QUARTZ: QA-based Unsupervised Abstractive Refinement for Task-oriented Dialogue Summarization

Mohamed Imed Eddine Ghebriout[1], **Gaël Guibon**[1,2], **Ivan Lerner**[3,4,5], **Emmanuel Vincent**[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Université Sorbonne Paris Nord, CNRS, Laboratoire d'Informatique de Paris Nord, LIPN, F-93430 Villetaneuse, France
[3]Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, F-75006 Paris, France
[4]HeKA, Inria Paris, F-75012 Paris, France
[5]Department of Medical Informatics, Assistance Publique Hôpitaux de Paris, Georges Pompidou European Hospital
{imed-eddine.ghebriout, gael.guibon}@loria.fr {ivan.lerner, emmanuel.vincent}@inria.fr

Abstract

Dialogue summarization aims to distill the core meaning of a conversation into a concise text. This is crucial for reducing the complexity and noise inherent in dialogue-heavy applications. While recent approaches typically train language models to mimic human-written summaries, such supervision is costly and often results in outputs that lack task-specific focus limiting their effectiveness in downstream applications, such as medical tasks. In this paper, we propose *QUARTZ*, a framework for task-oriented utility-based dialogue summarization. *QUARTZ* starts by generating multiple summaries and task-oriented question-answer pairs from a dialogue in a zero-shot manner using a pool of large language models (LLMs). The quality of the generated summaries is evaluated by having LLMs answer task-related questions before *(i)* selecting the best candidate answers and *(ii)* identifying the most informative summary based on these answers. Finally, we fine-tune the best LLM on the selected summaries. When validated on multiple datasets, *QUARTZ* demonstrates its effectiveness by achieving competitive results in various zero-shot settings, rivaling fully-supervised State-of-the-Art (SotA) methods. *Code and supplementary material are publicly available*[1].

1 Introduction

Automatic Text Summarization (ATS) is the task of generating a concise summary from a lengthy text. It can be *(extractive)* – selecting and concatenating key sentences – or *(abstractive)* – generating new condensed formulation that resemble human-written summaries (Lin and Ng, 2019). While early ATS methods relied on graph-based (Mihalcea and Tarau, 2004) and frequency-based techniques (Alsaedi et al., 2016), the emergence of large language models (LLMs) has transformed the field, achieving breakthroughs across domains, including healthcare (Van Veen et al., 2024). Dialogue summarization, a subfield of ATS, targets key information distillation from conversations. It has gained prominence due to its applicability in real-world settings such as customer service (Feigenblat et al., 2021), business meetings (Rennard et al., 2023), and healthcare (Abacha et al., 2023), where the input is often less structured than typical written texts like news articles. Summarizing such dialogues helps simplify complex interactions and supports downstream tasks like decision-making and automation. Dialogues introduces challenges absent in standard ATS: disfluencies, speaker turns, redundancy, verbosity, and loosely structured content. These characteristics make traditional ATS methods, designed for more structured text, less effective, often resulting in low-quality summaries (Zechner, 2002; Feng et al., 2022). Recently, the capability of LLMs to perform dialogue summarization has been assessed by framing it as a sequence-to-sequence task (Van Veen et al., 2024; Tian et al., 2024). Fine-tuning LLMs for task-oriented summarization has demonstrated promising results, even surpassing human performance (Van Veen et al., 2024). However, this approach has limited practicality due to the reliance on costly human-written summaries. Moreover, while the vast amount of text seen during training enables LLMs to be powerful *"multitask learners"*, their lack of robust mechanisms for maintaining topic focus and factual consistency makes them prone to topic-deviations, often resulting in incoherent summaries (Tonmoy et al., 2024).

In this paper we present *QUARTZ*, a framework for unsupervised, task-oriented abstractive dialogue summarization. Its task-oriented design enables better control over LLM outputs by maintaining topic focus and factual consistency, while its unsupervised nature eliminates the need for reference summaries or domain-specific knowledge beyond

[1]https://github.com/Mohamed-Imed-Eddine/QUARTZ

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14689–14706
November 4-9, 2025 ©2025 Association for Computational Linguistics

that used to design task-specific LLM prompts. Our framework encompasses the following novelties and contributions:

1. To the best of our knowledge, *QUARTZ* is the first framework to enhance LLM-based dialogue summarization in a reference-free setting.

2. *QUARTZ* generates multiple candidate summaries per dialogue and introduces a new two-stage selection process to identify the best summary given a set of task-oriented questions and LLM-derived answers from the summaries. In stage one, it identifies the best answers for each summary-question pair. In stage two, it selects the most informative summary based on these answers.

3. Even without any gold summary, fine-tuning on the selected generated summaries further boosts summarization effectiveness.

4. Our framework is applicable to real-world scenarios, for instance to provide assistance and facilitate the summarization of clinical conversations.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization has attracted growing interest due to its utility in domains such as healthcare and business. Unlike document summarization, it involves additional challenges like frequent semantic shifts, redundancies and multi-speaker dynamics. To address these, Zhao et al. (2021) leveraged dialogue state tracking to enhance coherence, while Tian et al. (2024) used a Mixture-of-Experts (MoE) (Jacobs et al., 1991) LLM to process utterances via specialized experts and fusing their representations to produce summaries. Despite these efforts, methods still struggle with coherence, information coverage and overall grasp of dialogue structure (Zhao et al., 2021). Task-oriented dialogue summarization offers a more focused perspective. For instance, Wang et al. (2023) proposed instruction-guided summarization by deriving targeted queries from reference summaries to steer LLM generation. However most task-oriented approaches require a decent amount of dialogue data with human-written summaries to achieve cutting-edge performance (Zou et al., 2021), low-resource settings have been addressed via data augmentation based on pre-trained language models. Ouyang et al. (2023) generated dialogue-summary pairs to create new training instances, and Liu et al. (2022) replaced sections from the input dialogues and summaries using generated text. In contrast, the area of unsupervised dialogue summarization, which we tackle in this paper, remains largely unexplored.

2.2 LLMs for Data Augmentation

Data Augmentation (DA) refers to strategies aimed at enhancing training data diversity without additional manual collection (Feng et al., 2021). It has been widely applied in Natural Language Processing (NLP) tasks, including low-resource machine translation (Xia et al., 2019), summary grounded conversation generation (Gunasekara et al., 2021), and Question Answering (QA) (Guo et al., 2023). Concerning the latter, Yang et al. (2019) proposed a DA method to enhance BERT fine-tuning for open-domain QA using distant supervision, and Riabi et al. (2021) employed a question generation model to augment examples for cross-lingual QA. The rise of LLMs has revolutionized DA, enabling the generation of high-quality synthetic data and reducing the data collection and labeling costs (Tang et al., 2023). Examples include GPT-3 generating synthetic medical dialogue summaries for models training (Chintagunta et al., 2021), and GPT-4 producing an instruction-following dataset for LLaMA fine-tuning (Peng et al., 2023).

2.3 QA Evaluation using LLMs

QA evaluation has emerged as an effective proxy for assessing factuality in summaries (Wang et al., 2024). It ensures that key information in the original text remains accessible and accurate in the summary by comparing candidate answers to predefined gold answers. Historically, lexical matching dominated QA evaluation through techniques like Exact-Match (Izacard and Grave, 2021) or N-gram matching (Chen et al., 2017). Subsequently, similarity metrics such as BERT Matching (Bulian et al., 2022) emerged to assess how closely a candidate answer aligns with the core meaning of the gold answer. More recently, LLMs have proven highly effective for this task, supported by studies demonstrating performance on par with human evaluators (Bavaresco et al., 2024; Törnberg, 2023; Song et al., 2024), offering a cost-effective and scalable alternative. This is typically achieved by conditioning the LLM's behavior using a system

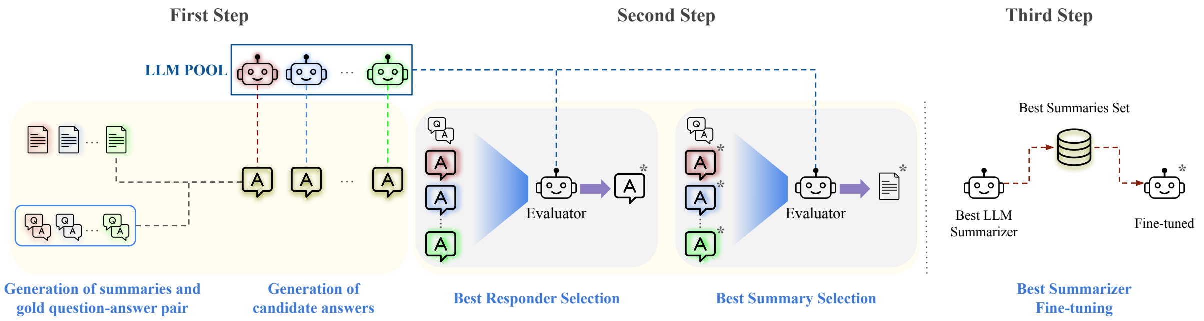

Figure 1: An overview of *QUARTZ* for unsupervised task-oriented dialogue summarization.

prompt, the question, the gold answer and the candidate answer to be assessed (Wang et al., 2024; Kamalloo et al., 2024). This process can be further enhanced through prompt engineering techniques such as In-Context Learning (Brown et al., 2020) and Chain-of-Thoughts (Wei et al., 2022). However, long contexts, can hinder LLM focus, leading to unreliable assessments. This phenomenon was inspected by Liu et al. (2024), who showed that LLMs process different parts of a long context inconsistently, and often neglect information in the middle. To solve this, repeated prompting and score aggregation tend to be a remedy (Tang et al., 2024).

3 Our Approach

The overall workflow of our approach is illustrated in Figure 1. Inspired by recent findings in summarization, showing that LLM summaries are significantly preferred by human evaluators (Pu et al., 2023), *QUARTZ* starts by **(First Step)** prompting a pool of LLMs – rather than a single model – helps mitigate model-specific tendencies and enables the generation of varied task-oriented summaries along with gold question-answer pairs from the input dialogue. Tailored prompts (Appendix D) are used to ensure focus on task-relevant information. We also utilize LLMs as responders to answer the generated questions using the summaries, rather than using the original dialogues. We then **(Second Step)** judge the quality of the generated summaries by assessing the candidate answers using a two-stage evaluation process. In the first stage, LLMs are employed as rankers to select the best candidate answer for each question and summary. In the second stage, LLM ranking reveals the best summary for each dialogue. Finally, **(Third Step)** the best LLM is fine-tuned on the selected summaries to further improve its task-oriented summarization performance. We detail the different stages of *QUARTZ* and its evaluation in the following.

3.1 First Step: Generation

Generation of Task-Oriented Summaries. A task-oriented dialogue is framed as a multi-turn conversation between two or more participants aimed at achieving a specific goal. We prompt each model l_{S} from the pool L to generate one summary $S_{i,l_{\mathrm{S}}}$ for each dialogue D_i, conditioning this process on a carefully crafted task prompt T that fetches the task-specific information.

Generation of Gold Question-Answer Pairs. Similarly, each model $l_{\mathrm{Q}} \in L$ is prompted to generate a set of gold question-answer (QA) pairs from each dialogue D_i, focusing on questions that ought to be answerable using the task-oriented dialogue summary. The gold QA pairs generated by all LLMs are merged into a single set of J_i gold QA pairs $(Q_{i,j}, A_{i,j})$ indexed by j.

Generation of Candidate Answers. With summaries and questions in place, we next derive candidate answers from the generated summaries. Indeed, a factual summary should retain the key information needed to answers all task-relevant questions. To do so, we prompt each model $l_{\mathrm{R}} \in L$ to generate one candidate answer $\hat{A}_{i,j,l_{\mathrm{S}},l_{\mathrm{R}}}$ for each question $Q_{i,j}$ from each summary $S_{i,l_{\mathrm{S}}}$. Recognizing the self-bias phenomenon – where LLMs may easily respond to its own generated question (Xu et al., 2024) – , and aware of the risk of amplifying this bias, we use all LLMs in the pool as responders, including the question generator LLM itself. This reduces bias and provides a fair way of answering the questions.

3.2 Second Step: Two-Stage Evaluation

First-Stage Evaluation *(Best Responder Selection).* At this stage we identify the best candidate answers for each question-summary pair. To do so, we use each model $l_\mathrm{E} \in L$ as an evaluator, receiving the question $Q_{i,j}$, the gold answer $A_{i,j}$, and the set of $|L|$ candidate answers $\{\hat{A}_{i,j,l_\mathrm{S},l_\mathrm{R}}\}_{l_\mathrm{R}\in L}$ obtained from summary S_{i,l_S}, and it is instructed to provide a ranking (defined as a permutation $\sigma : \{1, \ldots, |L|\} \mapsto \{1, \ldots, |L|\}$) for the candidate answers based on their relevance and correctness. To overcome the self-preference bias (Panickssery et al., 2024) by which an LLM evaluator scores its own outputs higher than others, we use all LLMs in the pool as evaluators, including the LLM responder itself. We do this by prompting the LLMs multiple times for ranking while randomly altering the order of the candidate answers and subsequently aggregating the output ranks. This ensures that no LLM has a systematic advantage over the others, and has been theoretically and empirically proven by Tang et al. (2024) to converge to the true ranking. Therefore, we prompt each LLM N times to obtain N estimated rankings $\hat{\sigma}_{i,j,l_\mathrm{S},l_\mathrm{E},n}$ indexed by n. The optimal ranking $\sigma_{i,j,l_\mathrm{S},l_\mathrm{E}}$ is the one whose sum of Kendall tau distances (Kendall, 1938) to all estimated rankings is minimum (Tang et al., 2024):

$$\sigma_{i,j,l_\mathrm{S},l_\mathrm{E}} := \arg\min_{\sigma} \sum_{n=1}^{N} d_\kappa(\hat{\sigma}_{i,j,l_\mathrm{S},l_\mathrm{E},n}, \sigma). \quad (1)$$

The Kendall tau distance $d_\kappa(.,.)$ quantifies the dissimilarity between two rankings, specifically representing the number of *discordant* pairs:

$$d_\kappa(\sigma_1, \sigma_2) := \sum_{k=1}^{|L|} \mathrm{inv}(\sigma_1^{-1} \circ \sigma_2)_k \quad (2)$$

where

$$\mathrm{inv}(\sigma)_k := \#\{k' : \sigma(k') > \sigma(k), k' < k\}. \quad (3)$$

After obtaining the optimal ranking $\sigma_{i,j,l_\mathrm{S},l_\mathrm{E}}$ for each question j, we compute the Mean Reciprocal Rank (MRR) (Radev et al., 2002) of each LLM responder l_R over all questions. The MRR is calculated as follows:

$$\mathrm{MRR}_{i,l_\mathrm{S},l_\mathrm{R},l_\mathrm{E}} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{1}{\sigma^{-1}_{i,j,l_\mathrm{S},l_\mathrm{E}}(l_\mathrm{R})}. \quad (4)$$

Note that the MRR still depends on the evaluator l_E. To select the best responder independently of the evaluator, we compute the total score

$$\mathrm{Score}_{i,l_\mathrm{S},l_\mathrm{R}} = \sum_{l_\mathrm{E}\in L} \alpha_{l_\mathrm{R},l_\mathrm{E}} \mathrm{MRR}_{i,l_\mathrm{S},l_\mathrm{R},l_\mathrm{E}} \quad (5)$$

where the weighting factor

$$\alpha_{l_\mathrm{R},l_\mathrm{E}} = \begin{cases} 0.8 & \text{if } l_\mathrm{R} = l_\mathrm{E} \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

penalizes when the same model serves as both responder and evaluator. Finally, for each dialogue D_i and summary S_{i,l_S}, we select the responder with the highest score and use the corresponding candidate answers $\hat{A}^*_{i,j,l_\mathrm{S}}$ as inputs to the second stage.

Second-Stage Evaluation *(Best Summary Selection).* After selecting the best candidate answers for each summary, we determine the best summary for each dialogue. Similarly to above, we use each model $l_\mathrm{E} \in L$ as an evaluator. The evaluator is provided with a question $Q_{i,j}$, its gold answer $A_{i,j}$, and the set of $|L|$ best candidate answers $\{\hat{A}^*_{i,j,l_\mathrm{S}}\}_{l_\mathrm{S}\in L}$ obtained from all summaries, and is instructed to rank them. We repeat this N times and derive the final ranking $\sigma_{i,j,l_\mathrm{E}}$ as above. The MRR of each LLM summarizer l_S is then computed over all questions as

$$\mathrm{MRR}_{i,l_\mathrm{S},l_\mathrm{E}} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{1}{\sigma^{-1}_{i,j,l_\mathrm{E}}(l_\mathrm{S})}. \quad (7)$$

To select the best summary independently of the evaluator, we compute the total score

$$\mathrm{Score}_{i,l_\mathrm{S}} = \sum_{l_\mathrm{E}\in L} \alpha_{l_\mathrm{S},l_\mathrm{E}} \mathrm{MRR}_{i,l_\mathrm{S},l_\mathrm{E}} \quad (8)$$

with a similar weighting factor as in Equation (6). Finally, for each dialogue D_i, we select the summary S^*_i with the highest score.

Dataset		# Dial.	Avg. Len.	Avg. Turns
SAMSum	Train	14,732	93.8	11.2
	Valid	818	91.6	10.8
	Test	819	95.5	11.3
DialogSum	Train	12,460	131.0	9.5
	Valid	500	129.3	9.4
	Test	1,500	134.5	9.7
MTS-Dialog	Train	1,201	87.99	8.9
	Valid	100	77.46	7.7
	Test	200	87.69	9.1
SimSAMU	-	61	502.47	50.63

Table 1: Dataset statistics. **"# Dial."** refers to the number of dialogues, **"Avg. Len."** represents the average number of tokens per dialogue, and **"Avg. Turns"** indicates the average number of turns per dialogue.

3.3 Third Step: Fine-Tuning

Eventually, we fine-tune the LLM summarizer l_S^* that produced most selected summaries. Although this step involves learning, it remains unsupervised, hence no reference summaries or human labels are used. Training pairs consist of input dialogues and their best selected summaries:

$$\max_{\theta} \sum_{i=1}^{I} \log P(S_i^* \mid D_i, T, \theta) \tag{9}$$

where θ are model parameters and T is the task-specific prompt.

4 Experimental Settings

In the following, we explain our experimental evaluation protocol, including the datasets we used and the implementation details.

4.1 Datasets

We conduct experiments on four task-oriented datasets, summarized in Table 1:
① SAMSum (Gliwa et al., 2019) contains messenger-style conversation between two or more people with their corresponding summaries created by linguists to simulate real-life and task-specific scenarios.
② DialogSum (Chen et al., 2021) includes real-life dialogues spanning diverse task-oriented scenarios (e.g., business negotiation and doctor visits), with a perspective to support downstream applications for both business and personal use.
③ MTS-Dialog (Abacha et al., 2023) features doctor-patient dialogues paired with real-world clinical notes covering multiple specialties (e.g., Neurology, Immunology). Each dialogue is accompanied by a header (e.g., diagnosis, exam) that guides medical report generation.
④ SimSAMU (NUN et al., 2025) is a French medical dispatch dialogue dataset comprising the transcripts of 3 hours of audio recordings of simulated emergency dispatch dialogues across various incidents. Given the characteristics of this dataset displayed in Table 1 (e.g., lengthy dialogues with multiple turns), task-oriented summaries help assess incident severity and support timely, accurate triage decisions.

4.2 Implementation Details

We utilize a pool of three LLMs: *Llama-3.1-8B-Instruct* (Dubey et al., 2024), *Gemma-2-9b-it* (Team et al., 2024), and *Qwen2-7B-Instruct* (Yang et al., 2024). Each model generates 10 to 15 gold QA pairs, which are merged into a unified set of J_i pairs. For LLM ranking, we found that rank consistency is satisfactory for $N = 5$. Further implementation details can be found in Appendix B.

4.3 Evaluation Protocol

We evaluate the test summaries using ROUGE (R-1, R-2 and R-L) (Lin, 2004), BLEU (Papineni et al., 2002) and BERT-Score (Zhang* et al., 2020). Jackknife resampling (Efron and Stein, 1981) is used to provide 95% confidence intervals. Evaluation also includes human judgments (coherence, consistency, fluency, relevance) and LLM-as-Judge scoring.

Baselines. As unsupervised SotA approaches for task oriented dialogue summarization are scarce (Section 2.1), we benchmark our framework mostly against fully supervised methods. Tian et al. (2024) proposed a role-oriented routing based MoE summarizer. Wang et al. (2023) proposed an instruction-tuning method for instructive dialogue summarization, whereas Zhang et al. (2022) introduced momentum calibration to better align the model generation with reference summaries. Abacha et al. (2023) leveraged augmented pre-training followed by fine-tuning on supervised data. In contrast, GPT3-ICL (Suri et al., 2023) prompted OpenAI GPT3 with similar conversations and summaries for each test sample. While in Li et al. (2025), adaptive augmentation fusion is performed throughout training epochs of summarization model on summaries generated by a DA model, guided by a subset of the reference summaries. We also evaluate DeepSeek-R1 distilled models (Guo et al., 2025), available in 8B and 14B sizes, to investigate the impact of the Chain of Thought reasoning process on task-oriented dialogue summarization.

5 Results

5.1 Quantitative Analysis

Comparison with SotA methods.
We begin our analysis using standard metrics capturing both N-gram overlap (R1, R-2, R-L and BLEU) and semantic similarity (BERT-Score). While we acknowledge that these metrics alone cannot fully assess summary quality, they provide a useful foundation for further in-depth analysis.
① SAMSum: Table 2 (top) shows that *QUARTZ*

Method	R-1 ↑	R-2 ↑	R-L ↑	BLEU ↑	BERT-Score ↑
Dataset ① SAMSum – Chit-Chat Domain (Casual, open-domain conversations)					
Supervised Dialogue Summarization					
MoE (Tian et al., 2024)	**55.93**	30.86	**52.02**	**26.03**	**75.66**
InstructDS (Wang et al., 2023)	55.30	**31.30**	46.70	-	55.50
MoCa (Zhang et al., 2022)	55.13	30.57	50.88	-	-
Unsupervised Dialogue Summarization					
Llama-3.1-8B-Instruct (Dubey et al., 2024)	29.93±0.98	10.89±0.93	22.38±0.78	4.85±0.33	58.19±0.51
DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025)	31.69±1.20	10.70±0.93	23.91±0.96	3.45±0.37	59.16±0.56
DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025)	36.04±1.14	12.79±0.98	27.68±0.99	5.56±0.51	61.67±0.56
AAF (Li et al., 2025)	44.57	19.04	35.65	-	-
QUARTZ (Best summarizer: Llama 3.1)	**61.37**±1.36	**38.54**±1.75	**53.12**±1.53	**31.07**±0.89	**77.95**±0.63
Dataset ② DialogSum – Task-Oriented Domain (Real-life dialogues across various scenarios)					
Supervised Dialogue Summarization					
MoE (Tian et al., 2024)	**49.82**	**24.80**	**47.34**	**18.41**	**68.48**
InstructDS (Wang et al., 2023)	47.8	22.2	39.4	-	47.0
Unsupervised Dialogue Summarization					
Llama-3.1-8B-Instruct (Dubey et al., 2024)	21.35±0.79	7.36±0.41	16.46±0.60	3.09±0.19	53.70±0.39
DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025)	27.11±0.77	7.53±0.74	20.37±1.09	2.79±0.21	54.27±0.40
DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025)	29.03±0.77	8.72±0.62	21.99±0.87	3.44±0.32	55.47±0.43
AAF (Li et al., 2025)	**41.38**	15.09	32.52	-	-
QUARTZ (Best summarizer: Llama 3.1)	39.73±1.08	**16.02**±0.87	**32.68**±0.92	**15.73**±0.65	**68.72**±0.44
Dataset ③ MTS-Dialog – Medical Domain (Doctor-patient interactions with clinical notes)					
Supervised Dialogue Summarization					
BART-GS-DA (Abacha et al., 2023)	**42.52**	**17.50**	**34.90**	-	**40.80**
Unsupervised Dialogue Summarization					
GPT3-ICL (Suri et al., 2023)	19.87	8.67	15.60	-	57.03
Llama-3.1-8B-Instruct (Dubey et al., 2024)	28.29±2.37	10.19±1.53	21.14±1.79	7.57±1.28	55.82±1.54
DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025)	17.76±2.31	5.68±1.14	13.02±1.71	2.46±0.60	47.54±1.49
DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025)	26.31±2.85	10.17±1.84	19.78±2.33	4.29±1.01	52.84±1.69
QUARTZ (Best summarizer: Llama 3.1)	**34.63**±2.88	**13.98**±2.19	**27.72**±2.60	**12.75**±2.12	**61.18**±1.73

Table 2: Performance evaluation of *QUARTZ* on the test sets of SAMSum, MTS-Dialog, and DialogSum.

outperforms both SotA supervised and unsupervised methods, achieving relative improvements of +19% BLEU and +3% BERT-Score w.r.t. Tian et al. (2024). Among baselines, DeepSeek-R1-Distill-14B performs the best, slightly ahead of others. The top summarizer (Llama-3.1-8B-Instruct) achieves +34% BERT-Score improvement after *QUARTZ*'s third step. Table 3 shows the different summarizers' performance on this dataset for the First Step (Generation) and illustrates the performance boost from *QUARTZ*'s Step (Evaluation) through collaborative selection.

② DialogSum: While all baselines achieve comparable results, DeepSeek-R1-Distill-14B emerges as the best among them with a BERT-Score of 55.47%. *QUARTZ* outperforms both the baselines and the previous SotA MoE (Tian et al., 2024) by achieving a BERT-Score of 68.72%±0.44.

③ MTS-Dialog: Here again the best summarizer is Llama-3.1-8B-Instruct and its performance improves by 31%, 68% and 9% in terms of Rouge-L, BLEU, and BERT-Score respectively. Notably, it surpasses BART-GS-DA (Abacha et al., 2023) by 50% BERT-Score while outperforming other unsupervised baselines, including GPT-3, Llama, and DeepSeek.

Summaries	R-1	R-2	R-L	BLEU	BERT-Score
Generation (First Step)					
Lama 3.1	**44.95**	19.45	**35.40**	**12.34**	**67.69**
Gemma 2	43.72	18.33	34.40	11.07	67.05
Qwen 2	44.15	**19.55**	35.16	11.64	67.65
Evaluation (Second Step)					
Best Selected	**50.43**	**25.06**	**41.11**	**15.82**	**71.59**

Table 3: Evaluation of the generated and the selected summaries on the SAMSum dataset.

Setting	Vanilla	0-Shot *QUARTZ*	10% Sup.	*QUARTZ*	*QUARTZ* + 10% Sup.	Full Sup.
BERT-Score	53.70	65.99	67.10	68.72	73.02	73.06
Rouge-L	16.46	29.16	29.38	32.68	37.80	37.73

Table 4: Impact of supervision variants on *QUARTZ*. **"Sup."** denotes the ratio of supervised training data used.

The impact of incorporating supervision. While *QUARTZ* can achieve superior performance to supervised methods (see Table 2), it can also leverage limited supervision (10% of the training set) through LoRA fine-tuning (*QUARTZ + 10% Sup.*) to reach near full-supervision results (Table 4). In contrast, training the same baseline (Llama-3.1-8B-Instruct) on the same 10% data fails to exceed a BERT-Score of 68%. Notably, *0-Shot QUARTZ* (i.e., without fine-tuning), surpasses the *Vanilla* baseline by $+23\%$ BERT-Score and $+77\%$ Rouge-L, demonstrating robustness in low-resource or fine-tuning-limited scenarios.

The benefits of a diverse LLM pool over a single model. Recent work (Subramaniam et al., 2025) highlights that using a pool of LLMs promotes diversity and fosters richer model interaction. This is clearly demonstrated in Table 3 where the selected summaries using the LLM pool outperform those generated by each individual model in the pool across all metrics. While Llama-3.1-8B-Instruct generated 58% of the best summaries across all datasets, the contributions of other models are still significant (see Appendix A.2, Figure 3). Figure 2 shows that increasing the LLM pool size ($|L|$) enhances summary diversity and performance, though with higher computational cost (see Appendix A.1).

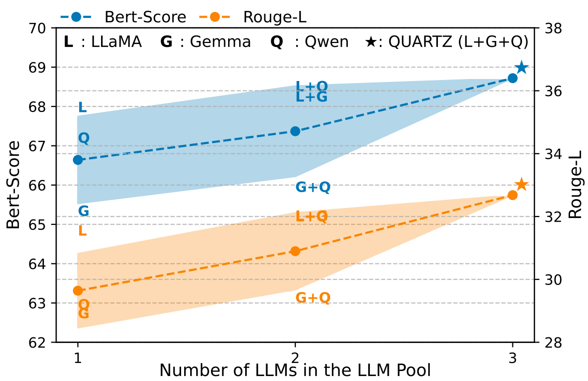

Figure 2: Impact of LLM pool size on summarization performance. ***Dotted lines*** represent the mean performance. ***Shaded regions*** indicate the standard deviation across the combinations (see Appendix A.1).

5.2 Qualitative Analysis

While *QUARTZ* promotes factually consistent and task-relevant summaries via QA-based evaluation, standard metrics (Section 5.1) fall short of capturing this. We therefore conduct a qualitative analysis to examine how *QUARTZ* improves factual consistency and task relevance.

Enhancing task relevance and factual soundness. The selected summaries consistently incorporate more task-relevant information, driven by task-specific QA generation and factual consistency throughout the ranking process of the candidate answers. These selected summaries often introduces new insights absent in other. Examples are provided in Appendix E, Figure 6.

Impact of fine-tuning on summarization quality. On the SimSAMU dataset, the medical expert involved in this study noted that fine-tuned summaries were clearer and directly to the point compared to those generated before fine-tuning. This is particularly valued by emergency regulators in time-sensitive situations. Table 5 shows key enhancements and potential degradations observed (more examples in Appendix E and Table 8).

6 LLM as a Judge

Motivated by recent findings that LLMs can serve as reliable proxies for human annotators (Song et al., 2024), we adopt the G-Eval framework (Liu et al., 2023) using two top-ranked open-weights judge models from the Judge Arena (Zheng et al., 2023): *Selene-1-Mini-Llama-3.1-8B* (Alexandru et al., 2025) and *Llama-3.3-70B-Instruct* (Dubey et al., 2024), to ensure reproducibility, contrary to the original GPT4-based setup. Following Liu et al. (2023) we consider the coherence (COH), consistency (CON), fluency (FLU) and relevance (REL) dimensions (as detailed in Appendix D.4). As shown in Table 6, both judges consistently assign lower scores to reference summaries, suggesting that LLM-generated summaries can already provide strong initial candidates. Although the judge models assign different absolute scores, their rel-

Zero-Shot *QUARTZ* (First Step + Second Step Only)	*QUARTZ* (All Three Steps)	Note
Possible psychiatric disorders or conditions that may be causing the patient's behavior, such as a psychotic disorder.	Psychotic disorder, **possibly schizophrenia or bipolar disorder**.	Conciseness, Specificity
Diagnostic Hypotheses: Self-inflicted stab wound.	Diagnostic Hypotheses: Self-inflicted stab wound, **potential suicidal intent**	Flags suicidal intent
The patient's fourteen-year-old daughter is experiencing abnormal behavior, including difficulty speaking and walking, after a fight with her mother. **The daughter smells of alcohol.**	Altered mental status and **ataxia** in a 14-year-old female following a fight with her mother.	Lacks alcohol detail, Includes medical terminology
The patient is the neighbor of the individual requiring medical attention. **The call was made from the neighbor's home**.	The caller is the patient's neighbor, **calling from the patient's home**.	Incorrect Call Context

Table 5: Impact of *QUARTZ*'s Third Step (Fine-Tuning) on model alignment. **Left**: Zero-Shot *QUARTZ* generates multiple summaries and selects the best ones. **Right**: Additional fine-tuning is applied to the selected summaries

ative rankings are globally preserved. The 0-Shot configuration (*QUARTZ*: Step 1 + Step 2) identifies summaries that perform well on individual metrics (e.g., CON, FLU), but finding candidates that excel across all dimensions remains challenging due to trade-offs in the LLMs-generated summaries set. Fine-tuning addresses this by steering generation toward better overall trade-offs.

Judge / Config	Selene-1-Mini-Llama-3.1-8B					Llama-3.3-70B-Instruct				
	COH	CON	FLU	REL	Avg	COH	CON	FLU	REL	Avg
Llama 3.1	3.79	4.12	3.54	3.86	3.82	4.26	4.34	3.92	4.29	4.20
Gemma	3.77	4.13	3.62	**4.04**	**3.89**	**4.41**	4.39	**4.12**	**4.52**	**4.36**
Qwen	**3.88**	4.12	3.63	3.96	**3.89**	4.29	4.37	3.98	4.41	4.26
0-Shot	3.86	**4.14**	**3.66**	3.86	3.88	4.28	**4.40**	4.00	4.36	4.26
Reference	3.47	3.91	3.31	3.52	3.55	3.89	3.92	3.60	3.86	3.81
QUARTZ	**3.89**	**4.18**	**3.66**	3.87	**3.90**	**4.47**	**4.41**	3.96	**4.53**	**4.34**

Table 6: LLM-as-judge evaluation scores for summaries using two models. COH, CON, FLU, REL, and Avg denote coherence, consistency, fluency, relevance, and average score, respectively.

7 Human Evaluation

Given the high cost of human evaluation, we adopted a two-phase protocol with four expert annotators: all computer scientists, including one with a medical degree to ensure informed assessment, particularly for clinical dialogues.

Phase One. Using the Potato annotation tool (Pei et al., 2022), annotators reviewed each dialogue along with two anonymized summaries, one generated by *QUARTZ* and the other being the reference. Without knowing the source of either summary, they were asked to choose the one they preferred based on overall quality. Notably, in 48% of cases, the annotators favored *QUARTZ* summaries over the reference with a Fleiss' kappa (Fleiss and Cohen, 1973) agreement of 0.14 (slight agreement). Although no "equal quality" option was provided, some annotators reported difficulty in choosing a winner, stating that both summaries were often equally informative. **Phase Two.** Annotators evaluated the final summaries produced by *QUARTZ* according to the four qualitative dimensions as in Section 6. The average scores are presented in Table 7, with annotators reaching 39.5% inter-annotator agreement across all evaluation criteria.

	COH	CON	FLU	REL	Avg
QUARTZ (1–5 Likert)	4.17	4.01	4.27	4.06	4.12
Cohen's kappa ([-1; 1])	0.12	0.17	0.14	0.09	0.13
Exact Agreement (%)	42	42	40	34	39.5

Table 7: Human evaluation for *QUARTZ* final summaries and Cohen's kappa and exact agreement rates.

8 Conclusion and Future Work

We introduced *QUARTZ*, a framework that harnesses a pool of LLMs for unsupervised dialogue summarization. *QUARTZ* begins by generating diverse summaries and task-related QA pairs, then employs a two-stage evaluation process to identify the most informative summaries based on answer quality. Finally, *QUARTZ* fine-tunes the top-performing LLM summarizer on the selected generated summaries. We evaluate it using statistical and embedding-based scorers, LLM as a judge and human evaluation. Experiments across multiple domains demonstrate the effectiveness of *QUARTZ*, consistently surpassing state-of-the-art supervised methods. We believe *QUARTZ* has broad practical value, particularly in domains like healthcare and meeting summarization, where it can help professionals efficiently structure key information and support tasks such as clinical documentation and entity extraction. For future work, we plan to unveil the impact of *iterative QUARTZ* (multi-iteration LLM selection and specialization) on summary quality, and explore other fine-tuning techniques to tackle the pool of LLMs in a different fashion.

9 Limitations

Although we focused on minimizing data annotation costs while maintaining high-quality summaries, some limitations remain. For the sake of simplicity, we did not act on the generated gold question-answer pairs. While we can explicitly incorporate questions relevant to the task, we think that it would be beneficial to establish some control over the questions based on their contribution to summary evaluation, as not all question-answer pairs carry the same level of importance. While *QUARTZ* is designed for unsupervised summarization, the lack of human oversight during training and its reliance on LLM performance may lead to discrepancies from human-preferred summaries or overlook subtle errors. Although human evaluations provided high scores across all dimensions (average ratings over 4 on a 1-5 Likert scale), inter-annotator agreement was low, indicating that even expert judgments might differ substantially. The employment of supplementary evaluation techniques is encouraged by this, which highlights the difficulties in assessing subjective components of summary quality.

10 Ethical Considerations

While this work demonstrates remarkable advancements toward reference-free dialogue summarization, it is important to be aware of the limitations and risks of relying on such approaches in particularly sensitive and high-stakes areas, especially in healthcare. Generative models present serious challenges, such as implicit assumptions embedded in their training data and difficulties in generalizing across wide and complex domains. These limitations will need drastic scrutiny when applying such models in high-stakes environments. For our triage-oriented summarization task, each decision made by the model was checked by a medical expert to ensure its accuracy and relevance. However, the integration of automated systems in healthcare raises ethical and legal concerns regarding accountability and privacy. These considerations further validate our decision to operate on open-source models. Unlike closed-source models, which may involve transferring sensitive private data to third-party companies and expose it to potential misuse, open-source approaches allow for greater transparency and control over the underlying data and processes. Thus, while LLMs have shown promising capabilities, their application in critical fields must be approached with caution to ensure that ethical standards are maintained, and that these systems do not inadvertently compromise patient safety or privacy.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. 2025. Atla selene mini: A general purpose evaluation model. *Preprint*, arXiv:2501.17195.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2016. Temporal TF-IDF: A high performance approach for event summarization in Twitter. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 515–521.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *CoRR*.

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.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware

GPT-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bradley Efron and Charles Stein. 1981. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm-a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, and David Konopnicki. 2021. Summary grounded conversation generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3748–3756.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. 2023. Improving small language models on PubMedQA via generative data augmentation. *arXiv preprint arXiv:2305.07804*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Ehsan Kamalloo, Shivani Upadhyay, and Jimmy Lin. 2024. Towards robust QA evaluation via open LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2811–2816.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Weihao Li, Dan Jiang, Han Zhang, Kejing Xiao, and Shaozhong Cao. 2025. An adaptive fusion-based data augmentation method for abstract dialogue summarization. *PeerJ Computer Science*, 11:e2845.

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

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Aimé NUN, BIROT Olivier, GUIBON Gaël, LAPOSTOLLE Frédéric, and LERNER Ivan. 2025. Simsamu - a french medical dispatch dialog open dataset. *Computer Methods and Programs in Biomedicine*, page 108857.

Siru Ouyang, Jiaao Chen, Jiawei Han, and Diyi Yang. 2023. Compositional data augmentation for abstractive conversation summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1471–1488.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

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

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Kunal Suri, Saumajit Saha, and Atul Singh. 2023. Healthmavericks@ mediqa-chat 2023: Benchmarking different transformer based models for clinical dialogue summarization. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 472–489.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Türe. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.

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

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for BERT fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. 2022. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. TODSum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91.

A LLM Pool Setup and Selection Dynamics

A.1 Details on LLM Pool Configurations

We presented in Figure 2 how the number $|L|$ of LLMs in the LLM pool impacts summarization quality. The LLMs are chosen among these 3 models: Llama-3.1-8B-Instruct, Gemma-2-9b-it, and Qwen2-7B-Instruct. For each pool size $|L|$ we evaluate all $\mathcal{C}_3^{|L|}$ possible model combinations. In Figure 2, dotted lines represent the mean values, while shaded regions indicate the standard deviation. For $|L| = 3$, there is only one possible combination, which corresponds to the actual *QUARTZ* configuration reported in the last row of Table 2 for the DialogSum dataset.

A.2 Prioritizing the Best LLM Summarizer for Fine-Tuning

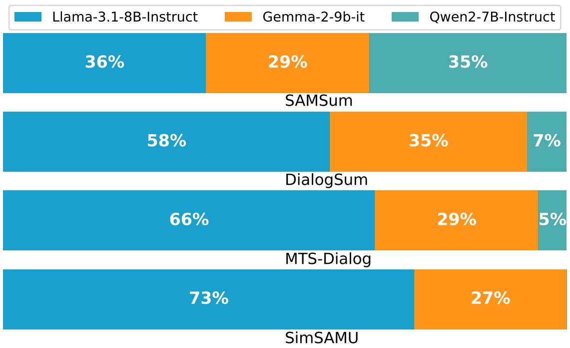

Figure 3: Win Rate of LLMs Across Datasets

QUARTZ's final step (see Section 3.3) aims to fine-tune the best LLM summarizer, i.e., the model that generated the majority of top-selected summaries. Figure 3 illustrates the proportion of best-selected summaries across datasets, with Llama-3.1-8B-Instruct contributing 58%, Gemma-2-9b-it 30%, and Qwen2-7B-Instruct 12%. Empirical findings indicate that fine-tuning the top-performing model yields the most effective summarization results, outperforming alternative selections for the final step.

B Implementation Details

We fine-tune the LLM summarizer using LoRA (Hu et al., 2022) for 3 epochs with rank $r_{\text{LoRA}} = 8$ and scaling factor $\alpha_{\text{LoRA}} = 16$. Optimization is performed using the AdamW optimizer with a learning rate of 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. A linear learning rate scheduler is applied. All experiments were conducted on a single A100-40GB GPU. For full *QUARTZ* pipeline (comprising the 3 steps), the total computation time amounts to 11 GPU hours. Once the best summarizer is obtained, it can be used for online inference and further optimized via inference optimization toolkits.

C Annotator Guidelines

To ensure reliable human evaluation, we conducted a two-phase annotation process with four expert annotators. In **Phase 1**, annotators were presented with a dialogue and two anonymized summaries (one from *QUARTZ* and one from the reference) and asked to select the preferred one. In **Phase 2**, annotators independently rated each *QUARTZ* 's summary along four quality dimensions—*Coherence*, *Consistency*, *Fluency*, and *Relevance* on a 5-point Likert scale. The criteria for scoring were as follows:

- **Coherence (1–5):** Assesses the overall logical flow and structure of the summary.
- **Consistency (1–5):** Evaluates factual alignment with the source dialogue.
- **Fluency (1–5):** Rates grammaticality, readability, and linguistic quality.
- **Relevance (1–5):** Measures how well the summary captures important content without redundancy.

D Prompt Templates

D.1 Summary Generation

SAMSum:

"instruction": You will be provided with a conversational exchange that simulates a natural messaging or chat-like interaction. Your task is to produce a short, concise and clear summary that captures the most important points and key information conveyed throughout the exchange.
"input": [Conversation]

MTS-Dialog:

"instruction": You are a medical scribe tasked with writing concise yet informative medical notes based on doctor-patient interactions. Your goal is to create clear and professional note text about from the patient doctor dialogue focusing on [Header]
"input": [Conversation]

SimSAMU:

Prompt design for this triage-related dataset was assisted by a medical expert to ensure that the generated notes are contextually relevant and support accurate triage decisions during critical incidents.

"instruction": You are an experienced emergency physician handling a telephone consultation. Your task is to summarize the following medical dialogue into a precise and structured clinical report. When summarizing, ensure the following:
- Use concise, clear, and professional language.
- Translate informal or everyday terms into appropriate medical terminology where possible.
- Maintain the structure provided below, leaving any sections blank if the required information is unavailable.
Format of the clinical report:
Complete the following sections in order.
1-Chief Complaint: The main medical issue prompting the call (e.g., chest pain).
2-Call Context: The relationship between the caller and the patient (e.g., patient themselves, spouse, bystander) and the location of the call (e.g., home, street).
3-Patient Context: Demographic information (age, sex), social situation (e.g., lives alone, resides in a retirement home), and degree of autonomy.
4-Usual Treatment: Current medications or treatments for known comorbidities.
5-Past Medical History: Relevant medical, allergic, or surgical history (e.g., diabetes, prior surgeries).
6-Patient Symptoms: Symptoms as reported, categorized into:
- General symptoms (e.g., fever, fatigue).
- Organ-specific symptoms (e.g., respiratory: shortness of breath).
7-History of Present Illness: A detailed, chronological account of the events leading to the call, describing their sequence and interrelations (free text).
8-Diagnostic Hypotheses: Possible diagnoses based on the information

provided (e.g., myocardial infarction).
9-Treatment Plan: Recommendations including medications, therapies, lifestyle advice, referrals, or additional diagnostic tests.
10-Triage Decision: The proposed course of action (e.g., remain at home, dispatch a doctor, self-transport to the emergency department, send an ambulance, or dispatch a medicalized ambulance)
"input": [Conversation]

D.2 QAs Generation

SimSAMU:

"instruction":Given this emergency call between a doctor and a patient, generate 10 to 15 question and answer pairs that are relevant to the medical triage and that should be present in the clinical record.
Do not repeat the same question or answer.
Do not ask questions that you can't answer based on the information provided in the dialogue.
Format the questions and answers as follows: Q1: <question1> A1: <answer1>
"input":[Conversation]

D.3 Response Evaluation

"instruction": You will be provided with a ground truth answer and a list of generated answers. Your task is to rank the generated answers based on their correctness and closeness to the ground truth answer. The ground truth answer appears first, followed by the list of generated answers. The ranking should be a three-element list of integers between 1 and 3, where 1 represents the most accurate answer and 3 represents the least accurate. If an answer is NOT_INCLUDED, place it at the end of the ranking.
"input": Question: [Question]
Ground Truth Answer: [Ground Truth Answer]
Possible answers:
1) [Answer_1]
2) [Answer_2]
3) [Answer_3]

D.4 LLM as a Judge

Coherence metric:

"instruction": You are a helpful assistant tasked with evaluating how coherent a summary is for a given dialogue. Your goal is to rate the summary based only on how well its sentences form a clear, logical, and well-structured presentation of the dialogue content. Assign a score from 1 to 5 based solely on *coherence*:
5: Excellent - the summary is well-organized, easy to follow, and logically structured.
4: Good - mostly coherent with only minor issues in flow or structure.
3: Fair - somewhat coherent but with noticeable issues in clarity or organization.
2: Poor - disorganized, unclear, or hard to follow.
1: Very poor - sentences feel disconnected or incoherent, severely impacting understanding.
Only reply with the number **1**, **2**, **3**, **4**, or **5**. Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>
"input": Dialogue: [Dialogue]
Summary: [Summary]

Consistency metric:

"instruction": You are a helpful assistant tasked with evaluating how factually consistent a summary is with a given dialogue. Your goal is to rate the summary based only on whether it accurately reflects the facts stated in the original dialogue without introducing unsupported or hallucinated information. Assign a score from 1 to 5 based solely on *consistency*:
5: Excellent — all statements in the summary are fully supported by the dialogue.
4: Good — minor inaccuracies or slight overgeneralizations, but mostly faithful to the dialogue.
3: Fair — some factual inconsistencies or minor hallucinations are present.
2: Poor — several statements in the summary are not supported or contradict the dialogue.
1: Very poor — the summary contains major

hallucinations or is largely inconsistent with the dialogue.
Only reply with the number **1**, **2**, **3**, **4**, or **5**. Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>
“input”: Dialogue: [Dialogue]
Summary: [Summary]

Fluency metric:

“instruction”: You are a helpful assistant tasked with evaluating how fluent a summary is for a given dialogue. Your goal is to rate the summary based only on grammar, spelling, punctuation, word choice, and sentence structure. Assign a score from 1 to 5 based solely on *fluency*:
5: Excellent – the summary is free of errors and reads very smoothly.
4: Good – the summary has minor errors but is easy to read.
3: Fair – the summary has some noticeable errors that slightly affect clarity or flow.
2: Poor – the summary has many errors that affect understanding or naturalness.
1: Very Poor – the summary contains frequent errors making it difficult to understand.
Only reply with the number **1**, **2**, **3**, **4**, or **5**. Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>
“input”: Dialogue: [Dialogue]
Summary: [Summary]

Relevance metric:

“instruction”: You are a helpful assistant tasked with evaluating how relevant a summary is for a given dialogue. Your goal is to rate the summary based only on how well it includes important information from the dialogue while avoiding irrelevant or redundant content. Assign a score from 1 to 5 based solely on *relevance*:
5: Excellent – the summary covers all key points and contains no irrelevant or redundant information.
4: Good – the summary covers most key points with minimal irrelevant or redundant content.
3: Fair – the summary covers some key points but includes noticeable irrelevant or redundant information.
2: Poor – the summary misses many key points and contains significant irrelevant or redundant content.
1: Very Poor – the summary largely fails to capture key points and is mostly irrelevant or redundant.
Only reply with the number **1**, **2**, **3**, **4**, or **5**. Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>
“input”: Dialogue: [Dialogue]
Summary: [Summary]

E Example *QUARTZ* Summaries

SimSAMU:

We use RED to highlight information present in the pre-fine-tuning summary but omitted after fine-tuning, and GREEN to indicate information added in the post-fine-tuning summary, in order to visualize differences between the two model outputs.

Example 01:

1-Chief Complaint: Self-inflicted stab wound.
2-Call Context: The patient’s cousin called for the brother, who is the patient. The call was made from the patient’s home.
3-Patient Context: Age and sex are unavailable. The patient lives at home with his brother. The degree of autonomy is unclear. It is mentioned that the patient’s brother is in a residence where the patient resides.
4-Usual Treatment: Unavailable.
5-Past Medical History: Unavailable.
6-Patient Symptoms: General symptoms: agitation, refusal to seek help. Organ-specific symptoms: unavailable.
7-History of Present Illness: The patient’s brother was found with blood on his pants after the patient saw him stab himself again. The brother refused to let the patient look at the wound and did not want him to call for help. The patient’s cousin called for help instead. The patient was reluctant to involve

the authorities, fearing his brother's reaction.
8-Diagnostic Hypotheses: Self-inflicted stab wound, potential suicidal intent.
9-Treatment Plan: Dispatch of firefighters to the scene to handle the situation. The patient is advised to stay with his brother and remind the authorities of any new developments.
10-Triage Decision: The emergency services decided to Dispatch of firefighters[2] to the scene to handle the situation .

Example 02:
1-Chief Complaint: Malaise, possibly a vagal malaise, episode of feeling like leaving without losing consciousness, characterized by a feeling of impending syncope without actual loss of consciousness.
2-Call Context: The patient is the daughter of the 85-year-old man experiencing the malaise, calling from her home.
3-Patient Context: The patient is an 85-year-old man with hypertension, living with his daughter, who is an emergency worker.
4-Usual Treatment: He takes treatments for hypertension.
5-Past Medical History: Hypertension.
6-Patient Symptoms: General symptoms: Malaise, feeling of impending syncope. Organ-specific symptoms: None mentioned.
7-History of Present Illness: The patient experienced a malaise while in the bathroom, characterized by a feeling of impending syncope. He called his daughter, who assisted him and he recovered., who arrived and turned him into a supine position with his legs elevated, which improved his condition.
8-Diagnostic Hypotheses: Vagal malaise.
9-Treatment Plan: The patient should be taken to the emergency room for further evaluation. He should bring his prescription for his usual treatments.
10-Triage Decision: The patient should be transported to the emergency room by his daughter, who should ensure his comfort during the transport.

[2]Firefighters are commonly responsible for responding to medical emergencies in France.

DialogSum:
In Figure 6, we present a dialogue example alongside the summaries and question-answer pairs generated by LLMs. Notably, the best-selected summary (LLM Summarizer: Gemma-2-9b-it) incorporates additional facts beyond both the generated and reference summaries. While this enhances task-oriented summarization by enriching the information provided, it is undervalued by surface-level metrics.

Dialogue

#Person1#: Hello Mrs. Parker, how have you been?
#Person2#: Hello Dr. Peters. Just fine thank you. Ricky and I are here for his vaccines.
#Person1#: Very well. Let's see, according to his vaccination record, Ricky has received his Polio, Tetanus and Hepatitis B shots. He is 14 months old, so he is due for Hepatitis A, Chickenpox and Measles shots.
#Person2#: What about Rubella and Mumps?
#Person1#: Well, I can only give him these for now, and after a couple of weeks I can administer the rest.
#Person2#: OK, great. Doctor, I think I also may need a Tetanus booster. Last time I got it was maybe fifteen years ago!
#Person1#: We will check our records and I'll have the nurse administer and the booster as well. Now, please hold Ricky's arm tight, this may sting a little.

Reference Summary

Mrs Parker takes Ricky for his vaccines. Dr. Peters checks the record and then gives Ricky a vaccine.

Generated Summaries

Llama: Ricky needs Hepatitis A, Chickenpox and Measles shots. Mrs. Parker also needs a Tetanus booster.

(Best Selected Summary) ***Gemma***: Dr. Peters is giving Ricky, **a 14-month-old boy**, his vaccines. Ricky is due for **Hepatitis A, Chickenpox and Measles shots**. His mother, **Mrs. Parker, also needs a Tetanus booster**.

Qwen: Dr. Peters administers vaccines for Ricky and offers a Tetanus booster to Mrs. Parker.

Generated QAs

Q1: Who takes Ricky for his vaccines?
A1: Mrs Parker
⋮
Q5: How old is Ricky??
A5: **14 months old**.
⋮
Q9: What vaccines can Dr. Peters give Ricky immediately?
A9: Dr. Peters can give Ricky the **Hepatitis A, Chickenpox and Measles shots**.
Q10: Does Mrs. Parker need a vaccine too?
A10: Yes, Mrs. Parker needs a **Tetanus booster**.

Figure 4: An example dialogue from the DialogSum dataset. **Left:** The original dialogue and the reference summary. **Right:** The generated summaries from each LLM in the pool along with the task-related QAs. Additional information introduced by the best-selected summary is highlighted in **bold**.

Dialogue

Kerri: Did you see the apartment?
Stella: yes brb
Kerri: ok
Stella: Sorry I was driving
Stella: I saw it yesterday but it was already dark so I want to go again to see it in the morning
Stella: First of all, its really big, 130 sqm
Kerri: WOW
Stella: I could split it into 2 apartments
Stella: 90 and 40 sqm more or less
Kerri: Or 80 and 50?
Stella: Not really, there's only one way to split it
Stella: So I could stay in the big one with 3 bedrooms and 2 bathrooms
Kerri: Sounds good
Stella: I want to see it in the morning
Stella: The agent said the bedrooms are very sunny in the morning
Stella: Maybe I'll go tomorrow
Kerri: Took some photos?
Stella: <file_photo>
Kerri: Nice!!!
Kerri: The kitchen looks fantastic
Kerri: Its huge
Stella: It's the size of my living room!
Stella: We obviously don't need 3 bedrooms
Stella: But I could have a decent home office at last
Kerri: Has Greg seen it?
Stella: No, he's away this week

Reference Summary

Stella wants to see the apartment in the morning. She plans to split it into 2 apartments.

Generated Summaries

Llama: Stella saw an apartment of 130 sqm. She wants to visit it in the morning to see it in daylight. The apartment has 3 bedrooms and 2 bathrooms. Stella can split it into two apartments, 90 and 40 sqm, but only one way. She took some photos of the apartment.

Gemma: Stella saw a large apartment (130 sqm) and is considering splitting it into two units. She likes the apartment and wants to see it again in the morning to assess the sunlight in the bedrooms. The apartment has a large kitchen and Stella envisions having a home office.

(Best Selected Summary) ***Qwen:*** Stella saw the apartment **yesterday** and wants to visit it again in the morning because she couldn't see it properly. She thinks the apartment is very big (**130 sqm**) and plans to split it into two smaller apartments (**90 and 40 sqm**). The bedrooms are **sunny** in the morning, and Stella wants to see the apartment again **tomorrow**. The kitchen is **huge**, and she could use one of the bedrooms as a home office. Stella hasn't shown the apartment to Greg yet because he's away this week.

Generated QAs

Q1: When did Stella see the apartment previously?
A1: She saw it **yesterday**.
Q2: What is the size of the apartment?
A2: 130 sqm.
Q3: How large would each apartment be if split?
A3: 90 sqm and 40 sqm.
⋮
Q10: Does the agent say anything about the bedrooms?
A10: The agent said the bedrooms are **very sunny** in the morning.

Figure 5: Another example dialogue from the DialogSum dataset. **Left:** The original dialogue and the reference summary. **Right:** The generated summaries from each LLM in the pool along with the task-related QAs. Additional information introduced by the best-selected summary is highlighted in **bold**.

Dialogue

Jimmy: Hey, guess what? My car's completely kaput!
Carrie: What happened?
Jimmy: Don't know, either the battery, alternator, engine or a combination of all three
Carrie: Where is it now?
Jimmy: Yesterday I managed to jump-start it and it was fine, but then Amy took the car and stopped to get something at the store, and then she couldn't get it to start again
Carrie: Sucks! Wanna borrow my Mercedes?
Jimmy: How are you gonna get it to me? You live so far from the city now
Carrie: I don't live that far! It's only like 20 km
Jimmy: I'll let you know
Carrie: I can drive it into the city, and then you can drive me back home, then you'll have the car :)
Jimmy: :) Hmm… I still gotta do something about my car
Carrie: Actually, you should buy mine and get rid of yours
Jimmy: :-/ Yeah, lately there's always something with it. Anyways, I'll let you know. I may need it, Amy's sister is coming to visit.
Carrie: Ok, see ya
Jimmy: Bye

Reference Summary

Jimmy's car is broken down. Jimmy will either borrow or buy Carrie's car. Jimmy needs the car as Amy's sister is coming to visit.

Generated Summaries

Llama: Jimmy's car broke down and he might need to borrow Carrie's Mercedes. Carrie suggested that Jimmy buy her car and get rid of his own.

(Best Selected Summary)Gemma: Jimmy's **car broke down**. Carrie offered to lend him her **Mercedes**, but Jimmy lives far from the city. Carrie suggested Jimmy **buy her car** instead. Jimmy is considering the offer but needs to figure out his car situation first.

Qwen: Jimmy's car broke down. Carrie offered to lend him her car, but Jimmy needs to sort out his car first. Jimmy might borrow Carrie's car when Amy's sister visits.

Generated QAs

Q1: What is wrong with Jimmy's car?
A1: It could be the **battery**, **alternator**, **engine**, or a combination of all three.
⋮
Q4: What vehicle does Carrie offer Jimmy
A4: A **Mercedes**.
⋮
Q10: How would Carrie deliver the Mercedes to Jimmy?
A10: She would drive it into the city and have Jimmy drive her back home.

Figure 6: Another example dialogue from the SAMSum dataset. **Left:** The original dialogue and the reference summary. **Right:** The generated summaries from each LLM in the pool along with the task-related QAs. Additional information introduced by the best-selected summary is highlighted in **bold**.

Zero-Shot *QUARTZ* (First Step + Second Step Only)	***QUARTZ* (All Three Steps)**	**Note**
Possible psychiatric disorders or conditions that may be causing the patient's behavior, such as a psychotic disorder.	Psychotic disorder, **possibly schizophrenia or bipolar disorder**.	Conciseness, Specificity
Diagnostic Hypotheses: Self-inflicted stab wound.	Diagnostic Hypotheses: Self-inflicted stab wound, **potential suicidal intent**	Flags suicidal intent
The patient's husband called for his wife, who is experiencing a medical issue.	The patient's husband called for his wife, who is being **followed by a psychiatrist for depression**.	Adds context
The patient is experiencing malaise and has had a fall, with symptoms including a spinning head and loss of balance.	Malaise, **syncope (fainting)**, and a fall.	Includes medical terminology
The patient also has a sore throat since the morning and a pain in the left arm.	The pain is located in the **upper abdomen and radiates to the left arm**.	Precise anatomical detail
Possible myocardial infarction, considering the chest pain and arm pain.	Possible diagnoses include peptic ulcer disease, gastritis, or a cardiac event given the radiation of pain to the left arm.	Expanded differential diagnosis
The doctor advised the patient to monitor their symptoms.	If symptoms worsen, **the patient should call the 15**.	Explicit emergency guidance *(15 is the french 911)*
Home surveillance, and regular checks on the child's condition.	Home surveillance, frequent checks **every 3 hours**, and monitoring for any changes in the child's condition.	clearer medical guidance
Stab wound to the upper arm.	Stab wound, **possible injury to underlying structures (e.g., nerves, blood vessels)**.	Improved clinical specificity
Triage Decision: Dispatch an ambulance.	Triage Decision: The patient is to be transported to the hospital by **car**.	Adjusted transport method
The patient's fourteen-year-old daughter is experiencing abnormal behavior, including difficulty speaking and walking, after a fight with her mother. **The daughter smells of alcohol.**	Altered mental status and **ataxia** in a 14-year-old female following a fight with her mother.	Lacks alcohol detail, Includes medical terminology
The patient is the neighbor of the individual requiring medical attention. **The call was made from the neighbor's home**.	The caller is the patient's neighbor, **calling from the patient's home**.	Incorrect Call Context

Table 8: Further summary examples before and after fine-tuning on the *QUARTZ* selected summaries on the SimSAMU dataset. **Left**: Zero-Shot *QUARTZ* generates multiple summaries and selects the best ones. **Right**: Additional fine-tuning is applied to the selected summaries.