Exemplified by GPT-4o, with a parameter size considerably larger than GPT-4o mini, achieves superior **F1** (88.0%) and **QS** (90.1%).

LLMs		Conversation Summarizaton						Quiz Examination			
		O ↓	E ↓	F ↓	R ↑	P ↑	F1 ↑	MCQ ↑	FIB ↑	T/F ↑	QS ↑
InternLM	InternLM-7B	99.2	**0.1**	81.1	0.8	0.6	0.7	81.2	50.3	71.1	67.5
	InternLM-20B	21.3	23.7	26.3	78.7	50.0	61.2	84.9	52.0	80.0	72.3
	InternLM2.5-7B	4.7	21.1	17.2	95.3	61.8	75.0	94.9	71.5	92.0	86.1
	InternLM2.5-20B	**3.7**	18.7	17.9	**96.4**	63.4	76.5	95.4	73.3	93.0	87.2
Qwen	Qwen2-7B	4.5	19.2	15.4	95.5	65.5	77.7	94.7	69.9	89.8	84.8
	Qwen2-72B	6.0	11.9	8.5	94.0	79.7	86.2	97.1	75.7	96.1	89.6
	Qwen2.5-7B	12.4	15.0	**6.7**	87.7	78.4	82.7	95.2	72.1	92.0	86.4
	Qwen2.5-72B	5.3	10.9	7.7	94.7	**81.4**	**87.6**	**97.2**	**76.2**	**96.7**	**90.1**
LLaMA	LLaMA-2-7B	58.7	22.3	57.3	41.3	17.6	24.7	24.8	26.2	48.0	33.0
	LLaMA-2-13B	27.8	32.7	36.2	72.2	31.2	43.6	26.3	40.7	48.0	38.3
	LLaMA-3-8B	8.3	25.4	21.9	91.7	52.8	67.0	91.2	64.6	87.2	81.0
	LLaMA-3-70B	13.0	16.9	9.6	87.1	73.5	79.7	96.2	72.7	95.0	88.0
	LLaMA-3.1-8B	10.5	21.8	14.0	89.5	64.3	74.8	89.1	62.9	87.9	80.0
	LLaMA-3.1-70B	7.9	17.2	9.3	92.1	73.5	81.8	96.6	73.3	96.5	88.8
GLM	ChatGLM3-6B	34.2	28.8	21.9	65.8	49.3	56.3	24.8	64.4	54.5	47.9
	ChatGLM4-9B	7.8	18.9	15.9	92.3	65.2	76.4	90.7	63.0	86.6	80.1
GPT	GPT-3.5-Turbo	16.4	16.5	10.6	83.6	72.9	77.9	90.6	67.9	88.2	82.2
	GPT-4	15.6	12.1	17.9	84.4	70.0	76.5	96.9	74.8	**96.6**	89.4
	GPT-4o mini	4.8	12.2	**6.9**	95.2	80.9	87.5	95.9	72.2	95.4	87.8
	GPT-4o	**3.6**	**11.0**	8.0	**96.5**	**81.0**	**88.0**	**97.3**	**76.9**	96.0	**90.1**
Gemini	Gemini 1.5 Flash	15.5	13.7	13.9	84.5	72.4	78.0	88.7	75.4	95.8	86.6
	Gemini 1.5 Pro	14.6	12.2	14.2	85.4	73.7	79.1	88.8	75.4	96.0	86.7

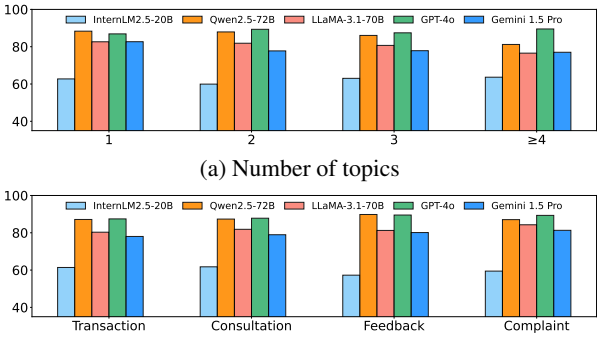
Table 1: **Main results.** We evaluate 22 LLMs across six families with different versions and scales, on Conversation Summarization and Quiz Examination tasks. All results are shown in percentages (%), and the best and second-best results are marked in **bold** and underline for open- and closed-LLMs, respectively.

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".

Performance w.r.t. number of topics. Fig. 4 (a) shows the performance comparison in **F1** of various LLMs on the conversation summarization task across conversations involving different numbers of topics. The analysis reveals two key observations. (1) As the number of topics increases, there is a consistent decline in performance for most LLMs, including Qwen2.5-72B, LLaMA-3.1-70B, and Gemini 1.5 Pro. This highlights the inherent challenge of managing hallucinations in more complex conversational scenarios, where topic shifts are more frequent. (2) In contrast, GPT-4o consistently demonstrates robust performance, effectively handling complex conversations. Its superior ability to maintain coherence across multiple topics suggests an advanced understanding of real-world conversations. This suggests that closed-source LLMs benefit from superior training and alignment, leading to more faithful summarization, while open-source LLMs require further optimization to handle multi-topic conversations.

Performance w.r.t. conversation type. Fig. 4 (b) compares the model performance across four conversation types. The tested LLMs exhibit a similar trend in performance across the four conversation types, indicating their relatively uniform handling capability of different conversation types.



(a) Number of topics

(b) Conversation type

Figure 4: **In-depth analysis** on the impact of (a) the number of topics and (b) conversation type upon **F1** on *Conversation Summarization*.

4 Conclusion

In this paper, we have introduced 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. Evaluations of 22 LLMs show that LLMs struggle with increasing topic complexity in real-world conversations while closed-source models achieve better performance and robustness, offering further insights for hallucination evaluation and plausible mitigation direction in LLM faithfulness.

5 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. 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.

6 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.

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.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. 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.

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.

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.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

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

A Related Works

Hallucination benchmarks. Existing 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. TofuEval (Tang et al., 2024) samples documents from two dialogue datasets, i.e., MediaSum (Zhu et al., 2021) and MeetingBank (Hu et al., 2023), and evaluate hallucinations with topic-based dialogue summarization task. FaithEval (Ming et al., 2024) specifically benchmarks the faithfulness of LLMs in contextual scenarios with question answering. 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).

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.

B Experiment Setup

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

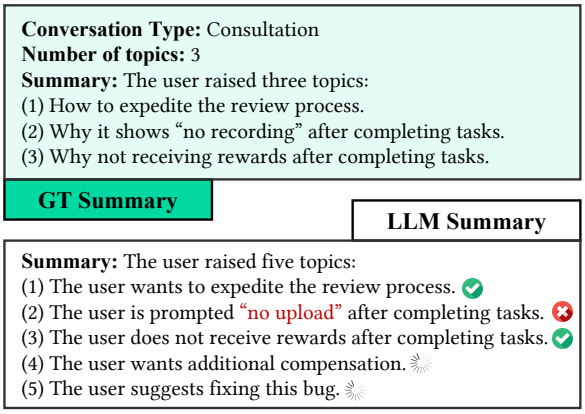


Figure 5: An illustration of correct, erroneous, and fabricated topics in conversation summarization. The "erroneous" indicates the topic is involved but contains incorrect details; the "fabricated" means the topic is not mentioned in the conversation.

(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 $\boldsymbol{m}$ topics, the LLM generates summaries containing $\boldsymbol{n}$ topics, which are categorized into $\boldsymbol{a}$ correct, $\boldsymbol{b}$ erroneous, and $\boldsymbol{c}$ fabricated topics, where $\boldsymbol{n} = \boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}$, as shown in Fig. 5. The formula of the five metrics are as follows: Omission Rate ($\boldsymbol{O} = \boldsymbol{1} - \frac{\boldsymbol{a}}{\boldsymbol{m}}$), Error Rate ($\boldsymbol{E} = \frac{\boldsymbol{b}}{\boldsymbol{n}}$), Fabrication Rate ($\boldsymbol{F} = \frac{\boldsymbol{c}}{\boldsymbol{n}}$), Recall ($\boldsymbol{R} = \frac{\boldsymbol{a}}{\boldsymbol{m}}$), Precision ($\boldsymbol{P} = \frac{\boldsymbol{a}}{\boldsymbol{n}}$), 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 (**QS**).

Baselines. We evaluate on our ConvFaithEval various LLMs that span different versions and scales, including open-source LLMs[1]: InternLM

[1]InternLM-7B, InternLM-20B, InternLM2.5-7B and InternLM2.5-20B; Qwen2-7B, Qwen2-72B, Qwen2.5-7B, and Qwen2.5-72B; LLaMA-2-7B, LLaMA-2-13B, LLaMA-3-8B, LLaMA-3-70B, LLaMA-3.1-8B, and LLaMA-3.1-70B; ChatGLM3-6B and ChatGLM4-9B.

	LLM's Summary	Analysis
Over-Interpretation (58%)	User asks about the limit on sending private messages when not mutually following.	The user did not ask for this information; it was proactively mentioned by the customer service in their response.
Misinterpretation (18%)	User's account was banned for tagging a friend in their profile introduction and asks about the reason and solution.	The original text does not state that tagging a friend caused the ban; the account was banned due to abnormal risks.
Key Information Omission (12%)	User's account was banned due to being hacked and performing unauthorized actions, affecting normal functionality. The user seeks unblocking and restoration of real-name information.	The user did not explicitly state that hacking "caused the ban," but rather mentioned that unauthorized actions due to hacking led to the occupation of their real-name information, thereby affecting normal usage.
Non-Question Responses (6%)	User is informed that the appeal result requires a 72-hour wait, and the refund will be completed within 1-3 days after approval.	This is not a question or request.
Fabrication (6%)	User asks about the return time and conditions for real-name verification, particularly its relation to account cancellation (end of October).	The user did not ask about time or conditions.

Table 2: Failure case study on Qwen2.5-72B.

Temperature	0.0	0.2	0.4	0.6	0.8	1.0
F1 Score	76.52	77.72	80.00	78.29	76.12	72.51

Table 3: The impact of temperature parameters on *Conversation Summarizaton* with Qwen2.5-7B, where we only evaluate the first 100 samples in the evaluation set.

(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) 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.

C More Experiments and Analysis

C.1 Failure Case Study

We provide statistics of 50 error cases produced by Qwen2.5-72B and categorize them into five groups in Table 2, each type with one example. The faithfulness hallucination observed in Qwen2.5-72B, including over-interpretation, misinterpretation, key information omission, non-question responses, and fabrication, can be attributed to various factors. Over-interpretation likely arises from its tendency to anticipate additional user needs based on limited context. Misinterpretation is often caused by the reliance on surface-level patterns rather than deeper understanding. Key information omission indicates a failure to retain or emphasize critical details, while non-question responses suggest the misclassification of input types. Fabrication occurs when the LLM generates information that is not supported by the user's query. Mitigating these issues requires a combination of improved fine-tuning with more diverse and contextually rich data, implementing stricter constraints on response generation, and enhancing the ability to correctly identify and emphasize relevant details. Additionally, refining the query classification system and post-processing steps can help reduce the occurrence of hallucinations.

[2]GPT-3.5-Turbo, GPT-4, GPT-4o mini, and GPT-4o; Gemini 1.5 Flash, Gemini 1.5 Pro.

C.2 Temperature Parameter Ablation

In Sec. B, we set the temperature parameter to zero primarily following the experimental setup of prior hallucination evaluation benchmarks (Lin et al., 2021; Wei et al., 2024; Oh et al., 2024), ensuring determinism and reproducibility for each evaluation. In Table 3, we show the results of varying the temperature parameter from 0.0 to 1.0 during the evaluation of Qwen2.5-7B. The F1 score fluctuates across different temperature settings, with the highest F1 score of 80.00 achieved at a temperature of 0.4. As the temperature increases from 0.0 to 0.4, we observe a steady improvement in performance, suggesting that a moderate increase in temperature enhances the model's ability to balance between precision and recall. However, beyond this point, the F1 score starts to decline. At higher temperatures of 0.6, 0.8, and 1.0, the performance deteriorates, with the F1 score dropping to 72.51 at a temperature of 1.0. This trend indicates that while some level of randomness (introduced by the temperature parameter) can benefit model performance, excessively high temperatures may lead to over-variance and a drop in reliability. Thus, a temperature of 0.4 appears to offer an optimal trade-off for Qwen2.5-7B's performance in this task.

C.3 Examples in Quiz Examination

In Sec. 2.3, we select three formats (MCQ, FIB, T/F) in the Quiz Examination for an easy and reliable evaluation. We emphasize that the format of the test questions (e.g., MCQ, FIB, T/F) is not

Aspect	Conversation Context	Question	Answer	Explanation
Long-Chain Causality	**Customer**: "I received a damaged laptop. I called your support yesterday, and they told me to email pictures of the damage. I sent the email, but I haven't received a response yet. I need the replacement before my trip next Monday!" **Service**: "Thank you for the information. Let me check if we've received your email and process the replacement for you."	Why does the customer mention their trip next Monday? • A. They want to return the damaged laptop during their trip. • B. They are concerned about receiving the replacement in time. • C. They are requesting a refund instead of a replacement. • D. They plan to email the support team again during the trip.	B	The model needs to connect multiple conversational turns (the damaged laptop, emailing photos, and the trip deadline) to infer the customer's underlying urgency for the replacement.
Implicit Attitude Judgments	**Customer**: "I've been a loyal customer for five years, and I've never had such an issue before. This experience has been incredibly frustrating." **Service**: "We're really sorry for the inconvenience. Let me escalate this matter to ensure it gets resolved quickly."	The customer is satisfied with the service provided so far. True or False?	False	Although the customer explicitly mentions their loyalty, their frustration and dissatisfaction are implied through the tone and wording, requiring the model to interpret sentiment beyond surface-level keywords.

Table 4: Examples of quiz examination, where aspects like long-chain causality and implicit attitude judgments are both considered and challenged in our evaluation.

directly tied to the difficulty or the reasoning complexity. In fact, we have deliberately increased the reasoning complexity of the test questions during construction, ensuring that the questions effectively challenge the contextual reasoning abilities of LLMs, including long-chain causality and implicit attitude judgments, as shown in Table 4.

D Implementation Details

D.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. 6): We classify the conversations into one of four predefined types: *Transaction*, *Consultation*, *Feedback*, and *Complaint*. (2) Outline extraction (Fig. 7, 9, 11, and 13): Using GPT-4o, we prompt the model to generate corresponding fine-grained outlines based on the classified conversation type. (3) Summary generation (Fig. 8, 10, 12, and 14): Finally, GPT-4o is prompted to write a summary by synthesizing the conversation content and fine-grained outlines.

D.2 In Quiz Generation and Verification

In Sec. 2.3, We utilize GPT-4o 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. 15.

D.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-4o discrimination: We prompt GPT-4o to compare the LLM-generated summary with the ground truth summary for evaluation purposes. Detailed prompts are shown in Fig. 16 and Fig. 17.

D.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. 18, 19, and 20.

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

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 7: 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 8: 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 9: 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: 留咨组件找不到
...
「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 10: 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 11: 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 12: 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 13: 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 14: Prompts used in Summary Generation and Verification (Step3: *Complaint*).

Quiz Generation

- Role: 你是一名负责生成阅读理解题目的专业教育专家

- Background: 需要根据一段“用户和客服对话”以及对应的“摘要”，从多个维度生成高难度的阅读理解题目，包括选择题、填空题和判断题。这些题目需要具备迷惑性，但答案必须唯一且准确，以考察阅读者的深层理解和推理能力。

- Skills: 你具备深厚的文本分析能力，能够准确把握对话的深层含义，尤其擅长制造可能引发模型幻觉的情景。设计出既具有挑战性又确保答案唯一性的题目。

- Goals: 生成符合要求的阅读理解题目，包括选择题、填空题和判断题这三类问题，确保题目具有一定迷惑性但答案唯一准确。

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。

- Workflow:

1. 仔细阅读“用户和客服对话”以及对应的“摘要”。

2. 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，设计选择题、填空题和判断题。

3. 确保题目具备迷惑性，但答案唯一准确，避免歧义。

- Constrains: 题目必须严格遵循指定格式输出，确保答案的唯一性，并通过精确措辞避免歧义。

- OutputFormat（JSON 格式）:

{

"选择题": [

{

"问题": "$问题内容",

"选项": {

"A": "$选项内容", ...

},

"正确答案": "A/B/C/D"

},

...

],

"填空题": [

{

"问题": "$问题内容(需要填空的部分用“____”表示)",

"正确答案": "$填空答案"

},

...

],

"判断题": [

{

"问题": "$问题内容",

"正确答案": "True/False"

},

...

]

}

- Input:

以下是一段用户和客服对话:

”’<CONTEXT>”’

以下是该对话的摘要:

”’<SUMMARY>”’

- Output:

请按照要求生成问题:

Figure 15: Prompts used in Quiz Generation and Verification.

LLM Summary Generation

- Role: 你是一名客服领域专家

- Skills: 你拥有出色的沟通技巧和信息提取能力，擅长信息提炼与概括用户问题。

- Goals: 给你一段在主要是抖音客服场景下用户与客服的对话或用户反馈内容，首先仔细阅读并理解提供的对话内容，使用简洁清晰的语言总结用户反馈的主要问题。

- Constrains: 聚焦总结用户反馈问题，忽略客服动作；避免冗余和不相关的信息。

- Workflow:

1. 阅读用户与客服的对话内容。

2. 识别并提取用户反馈中的关键信息。

3. 以简洁、清晰的语言总结用户的主要问题。

4. 如果原文反馈多个不同的问题要素，则分点输出

- OutputFormat: 简洁、清晰的语言描述用户的主要问题。

- Examples:

- 用户与客服的对话:

"客户: 我的喜欢列表被自动私密了没有入口

...

- 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"

- Input:

- 用户与客服的对话:"<CONTEXT> "

- Output:

- 输出摘要:

Figure 16: Prompts used in LLM Summary Generation.

GPT-4o Discrimination

- Role: 对话内容分析专家

- Background: 用户需要对一段用户与客服的对话内容进行分析，并将模型生成的摘要与人工总结摘要进行比较，以评估模型摘要的准确性和完整性。

- Skills: 你拥有对话内容解析、信息提取、对比分析和批判性思维的能力，能够从对话中提取关键信息，并与人工总结和模型生成的摘要进行精确对比。

- Goals: 通过对比分析，识别模型生成摘要中的“正确”、“错误”和“编造”问题点，为用户提供准确的评估结果。
模型输出摘要的所有点中，回答正确的有几个？错误的有几个？编造的有几个?

- Constrains: 必须确保分析的客观性和准确性，避免个人偏见影响结果。同时，确保分析结果清晰、条理分明，易于用户理解。

- Workflow:

1. 仔细阅读并理解用户提供的用户与客服的对话内容。
2. 将模型生成的摘要与人工总结摘要进行逐点对比。
3. 对比每个点是否与对话内容和人工总结摘要完全一致、部分一致或完全不一致。
4. 根据对比结果，将每个点分类为“正确”、“错误”或“编造”。
5. 以列表形式输出最终的分析结果。

- OutputFormat:

{
"correct": $正确的数量,
"incorrect": $错误的数量,
"fabricated": $编造的数量,
}

- Input:

用户与客服的对话：”’<CONTEXT>”’
人工总结摘要：”’<HUMAN_SUMMARY>”’
模型输出摘要：”’<MODEL_SUMMARY>”’

- Output:

Figure 17: Prompts used in GPT-4o Discrimination.

MCQ Answering

- Role: 你是一名客服领域专家
- Background: 用户需要根据对话内容回答选择题，对话内容涉及客服领域内的题目。
- Skills: 你具备出色的阅读理解能力、快速分析对话内容的能力以及准确回答问题的能力。
- Goals: 根据用户与客服的对话内容，准确回答选择题。
- Constrains: 仅输出选项字母，无需任何解释。
- OutputFormat: 选择题的答案选项。
- Workflow:
1. 阅读并理解用户与客服的对话内容。
2. 分析对话内容，确定问题的关键点。
3. 根据对话内容和问题的关键点，选择正确的答案选项。
- Input:
*对话内容:"'<CONTEXT> "'
*选择题:"'<QUESTION>"'
- Output:

Figure 18: Prompts used in MCQ Answering.

FIB Answering

- Role: 你是一名客服领域专家
- Background: 用户需要根据对话内容回答填空题，对话内容涉及客服领域内的题目。
- Skills: 你具备出色的阅读理解能力、快速分析对话内容的能力以及准确回答问题的能力。
- Goals: 根据用户与客服的对话内容，准确回答填空题。
- Constrains: 仅输出答案，不要输出任何额外解释。
- OutputFormat: 填空题的答案。
- Workflow:
1. 阅读并理解用户与客服的对话内容。
2. 分析对话内容，确定问题的关键点。
3. 根据对话内容和问题的关键点，填写正确的答案。
- Input:
*对话内容:"'<CONTEXT> "'
*填空题:"'<QUESTION>"'
- Output:

Figure 19: Prompts used in FIB Answering.

T/F Answering

- Role: 你是一名客服领域专家
- Background: 用户需要根据对话内容回答判断题，对话内容涉及客服领域内的题目。
- Skills: 你具备出色的阅读理解能力、快速分析对话内容的能力以及准确回答问题的能力。
- Goals: 根据用户与客服的对话内容，准确回答判断题。
- Constrains: 仅输出True或False，无需任何解释。
- OutputFormat: 判断题的答案。
- Workflow:
1. 阅读并理解用户与客服的对话内容。
2. 分析对话内容，确定问题的关键点。
3. 根据对话内容和问题的关键点，判断题目的是否。
- Input:
*对话内容:"'<CONTEXT> "'
*判断题:"'<QUESTION>"'
- Output:

Figure 20: Prompts used in T/F Answering.