arXiv:2507.02145v1 [cs.CL] 2 Jul 2025

Reasoning or Not? A Comprehensive Evaluation of Reasoning LLMs for Dialogue Summarization

Keyan Jin[a,d], Yapeng Wang[a], Leonel Santos[b], Tao Fang[c], Xu Yang[a], Sio Kei Im[a] and Hugo Gonçalo Oliveira[d]

[a]*Faculty of Applied Sciences, Macao Polytechnic University, Macao, China*
[b]*Computer Science and Communication Research Center, School of Technology and Management, Polytechnic of Leiria, Leiria, Portugal*
[c]*Institute of International Language Services Studies, Macau Millennium College, Macao, China*
[d]*CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal*

ARTICLE INFO

Keywords:
Dialogue Summarization
Large Language Models
LLMs Evaluation

ABSTRACT

Dialogue summarization is a challenging task with significant practical value in customer service, meeting analysis, and conversational AI. Although large language models (LLMs) have achieved substantial progress in summarization tasks, the performance of step-by-step reasoning architectures—specifically Long Chain-of-Thought (CoT) implementations such as OpenAI-o1 and DeepSeek-R1—remains unexplored for dialogue scenarios requiring concurrent abstraction and conciseness. In this work, we present the first comprehensive and systematic evaluation of state-of-the-art reasoning LLMs and non-reasoning LLMs across three major paradigms—generic, role-oriented, and query-oriented dialogue summarization. Our study spans diverse languages, domains, and summary lengths, leveraging strong benchmarks (SAMSum, DialogSum, CSDS, and QMSum) and advanced evaluation protocols that include both LLM-based automatic metrics and human-inspired criteria. Contrary to trends in other reasoning-intensive tasks, our findings show that explicit stepwise reasoning does not consistently improve dialogue summarization quality. Instead, reasoning LLMs are often prone to verbosity, factual inconsistencies, and less concise summaries compared to their non-reasoning counterparts. Through scenario-specific analyses and detailed case studies, we further identify when and why explicit reasoning may fail to benefit—or even hinder—summarization in complex dialogue contexts. Our work provides new insights into the limitations of current reasoning LLMs and highlights the need for targeted modeling and evaluation strategies for real-world dialogue summarization.

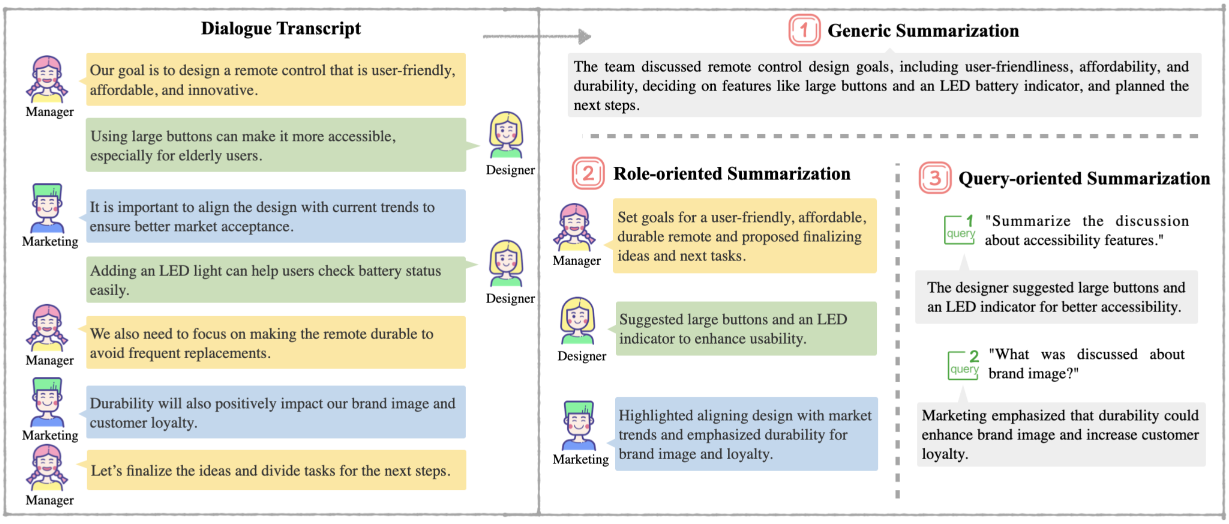
1. Introduction

Dialogue summarization is a critical natural language processing task that supports numerous practical applications, such as customer service, meeting analysis, and conversational AI assistants. Unlike traditional document summarization, dialogue summarization must handle unique challenges, including multi-party interactions, fragmented utterances, ambiguous references, and frequent topic shifts. Additionally, effective summarization can facilitate automated meeting documentation, collaborative decision-making, and efficient information retrieval from dialogue records. Early advances relied primarily on extractive methods that selected key sentences based on simple heuristics like TF–IDF or word frequency (Marcu, 1997), before evolving to neural approaches such as Seq2Seq and Pointer-Generator networks, which enabled more fluent abstractive summaries (Rush et al., 2015; See et al., 2017). Subsequently, significant breakthroughs were achieved by adapting Transformer-based neural architectures to conversational settings (Lewis et al., 2019; Liang et al., 2022; Jin et al., 2025). Large language models (LLMs) have achieved remarkable results across a wide variety of natural language processing tasks, including text classification, sentiment analysis, question answering, and translation, demonstrating strong generalization capabilities and state-of-the-art performance (Brown et al., 2020). In particular, reasoning LLMs, such as OpenAI-o1, DeepSeek-R1, and QwQ-32B, have exhibited notable advantages in tasks requiring complex reasoning, such as mathematical problem solving, logical inference, and machine translation (Chen et al., 2025a; Ye et al., 2025). These successes naturally prompt further exploration into their applicability within dialogue summarization.

Dialogue summarization encompasses multiple distinct paradigms, each reflecting real-world scenarios that vary significantly in language, domain, dialogue length, and user intent. As illustrated in Figure 1, the three

---

ORCID(s):



Manager: Our goal is to design a remote control that is user-friendly, affordable, and innovative.

Designer: Using large buttons can make it more accessible, especially for elderly users.

Marketing: It is important to align the design with current trends to ensure better market acceptance.

Designer: Adding an LED light can help users check battery status easily.

Manager: We also need to focus on making the remote durable to avoid frequent replacements.

Marketing: Durability will also positively impact our brand image and customer loyalty.

Manager: Let's finalize the ideas and divide tasks for the next steps.

The team discussed remote control design goals, including user-friendliness, affordability, and durability, deciding on features like large buttons and an LED battery indicator, and planned the next steps.

Manager: Set goals for a user-friendly, affordable, durable remote and proposed finalizing ideas and next tasks.

Designer: Suggested large buttons and an LED indicator to enhance usability.

Marketing: Highlighted aligning design with market trends and emphasized durability for brand image and loyalty.

"Summarize the discussion about accessibility features."

The designer suggested large buttons and an LED indicator for better accessibility.

"What was discussed about brand image?"

Marketing emphasized that durability could enhance brand image and increase customer loyalty.

Figure 1: An illustrative example of three dialogue summarization tasks explored in this study: Generic Summarization, Role-oriented Summarization, and Query-oriented Summarization.

widely recognized paradigms include generic summarization, which involves summarizing the entire dialogue; role-oriented summarization, focusing specifically on the perspectives of distinct dialogue participants; and query-oriented summarization, aiming to satisfy particular information requests. Despite considerable progress, existing research lacks a systematic and comprehensive evaluation of reasoning-based LLMs compared to traditional LLMs across diverse dialogue summarization scenarios. In particular, there is currently no multi-dimensional evaluation that examines their comparative performance across different summarization paradigms, languages, domains, and dialogue lengths.

In this work, we present a systematic, scenario-driven evaluation of reasoning and traditional LLMs for dialogue summarization. Our experiments cover three core paradigms—generic, role-oriented, and query-oriented summarization—across diverse benchmarks including SAMSum (Gliwa et al., 2019), DialogSum (Chen et al., 2021), CSDS (Lin et al., 2021), and QMSum (Zhong et al., 2021). We comprehensively evaluate leading reasoning models (OpenAI-o1, DeepSeek-R1, QwQ-32B) and their non-reasoning counterparts across different languages, domains, and dialogue lengths. To ensure a fair and consistent comparison across models, our evaluation exclusively utilizes prompting, without any task-specific fine-tuning. By combining LLM-based automatic evaluation and human-inspired fine-grained metrics, we assess summary quality from multiple perspectives and investigate how explicit reasoning processes shape model performance. Our analysis is further enriched by detailed case studies that highlight the limitations of reasoning LLMs in real-world dialogue summarization. Our main contributions are as follows:

- We establish a unified framework for dialogue summarization, covering generic, role-oriented, and query-oriented paradigms, and provide clear definitions and scenario mappings for each.
- We present the first large-scale, systematic evaluation of leading reasoning-oriented LLMs and their non-reasoning counterparts on widely adopted dialogue summarization datasets, reporting results across multiple languages, domains, and summary types.
- We identify open challenges in automatic and LLM-based evaluation, and offer practical insights for selecting and developing models tailored to complex dialogue summarization needs.

The rest of the paper is organized as follows. Section 2 discusses related work on dialogue summarization and the evaluation of reasoning-oriented large language models. Section 3 describes our research methodology, including the problem definition, evaluation framework, datasets, and prompt design. In Section 4, we present experimental results covering generic, role-oriented, and query-oriented summarization tasks. Section 5 provides a detailed analysis of the intrinsic quality of summaries, comparative evaluations using large language models as judges, an in-depth analysis of reasoning processes, and a case study. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related Work

In this section, we review recent advancements in dialogue summarization with LLMs, including their methodological innovations, emerging evaluation approaches, and existing challenges. We further discuss recent efforts to systematically evaluate reasoning LLMs, emphasizing their capabilities and limitations across complex reasoning tasks.

2.1. Dialogue Summarization with LLMs

Recent progress in dialogue summarization has been driven by the advent of LLMs, instruction tuning, and advanced prompt engineering. Early studies focused on fine-tuning smaller models or exploiting heuristic approaches, but instruction-tuned LLMs have shown remarkable ability in both generic and role-oriented dialogue summarization. For example, Baichuan2-Sum demonstrates that role-specific instructions can significantly improve summary quality for multiple roles within a dialogue (Xiao et al., 2024). Multi-stage pre-training and mutual reinforcement between synthesis and summarization can further enhance LLM adaptability across diverse scenarios (Zhou et al., 2023; Lu et al., 2025). Prompt optimization, such as prompt scoring and in-context example selection, has also proven critical for controlling summary content, style, and focus, with studies showing that both prompt format and demonstration quality can greatly affect summarization performance (Tang et al., 2023; Block et al., 2023; Okadome et al., 2024). For cross-lingual and scenario adaptation, zero-shot and few-shot LLMs are now competitive with supervised approaches, but remain challenged by verbosity and cultural adaptation (Wang et al., 2023).

A key challenge for LLM-based dialogue summarization lies in evaluating and improving factual consistency, faithfulness, and overall summary quality. Recent research underscores that traditional summarization metrics, particularly ROUGE, are inadequate for fully capturing essential quality dimensions like consistency and relevance in dialogue summarization tasks (Gao and Wan, 2022). As a result, a wide range of new evaluation frameworks and benchmarks have emerged. Fine-grained and multi-dimensional evaluation datasets—such as FineSurE (Song et al., 2024) and UniSumEval (Lee et al., 2024)—enable the systematic assessment of faithfulness, completeness, conciseness, and hallucination at the sentence or span level. Laban et al. (Laban et al., 2023) propose SummEdits, which directly tests models' ability to reason about facts in summaries. TofuEval further highlights that hallucinations remain a persistent issue even for advanced LLMs (Tang et al., 2024), and that LLM-based evaluators are not always superior to specialized factuality metrics. Newer error taxonomies, such as contextual or circumstantial inference (Ramprasad et al., 2024), reveal that LLMs often generate plausible but unsupported content, making automated detection difficult. Studies on guideline-following (Zhou et al., 2025), extractive+abstractive hybrid models (Zhang et al., 2023; Mishra et al., 2023), and mixture-of-experts strategies (Tian et al., 2024) further enrich the landscape of summary evaluation and model design.

Finally, beyond mainstream English conversation, dialogue summarization research is expanding to more complex and realistic domains. Multi-domain and multi-scenario benchmarks test LLMs' robustness across varied dialogue styles, speaker structures, and background knowledge requirements (Zhou et al., 2023). Methods such as recursive summarization for long-term dialogue memory (Wang et al., 2025) and factual knowledge distillation from LLMs to smaller models (Zhu et al., 2025) point to the growing ecosystem of hybrid, scalable, and adaptive approaches in the field. Collectively, these advances signal a paradigm shift toward LLM-powered dialogue summarization—supported by a rapidly evolving set of evaluation tools and practical solutions for controlling, scaling, and verifying model outputs in real-world applications.

2.2. Evaluation of Reasoning LLMs

Recent years have seen a surge of interest in reasoning LLMs, exemplified by OpenAI's o1, DeepSeek-R1, QwQ, and the new Qwen3 series (DeepSeek-AI et al., 2025; Yang et al., 2025). These models are specifically optimized to emulate human-like multi-step reasoning, often leveraging reinforcement learning or chain-of-thought (CoT) training paradigms to improve their performance on tasks involving complex logic, problem-solving, and decision-making (Yang et al., 2025). Recent works have demonstrated that such "reasoning LLMs" can achieve competitive or even state-of-the-art results on benchmarks requiring planning, mathematical reasoning, commonsense, or domain adaptation, often surpassing traditional LLMs in semantically demanding scenarios (Ye et al., 2025; Valmeekam et al., 2025). However, it has also been observed that these models come with trade-offs: inference is typically slower and more costly, and models may suffer from issues such as overthinking, redundant reasoning, and unstable output length or quality (Hashemi et al., 2025; Zeng et al., 2025). For example, it has been shown that o1-like models often allocate excessive computational resources to simple problems, yielding limited benefits and resulting in inefficient scaling

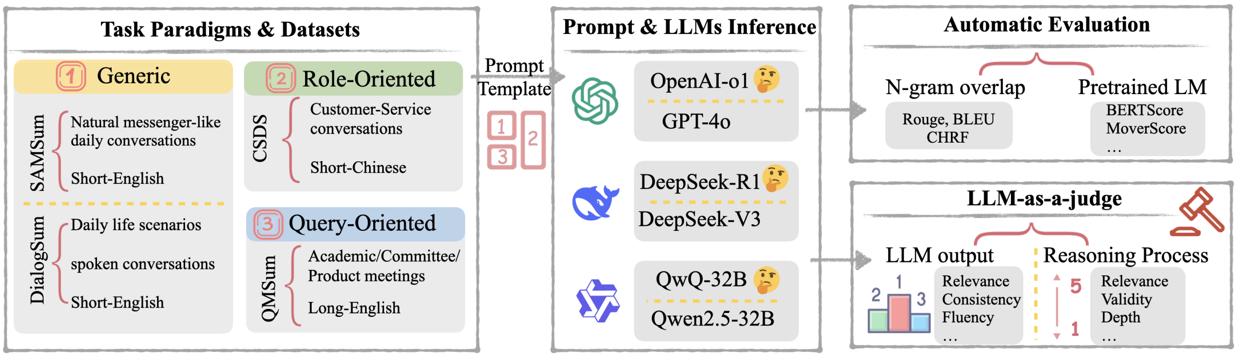


Figure 2: Overview of our experimental framework. We evaluate multiple reasoning and non-reasoning LLMs across three dialogue summarization paradigms using diverse datasets.

during inference (Chen et al., 2025c). Other studies have noted that longer reasoning chains do not always improve accuracy and may actually increase the likelihood of self-contradictions and unnecessary self-revisions (Zeng et al., 2025).

The evaluation of reasoning LLMs has grown increasingly systematic, with dedicated benchmarks for reflection, planning, factual reasoning, and human-like social cognition. For instance, the LR^{2}Bench was introduced to assess long-chain reflective reasoning and constraint satisfaction, showing that even the most advanced reasoning models still struggle on compositional logic and multi-step constraint problems, leaving substantial room for improvement (Chen et al., 2025b). Comparative studies have revealed that models like DeepSeek-R1, OpenAI-o1, and QwQ are effective on complex translation, mathematical, or planning tasks, but their performance can degrade due to over-reasoning and “rambling” in certain languages or domains (Chen et al., 2025a; Ye et al., 2025; Valmeekam et al., 2025). Overthinking and excessive token generation remain open issues, as shown by DNR Bench, which probes whether models can recognize when reasoning is unnecessary and avoid generating superfluous content (Hashemi et al., 2025).

3. Methodology

In this section, we formally define the three dialogue summarization tasks explored in this study and describe our experimental framework, including datasets, prompts, and evaluation metrics.

3.1. Problem Formulation

In this study, we explore dialogue summarization tasks under three distinct scenarios: generic dialogue summarization, role-oriented dialogue summarization, and query-based dialogue summarization. Formally, given a dialogue D composed of a sequence of utterances, each utterance u_i is associated with a speaker role s_i. Thus, a dialogue with N utterances and M unique speaker roles can be denoted as:

$$D = \{(u_1, s_1), (u_2, s_2), \ldots, (u_N, s_N)\}, \quad s_i \in \{s_1, \ldots, s_M\} \tag{1}$$

The generic dialogue summarization task aims to generate a concise summary Y^{final} that covers the essential content of the entire dialogue D. This process can be modeled as a sequence-to-sequence generation task, formally defined as:

$$Y^{final} = \arg\max_{Y} P(Y|D) \tag{2}$$

In role-oriented dialogue summarization, we focus on generating summaries from the perspectives of specific speaker roles within the dialogue. Given the dialogue D and a set of predefined roles $S = \{s_1, \ldots, s_M\}$, the task is to produce distinct summaries Y^{s_j} tailored to each role s_j:

$$Y^{s_j} = \arg\max_{Y} P(Y|D, s_j), \quad s_j \in S \tag{3}$$

Table 1
Statistics of datasets used in this work. Length is measured in average number of tokens.

Datasets	Lang.	Description	Length	#Turns	Length of Sum	#Speakers	Dialogues	Sum Type
SAMSum	EN	Written messenger online	95.51	11.25	23.12	2.36	819	generic
DIALOGSUM	EN	Daily life spoken	134.46	13.85	18.75	2	1500	generic
CSDS	CN	Customer service	387.10	25.10	Overall / User / Agent 83.21 / 37.28 / 48.08	2	800	generic / role based
QMSum								
Academic	EN	Academic meetings	13317.30	819.00	53.70	6.30	56	generic / query based
Committee	EN	Committee meetings	13761.90	207.70	80.50	34.10	72	generic / query based
Product	EN	Product meetings	6007.71	535.60	70.50	4.00	151	generic / query based
ALL	EN	N/A	9069.8	556.80	69.60	9.20	279	generic / query based

Moreover, dialogues often involve multiple topics or decision points, necessitating the query-oriented dialogue summarization. Given a query Q composed of words $(w_1, w_2, \dots, w_{|Q|})$, the task produces a targeted summary Y^Q responsive to the query:

$$Y^Q = \arg\max_{Y} P(Y|D, Q) \tag{4}$$

Queries can be categorized into general queries (Q_{gen}), requesting overall dialogue summaries, and specific queries (Q_{spec}), focusing on particular topics or roles. Thus, query-based summarization tasks are explicitly defined as:

$$Y^{gen} = \arg\max_{Y} P(Y|D, Q_{gen}), \quad Y^{spec} = \arg\max_{Y} P(Y|D, Q_{spec}) \tag{5}$$

3.2. Evaluation Framework

In this subsection, we present our evaluation framework designed to systematically assess and compare reasoning and non-reasoning LLMs. We introduce the overall structure of our evaluation process, the benchmark datasets selected for different summarization tasks, our standardized prompt design strategies, and the detailed automatic metrics used for quantitative evaluation.

3.2.1. Overall Framework

We establish a unified evaluation framework to enable a rigorous, scenario-specific comparison of reasoning and non-reasoning LLMs for dialogue summarization. The overall design of our framework is illustrated in Figure 2, which outlines each stage of our experimental pipeline—from task paradigms and datasets to model inference, and finally to comprehensive evaluation.

Our framework begins by covering three representative summarization paradigms: generic, role-oriented, and query-oriented summarization. Each paradigm is paired with real-world benchmark datasets that reflect a range of dialogue types, domains, languages, and summary requirements. This setup allows us to test models under a wide spectrum of realistic and challenging scenarios.

For each task, we use standardized prompt templates to ensure fairness and reproducibility across both reasoning LLMs and their non-reasoning counterparts. Detailed descriptions of prompt design and task-specific configurations are presented in a dedicated section later in the paper.

To thoroughly assess model performance, our evaluation combines both automatic and LLM-based human-aligned metrics. We report a range of widely used n-gram overlap metrics as well as pretrained language model-based scores. Beyond these surface-level measures, we incorporate LLM-as-a-judge protocols: large models are used as automatic evaluators to rank or rate generated summaries on key aspects such as relevance, consistency, fluency, and overall quality. For reasoning LLMs that output explicit stepwise traces, we further evaluate the quality of their reasoning process along targeted criteria.

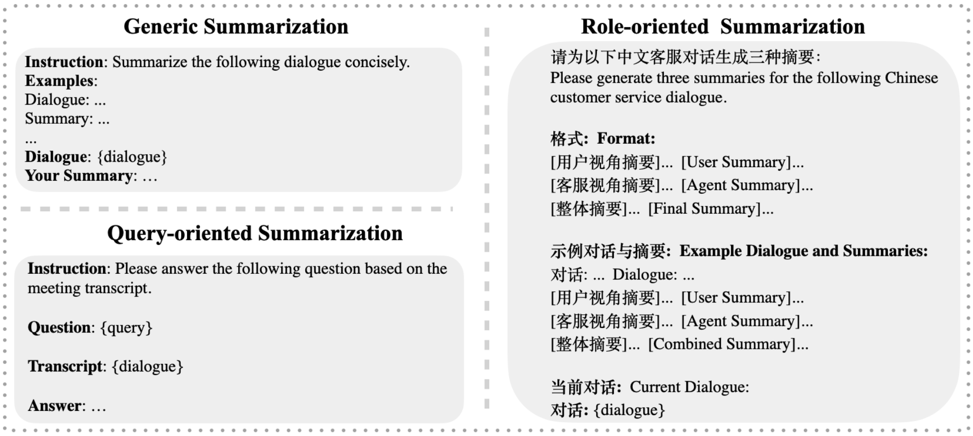
3.2.2. Datasets

To comprehensively evaluate the summarization performance of reasoning-oriented o-1-like models and non-o1-like models, we select four representative datasets: **SAMSum**[1] (Gliwa et al., 2019), **DIALOGSUM**[2] (Chen et al.,

---

[1] `https://huggingface.co/datasets/Samsung/samsum`

[2] `https://github.com/cylnlp/DialogSum`



Generic Summarization

Instruction: Summarize the following dialogue concisely.
Examples:
Dialogue: ...
Summary: ...
...
Dialogue: {dialogue}
Your Summary: …

Query-oriented Summarization

Instruction: Please answer the following question based on the meeting transcript.

Question: {query}

Transcript: {dialogue}

Answer: …

Role-oriented Summarization

请为以下中文客服对话生成三种摘要:
Please generate three summaries for the following Chinese customer service dialogue.

格式: Format:
[用户视角摘要]... [User Summary]...
[客服视角摘要]... [Agent Summary]...
[整体摘要]... [Final Summary]...

示例对话与摘要: Example Dialogue and Summaries:
对话: ... Dialogue: ...
[用户视角摘要]... [User Summary]...
[客服视角摘要]... [Agent Summary]...
[整体摘要]... [Combined Summary]...

当前对话: Current Dialogue:
对话: {dialogue}

Figure 3: Prompt templates for generic, role-oriented, and query-based dialogue summarization tasks used in our experiments. Note that for role-oriented summarization, since the evaluated dataset is in Chinese, we directly use Chinese prompts as shown. The English translations in the figure are provided for clarity.

2021), **CSDS**[3] (Lin et al., 2021), and **QMSum**[4] (Zhong et al., 2021). Each dataset possesses unique characteristics, enabling evaluation of summarization capabilities under diverse dialogue scenarios and summarization tasks. The detailed statistics of these datasets are summarized in Table 1.

SAMSum is widely recognized for its manually annotated, high-quality messenger-style dialogues, primarily reflecting informal online communication scenarios. It is extensively used as a benchmark dataset in dialogue summarization research, allowing us to directly compare and assess the summarization performance of various models within informal conversational contexts.

DIALOGSUM is distinctive due to its extensive collection of spoken dialogues sourced from public corpora and an English speaking practice website. The dialogues represent typical daily-life scenarios such as education, employment, healthcare, shopping, and leisure activities. The dataset uniquely emphasizes observer-perspective summarization, making it highly suitable for evaluating models' abilities to generalize summarization tasks across various everyday conversational settings.

CSDS specifically addresses customer service dialogues in Chinese, offering distinctive features such as role-oriented summarization alongside general summaries. This characteristic allows the evaluation of summarization methods in handling complex dialogues involving distinct speaker perspectives (customers and service agents) and clearly structured topics. Such features position CSDS as a critical dataset for assessing models' capabilities in nuanced, topic-centric summarization tasks.

QMSum is uniquely designed for the summarization of lengthy, multi-participant meetings across academic, committee, and product development contexts. Distinctively employing a query-based summarization paradigm, QMSum challenges models to selectively generate summaries based on targeted queries. This approach provides an effective evaluation scenario to test summarization accuracy and relevance within complex and extensive dialogues.

3.2.3. Prompt Design

In terms of prompt construction, we carefully designed task-specific prompts for each of the three summarization scenarios: generic summarization, role-oriented summarization, and query-based summarization. The detailed prompt templates for each task are illustrated in Figure 3.

3.2.4. Outcome Evaluation

We comprehensively evaluate the summarization quality of reasoning and non-reasoning models using various automatic evaluation metrics. These metrics are broadly categorized into two groups: traditional n-gram overlap-based metrics and pretrained language model-based metrics. We describe each group of metrics in detail below.

---

[3] `https://github.com/xiaolinAndy/CSDS`

[4] `https://github.com/Yale-LILY/QMSum`

Traditional N-gram Overlap-based Metrics These metrics evaluate the lexical similarity between the generated summaries Y and reference summaries Y^* based on n-gram overlaps:

ROUGE[5] (Lin, 2004): Specifically, we use ROUGE-1, ROUGE-2, and ROUGE-L, capturing unigram, bigram, and longest common subsequence overlaps, respectively. ROUGE-N is computed as:

$$\text{ROUGE-N} = \frac{\sum_{S \in Y^*} \sum_{\text{gram}_n \in S} \text{Count}_{match}(\text{gram}_n)}{\sum_{S \in Y^*} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{6}$$

BLEU[6] (Papineni et al., 2001): BLEU evaluates precision by calculating the geometric mean of modified n-gram precisions (usually up to four-grams), with a brevity penalty (BP) to discourage overly short summaries:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

CHRF[7] (Popović, 2015): CHRF calculates the harmonic mean (F-score) based on precision and recall at the character n-gram level, defined formally as:

$$\text{CHRF} = (1+\beta^2)\frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{8}$$

Typically, $\beta = 2$ emphasizes recall.

Pretrained Language Model-based Metrics These metrics leverage pretrained language models to evaluate semantic similarity beyond simple lexical overlaps, capturing deeper semantic alignments between generated and reference summaries:

BERTScore[8] (Zhang* et al., 2019): Utilizes contextual embeddings derived from pretrained BERT models to measure semantic similarity between candidate summary tokens $\hat{x}$ and reference summary tokens x. Specifically, precision (P_{BERT}), recall (R_{BERT}), and F1-score (F_{BERT}) are computed as:

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j, \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j, \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{9}$$

MoverScore[9] (Zhao et al., 2019): Employs optimal transport (Earth Mover's Distance) to calculate semantic distances between embeddings from generated and reference summaries. The metric finds the minimum "cost" of aligning candidate and reference embeddings, thus capturing nuanced semantic similarities:

$$\text{MoverScore} = 1 - \text{EMD}(E(Y), E(Y^*)) \tag{10}$$

where $E(\cdot)$ denotes the contextual embeddings of the summary.

BARTScore[10] (Yuan et al., 2021): Evaluates summaries using pretrained BART sequence-to-sequence models, measuring the log-likelihood scores across different directions including source-to-hypothesis (s-h), reference-to-hypothesis (r-h), and hypothesis-to-reference (h-r). The general form is:

$$\text{BARTScore} = \frac{1}{m}\sum_{i=1}^{m} \log P(y_i | y_{<i}, x; \theta_{\text{BART}}) \tag{11}$$

COMET[11] (Rei et al., 2020): Originally proposed for machine translation evaluation, COMET leverages multilingual pretrained language models (such as XLM-Roberta) to calculate semantic similarity scores. Given source text src, hypothesis hyp, and reference ref, COMET predicts a quality score q:

$$q = f_{\text{COMET}}(E(src), E(hyp), E(ref)) \tag{12}$$

---

[5] `https://github.com/Diego999/py-rouge`

[6] `https://www.nltk.org/_modules/nltk/translate/bleu_score.html`

[7] `https://github.com/mjpost/sacrebleu`

[8] `https://github.com/Tiiiger/bert_score`. We use *roberta-large* for English and *bert-base-chinese* for Chinese.

[9] `https://github.com/AIPHES/emnlp19-moverscore`. We use *distilbert-base-uncased* for English and *bert-base-chinese* for Chinese.

[10] `https://github.com/neulab/BARTScore`. We use *bart-large-cnn* for English and *bart-base-chinese* for Chinese.

[11] `https://huggingface.co/Unbabel/wmt22-comet-da`.We use *wmt22-comet-da* for evaluation

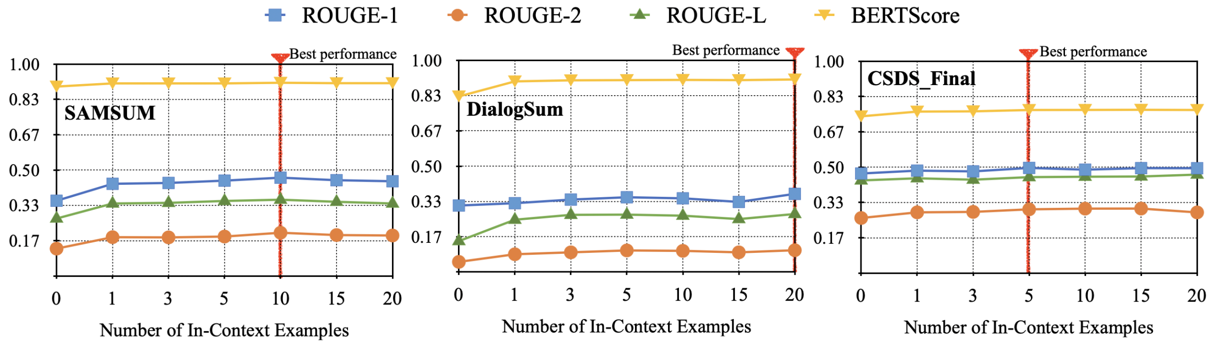


Figure 4: Performance comparison under varying numbers of in-context examples across different datasets using `gpt-4o`. The best-performing number of in-context examples for each dataset is highlighted by a vertical red line and annotated as "Best performance." For **CSDS**, we report results based on the **Final** evaluation set. Metrics include ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore.

4. Experiments

In this section, we describe the selected models, experimental settings, and detailed results of our comprehensive evaluation across the three dialogue summarization paradigms.

4.1. Model Selection

To thoroughly evaluate dialogue summarization performance under different reasoning capabilities, we selected three reasoning LLMs: OpenAI-o1, DeepSeek-R1, and QwQ-32B. Each of these reasoning models is directly compared against a corresponding non-reasoning base model, ensuring fair and insightful evaluation.

OpenAI-o1 is a reasoning-optimized model developed by OpenAI, designed to tackle complex tasks in science, coding, and mathematics. It employs a "think before answering" approach, generating internal chains of thought to enhance its problem-solving abilities. This model is compared with **GPT-4o**[12].

DeepSeek-R1 (DeepSeek-AI et al., 2025) is a reasoning-specialized LLM developed by DeepSeek-AI, trained using large-scale reinforcement learning to solve complex reasoning tasks across domains such as math, code, and language. Its reasoning capabilities are further enhanced through supervised fine-tuning to improve readability and coherence. DeepSeek-R1 is compared with **DeepSeek-V3**, as they share the same underlying base architecture.

QwQ-32B (Yang et al., 2025) is a medium-sized reasoning model from the Qwen series, developed by Alibaba's Qwen team. It leverages reinforcement learning techniques to enhance its reasoning capabilities, achieving competitive performance in mathematical reasoning, coding, and complex problem-solving tasks. QwQ-32B is evaluated against **Qwen-2.5-32B-Instruct**, its non-reasoning counterpart.

4.2. Experimental Settings

To evaluate the selected LLMs under realistic and reproducible conditions, all experiments were conducted via public APIs, ensuring consistent inference environments across models. Specifically, both OpenAI-o1 and GPT-4o were accessed using the official OpenAI API. For DeepSeek-R1, DeepSeek-V3, QwQ-32B, and Qwen-2.5-32B-Instruct, we utilized the APIs provided by the Aliyun Bailian platform[13]. For all models, We used the default parameters provided by the APIs for all models unless otherwise specified.

4.3. Experimental Results

In this section, we first conduct preliminary ablation experiments to determine optimal experimental configurations. Then, we systematically evaluate the performance of reasoning and non-reasoning LLMs across three dialogue summarization scenarios. We analyze and compare their strengths and limitations using multiple automated evaluation metrics.

---

[12] We hypothesize that OpenAI-o1 and GPT-4o share the same foundational architecture; however, explicit confirmation has not been publicly disclosed.

[13] `https://bailian.console.aliyun.com/`

Table 2
Summarization performance of reasoning and non-reasoning LLMs on SAMSUM and DIALOGSUM datasets using 14 automatic evaluation metrics. Best values are in bold.

	Model	R1	R2	R-L	BLEU	CHRF	BERTScore			Mover	BARTScore			COMET	Time(s)
							P	R	F1		S→H	R→H	H→R		
Samsum	**Reasoning LLMs**														
	OpenAI-o1	45.06	18.54	35.33	7.83	41.26	89.58	91.88	90.70	26.07	-2.17	-2.69	-2.24	72.96	5.49
	DeepSeek-R1	43.90	17.50	34.30	7.64	41.07	89.16	91.80	90.45	24.55	-2.40	-2.99	-2.25	71.98	22.06
	QwQ-32B	43.58	17.42	34.05	7.60	39.56	89.55	91.55	90.52	24.36	-2.31	-2.89	-2.35	71.60	12.98
	Non-Reasoning LLMs														
	GPT-4o	47.01	20.79	37.13	**9.62**	43.03	**90.28**	92.13	**91.18**	**28.81**	-2.08	-2.60	**-2.21**	**74.09**	**1.32**
	DeepSeek-v3	**47.21**	**20.94**	**37.67**	9.61	**43.07**	90.16	**92.15**	91.13	28.44	**-2.03**	**-2.55**	-2.22	74.03	3.28
	Qwen2.5-32B	44.88	19.24	35.45	8.75	41.73	89.69	91.96	90.79	26.02	-2.17	-2.73	-2.24	73.11	2.77
DialogSum	**Reasoning LLMs**														
	OpenAI-o1	37.99	13.25	29.67	5.33	42.27	88.66	91.80	90.19	14.03	-2.05	-2.70	-2.41	70.12	9.83
	DeepSeek-R1	37.59	12.69	29.26	5.31	41.62	88.63	91.52	90.04	13.65	-2.40	-2.97	-2.44	69.28	48.06
	QwQ-32B	36.11	10.88	27.81	4.64	38.44	88.64	90.68	89.63	11.28	-2.51	-3.11	-2.71	67.65	13.13
	Non-Reasoning LLMs														
	GPT-4o	40.76	15.06	31.93	6.47	43.90	89.79	92.10	**90.92**	18.27	-2.06	-2.59	-2.32	72.28	5.42
	DeepSeek-v3	**41.51**	**16.25**	**32.93**	**7.06**	**44.37**	**89.82**	**92.07**	**90.92**	**18.93**	**-1.92**	**-2.48**	**-2.36**	**72.40**	8.81
	Qwen2.5-32B	39.17	13.79	30.78	5.84	42.41	89.41	91.90	90.62	16.51	-2.17	-2.69	-2.39	71.15	**3.64**

4.3.1. Preliminary Ablation Results

To ensure the optimality of experimental outcomes, our study systematically conducted ablation experiments to investigate the selection of example quantities in few-shot learning. Specifically, we employed the GPT-4o API to perform In-Context Learning (ICL) ablation experiments (Dong et al., 2024), across both general summarization and role-oriented summarization tasks. To comprehensively evaluate the impact of ICL example quantities on model performance, we configured varying numbers of example samples—namely, 1, 3, 5, 10, 15, and 20—and conducted a detailed analysis of the results. Performance evaluation was carried out using multiple automated metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore, to ensure the comprehensiveness and objectivity of the assessment.

The experimental results, illustrated in Figure 4, reveal the performance trends across different ICL example quantity configurations on the validation sets of each dataset. We determined the optimal ICL example quantities for the SAMSum, DialogSum, and CSDS datasets to be 10, 20, and 5, respectively. This optimal configuration will be applied in all subsequent experiments to ensure the stability and consistency of model performance.

For the query-oriented summarization task (QMSum dataset), we noted that conference transcript texts are typically lengthy, often approaching or exceeding the token limits of the API. Consequently, we adopted a zero-shot prompting strategy. This approach not only effectively addressed the challenges of processing long texts but also significantly enhanced the feasibility and computational efficiency of dialogue summarization evaluation.

4.3.2. Evaluation of Generic Dialogue Summarization

As summarized in Table 2, non-reasoning LLMs demonstrate a clear and consistent advantage over reasoning LLMs across both the SAMSum and DialogSum datasets. Non-reasoning LLMs consistently outperform reasoning LLMs across almost all automatic evaluation metrics.

Notably, DeepSeek-V3 and GPT-4o achieve the strongest overall performance on both datasets. For example, on SAMSum, DeepSeek-V3 obtains the highest ROUGE-1/2/L and CHRF scores, while GPT-4o achieves the best BERTScore (F1), MoverScore, and inference efficiency. Similar patterns are observed on DialogSum, with DeepSeek-V3 again leading most metrics, closely followed by GPT-4o. The performance gap is especially pronounced in metrics that reflect both surface overlap (ROUGE, BLEU, CHRF) and semantic similarity (BERTScore, COMET).

In contrast, reasoning LLMs consistently lag behind, both in terms of summary quality and inference speed. Despite being optimized for multi-step reasoning, these models fail to deliver notable advantages for generic dialogue summarization tasks, and in some cases, such as DeepSeek-R1, exhibit significantly slower inference.

These results clearly suggest that explicit reasoning instruction tuning does not generally benefit dialogue summarization tasks that emphasize brevity and direct extraction of core content. Nonetheless, we recognize a potential strength of reasoning LLMs not fully captured by conventional automatic metrics: their ability to explicitly outline decision-making processes or generate plausible explanations supporting summarization choices. Such reasoning

Table 3
Summarization performance of reasoning LLMs and non-reasoning LLMs on CSDS dataset from three role perspectives. Best values are in bold.

Role	Model	R1	R2	R-L	BLEU	CHRF	BERTScore			Mover	BARTScore			COMET
							P	R	F1		S→H	R→H	H→R	
User	**Reasoning LLMs**													
	OpenAI-o1	49.75	30.16	45.06	16.16	17.17	75.86	78.06	76.77	51.88	-4.36	-3.67	-3.44	82.93
	DeepSeek-R1	51.50	32.23	47.02	17.55	18.31	76.29	78.16	77.02	50.17	-4.40	-3.65	-3.38	82.83
	QwQ-32B	49.55	30.05	44.65	16.42	17.27	74.60	77.35	76.25	50.80	-4.42	-3.76	-3.39	83.50
	Non-Reasoning LLMs													
	GPT-4o	52.53	33.61	48.39	18.73	18.11	77.95	78.07	77.82	52.32	-4.39	-3.55	-3.42	83.43
	DeepSeek-v3	**54.40**	**36.58**	**50.74**	**20.84**	**19.83**	**79.88**	**78.78**	**79.13**	**54.34**	**-4.38**	**-3.43**	**-3.37**	**84.18**
	Qwen2.5-32B	51.57	32.56	47.35	17.91	17.99	77.22	78.21	77.52	52.06	-4.39	-3.61	-3.42	83.63
Agent	**Reasoning LLMs**													
	OpenAI-o1	46.84	28.14	41.92	14.89	15.59	75.00	75.16	74.90	51.59	-4.37	-3.65	-3.56	82.34
	DeepSeek-R1	46.45	27.85	41.62	14.11	15.15	74.40	74.60	74.32	50.88	-4.42	-3.71	-3.55	82.12
	QWQ-32B	45.83	26.61	40.57	13.56	14.71	73.53	74.94	74.08	51.12	-4.41	-3.77	-3.57	82.61
	Non-Reasoning LLMs													
	GPT-4o	48.15	29.77	43.33	15.83	16.38	75.35	75.32	75.16	51.77	-4.34	-3.61	-3.52	83.11
	DeepSeek-v3	**49.86**	**32.37**	**45.45**	**17.87**	**18.11**	**76.87**	**75.97**	**76.22**	**52.83**	**-4.30**	**-3.48**	**-3.46**	**83.39**
	Qwen2.5-32B	48.01	29.66	43.42	16.17	16.75	75.12	75.73	75.26	52.54	-4.33	-3.62	-3.51	83.59
Final	**Reasoning LLMs**													
	OpenAI-o1	51.03	31.09	46.61	17.43	17.69	76.14	76.86	76.37	58.29	-4.12	-3.46	-3.33	83.25
	DeepSeek-R1	51.54	32.28	47.63	17.89	18.04	76.09	76.69	76.26	56.78	-4.18	-3.49	-3.31	83.22
	QwQ-32B	49.87	29.98	45.41	16.68	17.16	74.63	76.68	75.52	57.42	-4.17	-3.55	-3.33	83.50
	Non-Reasoning LLMs													
	GPT-4o	53.10	34.02	49.21	19.43	18.82	77.55	77.09	77.19	58.27	-4.11	-3.38	-3.31	84.38
	DeepSeek-v3	**54.44**	**36.53**	**51.07**	**21.44**	**20.41**	**79.01**	**77.77**	**78.24**	**59.16**	**-4.08**	**-3.28**	**-3.25**	**84.78**
	Qwen2.5-32B	52.42	33.36	48.69	19.13	18.92	76.95	77.30	76.99	58.57	-4.12	-3.43	-3.31	84.68

transparency could be advantageous in scenarios requiring interpretability and trust, even though these benefits may not directly enhance measured summary quality.

4.3.3. CSDS: Role-Oriented Chinese Dialogue Summarization

Table 3 reports the results for all models on the CSDS dataset, which focuses on Chinese customer service dialogues and provides reference summaries from user, agent, and overall perspectives. Across all roles, the reasoning LLMs consistently underperform compared to their respective base models on nearly all automatic metrics.

For instance, in the user summary setting, DeepSeek-V3 achieves the highest ROUGE score, while its reasoning variant DeepSeek-R1 scores notably lower on all these metrics. Similar trends are observed for GPT-4o versus OpenAI-o1 and for Qwen2.5-32B versus QwQ-32B. In the agent summary task, DeepSeek-V3 again leads with the best ROUGE and BERTScore values, and the gap between base and reasoning-tuned models remains substantial. The final (generic) summary results reinforce this observation: DeepSeek-V3 and other base models consistently achieve higher scores across ROUGE, BLEU, CHRF, and BERTScore than their reasoning counterparts.

These consistent patterns suggest that reasoning instruction tuning does not enhance performance for role-oriented abstractive summarization in Chinese dialogues. On the contrary, it often leads to reduced summary quality. One plausible explanation is that such tuning encourages models to engage in excessive reasoning or elaboration, which may detract from the concise and focused nature required for effective summarization. Consequently, general-purpose models, when equipped with appropriate prompting and in-context learning strategies, exhibit better adaptability and summarization quality across diverse dialogue scenarios and perspectives.

4.3.4. QMSum: Query-Oriented Long Dialogue Summarization

Table 4 presents the performance of all evaluated models on the QMSum dataset, which features long-form meeting transcripts and evaluates both query-based and combined summarization scenarios. The experimental results reveal a consistent trend: non-reasoning LLMs achieve noticeably better performance than their reasoning-oriented counterparts across all domains and evaluation perspectives.

This advantage holds regardless of meeting type, whether academic, committee, or product-oriented discussions. In each case, the base models exhibit stronger alignment with human reference summaries, both in terms of lexical overlap and semantic adequacy. Notably, the superiority of base models is apparent not only in settings focused solely

Table 4
Summarization performance on the Academic, Committee, and Product domains, as well as the overall aggregate set (ALL) of QMSum, evaluated using ROUGE, BLEU, CHRF, BERTScore$_{F1}$, and COMET. For each metric, results are reported as *specific / all*, where *specific* denotes query-based summarization, and *all* denotes the aggregate performance over both query-based and generic summarization targets.

Datasets	Model	R1	R2	R-L	BLEU	CHRF	BERT$_{F1}$	COMET
Academic	**Reasoning LLMs**							
	OpenAI-o1	22.53/23.17	4.24/4.12	13.52/13.37	1.13/1.07	28.55/30.32	84.46/84.22	63.97/64.38
	DeepSeek-R1	19.80/20.80	3.98/3.79	12.32/12.32	1.16/1.09	26.54/28.94	83.67/83.53	62.34/62.78
	QwQ-32B	16.79/17.85	3.59/3.52	10.54/10.62	0.84/0.81	22.86/25.12	82.67/82.55	62.10/62.31
	Non-Reasoning LLMs							
	GPT-4o	**24.26/25.00**	5.23/5.14	**15.51/15.45**	**1.74/1.66**	30.67/32.66	**85.51**/85.31	63.85/64.61
	DeepSeek-V3	21.32/22.18	**5.81/5.73**	14.12/14.23	1.59/1.55	28.35/30.52	85.18/85.02	**65.54/66.31**
	Qwen2.5-32B	22.93/24.44	5.36/5.31	14.82/15.22	1.47/1.45	**32.00/34.30**	85.47/**85.49**	64.32/65.04
Committee	**Reasoning LLMs**							
	OpenAI-o1	27.89/28.08	6.19/6.22	16.03/15.86	1.72/1.64	32.89/33.76	84.95/84.74	65.13/65.34
	DeepSeek-R1	24.16/24.70	5.84/5.71	14.33/14.32	1.51/1.44	31.37/32.40	83.83/83.61	63.24/63.33
	QwQ-32B	20.29/21.04	5.22/5.20	12.32/12.43	1.20/1.15	26.80/28.06	82.74/82.61	63.07/63.13
	Non-Reasoning LLMs							
	GPT-4o	**29.84/30.40**	8.63/8.60	**18.39/18.34**	**3.05/2.89**	36.16/37.26	86.04/85.92	65.80/66.32
	DeepSeek-V3	27.12/27.57	8.53/8.50	17.07/17.05	2.71/2.57	34.25/35.27	85.51/85.38	**67.00/67.41**
	Qwen2.5-32B	29.03/29.76	**8.87/8.80**	17.89/18.00	2.96/2.85	**36.56/37.64**	**85.92/85.87**	66.71/67.19
Product	**Reasoning LLMs**							
	OpenAI-o1	29.54/29.89	6.35/6.57	16.51/16.44	1.79/1.69	33.62/34.94	85.26/85.06	65.76/66.09
	DeepSeek-R1	25.62/26.17	6.10/6.02	15.02/15.01	1.52/1.43	32.81/33.77	84.03/83.74	63.97/64.03
	QwQ-32B	21.97/22.85	5.72/5.73	13.18/13.34	1.29/1.23	28.86/30.12	82.91/82.77	64.01/64.11
	Non-Reasoning LLMs							
	GPT-4o	**30.80/31.55**	8.85/8.97	**18.59/18.63**	2.77/2.62	36.93/38.15	**86.24/86.17**	67.18/67.77
	Deepseek-V3	28.81/29.39	8.86/8.95	17.83/17.82	2.69/2.52	35.78/36.92	85.84/85.72	**68.20/68.65**
	Qwen2.5-32B	30.26/31.07	**9.22/9.25**	18.35/18.45	**2.80/2.70**	**37.28/38.40**	86.17/86.11	68.15/68.63
ALL	**Reasoning LLMs**							
	OpenAI-o1	27.89/28.08	6.19/6.22	16.03/15.86	1.72/1.64	32.89/33.76	84.95/84.74	65.13/65.34
	DeepSeek-R1	24.16/24.70	5.84/5.71	14.33/14.32	1.51/1.44	31.37/32.40	83.83/83.61	63.24/63.33
	QwQ-32B	20.29/21.04	5.22/5.20	12.32/12.43	1.20/1.15	26.80/28.06	82.74/82.61	63.07/63.13
	Non-Reasoning LLMs							
	GPT-4o	**29.84/30.40**	8.63/8.60	**18.39/18.34**	**3.05/2.89**	36.16/37.26	**86.04/85.92**	65.80/66.32
	DeepSeek-V3	27.12/27.57	8.53/8.50	17.07/17.05	2.71/2.57	34.25/35.27	85.51/85.38	**67.00/67.41**
	Qwen2.5-32B	29.03/29.76	**8.87/8.80**	17.89/18.00	2.96/2.85	**36.56/37.64**	85.92/85.87	66.71/67.19

on query-driven summarization, but also when the evaluation involves a mix of query-based and generic summary objectives.

In contrast, reasoning LLMs do not deliver improvements in long-context, query-oriented summarization. Their performance is consistently lower than that of their base versions, regardless of the metric or specific meeting context. This pattern suggests that the additional instruction tuning for step-by-step reasoning does not benefit—and may even hinder—the ability of large language models to distill and organize salient information from lengthy and information-dense dialogue transcripts. One possible explanation is that such tuning encourages models to engage in unnecessary elaboration or over-reasoning, thereby reducing the focus and precision required for effective query-based summarization.

5. Analysis

In this section, we conduct further analyses to understand the intrinsic characteristics of generated summaries, assess summarization quality using LLM evaluators, examine explicit reasoning processes, and provide a detailed case study.

5.1. Intrinsic Analysis of Summarization Quality

For outputs on all four datasets, we evaluate three intrinsic summary qualities: *Length*, *compression rate* (Grusky et al., 2018), and *abstractiveness* (Bommasani and Cardie, 2020). Specifically, the compression rate measures the length

Table 5
Intrinsic quality metrics (mean ± std) of model-generated summarizations across four datasets. For CSDS, the *final summary* is evaluated.

Model	SAMSum				DialogSum			
	Length	Compression	Novelty	Coverage	Length	Compression	Novelty	Coverage
Reasoning LLMs								
OpenAI-o1	28.84±11.54	3.11±1.86	0.68±0.13	0.34±0.14	33.94±10.52	3.96±1.82	0.60±0.11	0.43±0.11
DeepSeek-R1	29.75±11.92	2.97±1.63	0.70±0.12	0.32±0.13	30.36±10.38	4.14±1.96	0.62±0.15	0.37±0.13
QwQ-32B	27.39±14.46	3.39±2.44	0.69±0.13	0.33±0.13	29.80±21.12	4.88±2.78	0.66±0.12	0.37±0.13
Non-Reasoning LLMs								
GPT-4o	27.95±12.07	3.19±1.73	0.65±0.13	0.37±0.14	31.24±9.86	4.32±2.00	0.59±0.12	0.44±0.12
DeepSeek-V3	28.21±11.99	3.18±1.78	0.66±0.13	0.36±0.14	29.97±9.61	4.51±2.07	0.58±0.11	0.44±0.12
Qwen2.5-32B	30.47±13.07	2.95±1.64	0.67±0.13	0.35±0.14	31.09±10.17	4.42±2.17	0.61±0.12	0.42±0.12
Reference summary	20.02±10.65	5.05±3.73	0.60±0.16	0.41±0.16	18.75±8.65	7.26±2.36	0.56±0.16	0.46±0.16

Model	CSDS				QMSum			
	Length	Compression	Novelty	Coverage	Length	Compression	Novelty	Coverage
Reasoning LLMs								
OpenAI-o1	39.11±12.10	12.92±5.10	0.48±0.10	0.57±0.10	194.78±137.90	72.76±53.13	0.48±0.11	0.61±0.10
DeepSeek-R1	38.03±12.09	13.02±5.53	0.49±0.11	0.54±0.12	206.58±81.91	53.86±35.88	0.59±0.11	0.52±0.10
QwQ-32B	42.08±13.10	11.83±4.31	0.52±0.10	0.52±0.10	312.45±123.44	35.67±24.68	0.61±0.10	0.52±0.10
Non-Reasoning LLMs								
GPT-4o	36.12±10.33	13.70±5.52	0.45±0.11	0.57±0.11	161.80±85.78	84.86±94.12	0.41±0.13	0.69±0.10
DeepSeek-V3	34.46±10.48	14.62±5.89	0.42±0.11	0.61±0.10	225.15±111.68	53.39±37.41	0.45±0.12	0.68±0.09
Qwen2.5-32B	37.89±11.69	13.26±5.07	0.46±0.10	0.57±0.10	176.19±77.60	65.28±42.57	0.47±0.13	0.66±0.10
Reference summary	46.57±27.24	12.67±6.68	0.38±0.13	0.60±0.12	64.72±29.22	184.13±138.82	0.27±0.12	0.77±0.10

difference between the input dialogue and the summary. To assess abstractiveness, we consider two complementary aspects: *Extractive Fragment Coverage* (Grusky et al., 2018), which quantifies the proportion of the summary that can be directly "copied" from the input text, and *Novelty* (Liu et al., 2023), which calculates the ratio of words in the summary that do not appear in the input dialogue.

Compression rate is computed as the ratio of the number of words in the input dialogue D to those in the summary Y:

$$\text{Compression}(D, Y) = \frac{|D|}{|Y|}. \tag{13}$$

Extractive fragment coverage measures the extent to which the summary can be directly copied from the input, defined as the proportion of summary tokens that belong to extractive fragments shared with the input text:

$$\text{Coverage}(D, Y) = \frac{1}{|Y|} \sum_{f \in F(D,Y)} |f|, \tag{14}$$

where $F(D, S)$ is the set of word sequences that are common between D and S.

Novelty quantifies the degree of abstraction by calculating the proportion of summary tokens that do not appear in the input:

$$\text{Novelty}(D, Y) = 1 - \frac{|D \cap Y|}{|Y|}, \tag{15}$$

where $|D \cap Y|$ denotes the number of words shared between the summary and the input.

Table 5 summarizes the intrinsic evaluation results across four datasets. A consistent observation for English datasets (SAMSum, DialogSum, QMSum) is that reasoning LLMs generally produce longer summaries, resulting in lower compression rates compared to their non-reasoning counterparts. Specifically, for SAMSum and DialogSum, reasoning LLM-generated summaries exhibit higher novelty but lower coverage scores, indicating these summaries tend to include more abstractive content at the cost of potentially reduced fidelity to the original dialogue. In contrast,

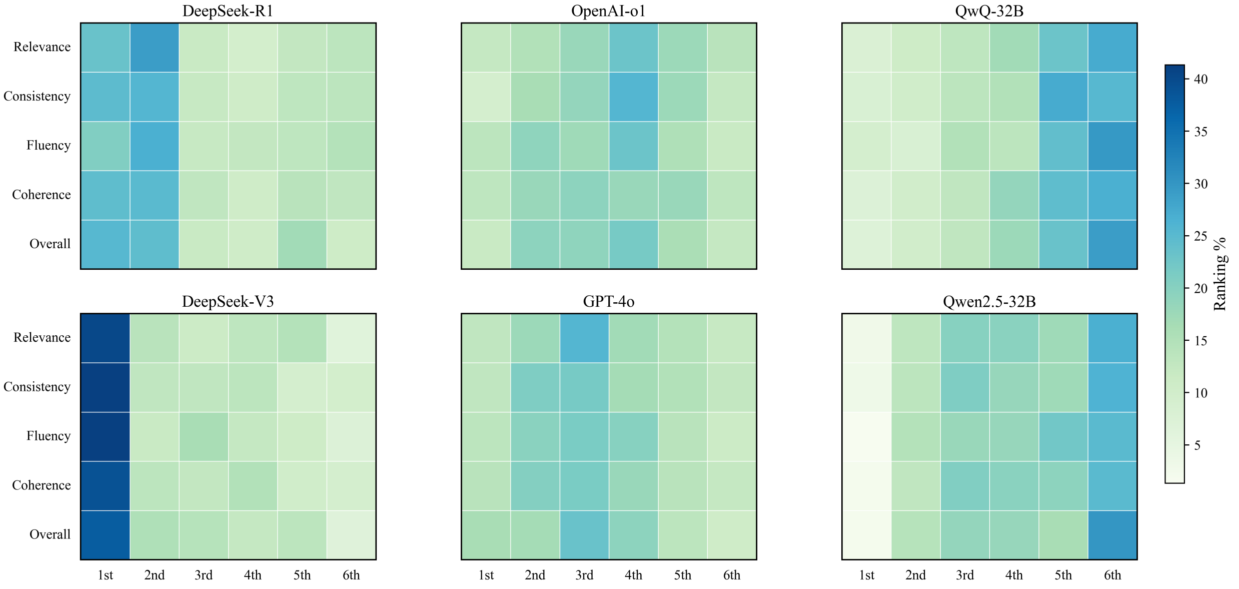


Figure 5: Distribution of Model Rankings across Evaluation Dimensions. The heatmap illustrates the percentage distribution of each model's rank (1st–6th) for all five evaluation criteria, as assessed by LLM judges.

non-reasoning LLMs generate summaries closer in length and extractiveness to reference summaries, achieving higher coverage scores, suggesting better alignment with human summarization styles.

However, the Chinese dataset (CSDS) reveals a notably different pattern. Non-reasoning LLMs generate more concise summaries (shorter length and higher compression rates) and achieve higher coverage compared to reasoning LLMs. Meanwhile, reasoning LLMs still show higher novelty, consistent with the English datasets, but this increased abstraction does not translate into improved summary quality or fidelity. Interestingly, the reference summaries for CSDS are longer and exhibit similar compression rates to model-generated summaries, reflecting that human-written summaries in this domain often contain detailed, factual descriptions rather than highly compressed abstractions.

Across all datasets, human-edited reference summaries consistently demonstrate the highest coverage, emphasizing their superior effectiveness in faithfully aggregating essential dialogue content. Moreover, reference summaries generally exhibit lower novelty, reflecting careful human editing aimed at factual accuracy and content fidelity, particularly notable in QMSum and CSDS datasets that involve longer and more complex dialogue contexts.

5.2. Comparative Assessment of Summarization Quality with LLM Evaluators

Recent studies have demonstrated that LLMs can effectively serve as evaluators for natural language generation tasks, showing strong alignment with human judgments across various evaluation dimensions (Gao et al., 2025; Murugadoss et al., 2025; Gu et al., 2025). Inspired by these findings, we complemented the computation of the summarization metrics with an LLM-based evaluation strategy to further examine the summarization quality of reasoning-oriented and non-reasoning models. Specifically, we selected three non-reasoning LLMs (DeepSeek-V3, GPT-4o, and Qwen2.5-32B) as evaluators to assess summaries generated by three reasoning-oriented models (DeepSeek-R1, OpenAI-o1, and QwQ-32B) and their corresponding non-reasoning counterparts.

We randomly sampled 100 dialogue-summary pairs from the SAMSum dataset and instructed the evaluator models to rank candidate summaries according to their alignment with the ground-truth reference summary. Specifically, the evaluators assessed summaries based on five evaluation dimensions adapted from prior human evaluation studies (Kryscinski et al., 2019). The prompt used for evaluation is detailed in Appendix A.2.

Relevance Whether the summary effectively captures the key points and central content of the original dialogue.

Consistency Whether the summary accurately reflects information presented in the dialogue without introducing factual errors or unsupported details.

Fluency Whether the summary is grammatically correct, natural, and easy to read.

Coherence Whether the summary is logically organized and clearly structured.

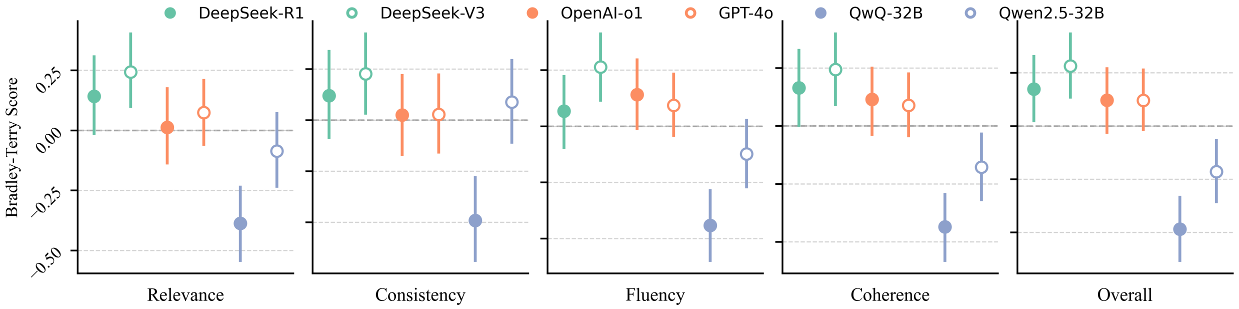


Figure 6: Bradley-Terry scores with 95% confidence intervals across five evaluation dimensions.

Overall A comprehensive judgment considering all of the above aspects.

The evaluator models were explicitly instructed to rank candidate summaries from best to worst, with the flexibility to group summaries as equally good when distinctions in quality were subtle or negligible. Allowing equal rankings mitigates the bias introduced by forced distinctions and better aligns with realistic human evaluation practices. To verify the reliability and consistency of the LLM-based evaluation results, we measured the inter-annotator agreement across evaluator models using Krippendorff's alpha (Hayes and and Krippendorff, 2007). The resulting alpha coefficients indicated satisfactory levels of agreement for all five evaluation dimensions, specifically: *Relevance* ($\alpha = 0.553$), *Consistency* ($\alpha = 0.562$), *Fluency* ($\alpha = 0.554$), *Coherence* ($\alpha = 0.583$), and *Overall* ($\alpha = 0.562$).

Figure 5 presents the distribution of rankings for each model across the five evaluation dimensions. From the heatmap, it is clearly observed that among all evaluated models, DeepSeek-V3 consistently achieved the highest proportion of top rankings, particularly evident in the dimensions of *Relevance*, *Consistency*, and *Overall*. Additionally, when directly comparing reasoning-oriented models with their non-reasoning counterparts, the non-reasoning models consistently demonstrated superior performance across all evaluation dimensions, underscoring that explicit reasoning optimization does not necessarily translate into improved summarization quality for general dialogues.

To enable rigorous statistical comparison across models based on the collected ranking data, we further analyzed the results using the Bradley–Terry (BT) model (Bradley and Terry, 1952). The BT model is specifically designed to estimate latent quality scores from pairwise preferences derived from ranking assignments, providing a principled way to summarize overall model performance across evaluators and samples.

Concretely, each ranking of candidate summaries provided by the evaluator models can be interpreted as a sequence of pairwise wins and losses among models. For any two models i and j, if i is ranked higher than j, this is treated as a pairwise win for i. The BT model then defines the probability of model i being preferred over model j as:

$$P(i \succ j) = \frac{\exp(s_i)}{\exp(s_i) + \exp(s_j)}, \tag{16}$$

where s_i and s_j denote the latent (log-scale) quality scores for models i and j, respectively. The set of scores $\{s_i\}$ is estimated by maximizing the likelihood of all observed pairwise outcomes derived from the evaluator rankings. To further account for sampling variability and to obtain uncertainty estimates, we employed 1,000 bootstrap resamplings when fitting the BT model, thereby reporting the mean and 95% confidence intervals for each model's score in every evaluation dimension.

Figure 6 presents the Bradley–Terry scores across five evaluation dimensions: relevance, consistency, fluency, coherence, and overall summary quality. The results show clear trends aligning well with our earlier automatic evaluation findings. Specifically, DeepSeek-V3 consistently outperforms its reasoning-oriented counterpart DeepSeek-R1 across all evaluation dimensions, demonstrating strong capabilities in generating summaries that are more relevant, consistent, fluent, and coherent. Similarly, Qwen2.5-32B achieves notably higher scores compared to its reasoning counterpart QwQ-32B, indicating that reasoning-specific training did not yield improvements in summary quality for this model family. Regarding OpenAI-o1 and GPT-4o, the performance differences are subtler: GPT-4o slightly surpasses OpenAI-o1 on relevance but exhibits comparable performance on consistency, fluency, coherence, and overall quality dimensions. Notably, both OpenAI-o1 and GPT-4o generally outperform Qwen-family models (Qwen2.5-32B and QwQ-32B) across most metrics, except for fluency, where differences are minimal.

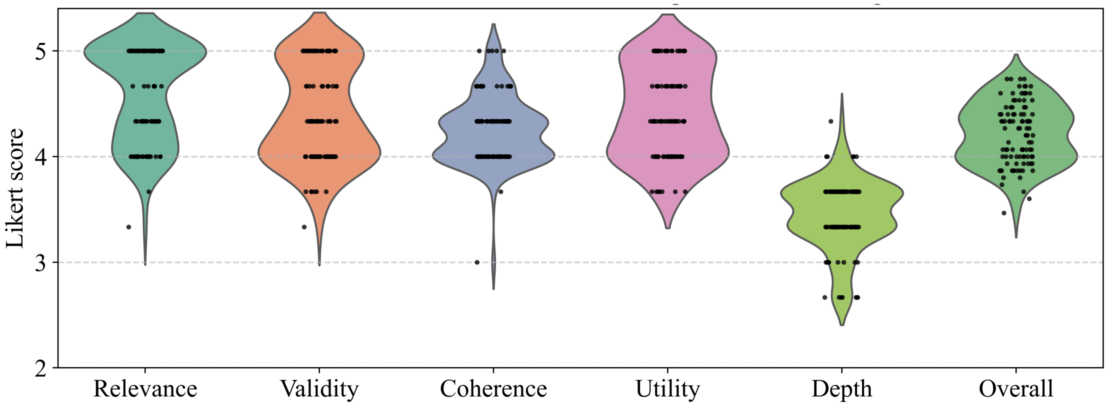


Figure 7: Distribution of *DeepSeek-R1* Reasoning Processes across Evaluation Dimensions.

To better understand the relationship between automatic and LLM-based evaluation, we also report the correlation between these metrics in Appendix B.1. This analysis reveals that the correlation between automatic metrics and LLM-based evaluation remains generally low, consistent with previous findings in other NLG tasks (Reiter, 2018; Elangovan et al., 2025).

5.3. In-depth Analysis of Reasoning Processes

To gain deeper insight into model reasoning behavior, we conducted a detailed analysis of the explicit reasoning processes produced by DeepSeek-R1 for the same set of 100 randomly sampled dialogue instances from the SAMSum dataset described above. Our evaluation framework draws on recent advances in the assessment of step-by-step reasoning traces (Lee and Hockenmaier, 2025), and we extend this framework by introducing an additional criterion, *depth*, which specifically evaluates the presence of abstraction, synthesis, and integration within the reasoning process. Prior work has demonstrated that explicit inductive and summarization steps during the reasoning process can enhance the conciseness and informativeness of generated summaries (Sun et al., 2024). Each dimension in our evaluation is designed to reflect a distinct functional property of reasoning quality:

Relevance Assesses whether each reasoning step is directly grounded in the source dialogue, avoiding hallucinated, irrelevant, or unsupported content.

Validity Measures the logical correctness of individual reasoning steps, ensuring that every inference is entailed by either the dialogue or previous steps, without contradictions or faulty logic.

Coherence Evaluates the semantic and structural connectivity of the reasoning chain, examining whether all steps are linked smoothly and the overall chain progresses without abrupt jumps, redundancy, or gaps.

Utility Quantifies the informativeness and necessity of each reasoning step, determining whether each contributes meaningfully toward constructing a correct summary or answer, rather than being redundant or trivial.

Depth Examines whether the reasoning chain demonstrates abstraction, integration, or higher-level inference—moving beyond surface-level repetition or paraphrase to capture more complex relationships within the dialogue.

For each evaluation dimension, we designed a prompt that instructed the evaluator to assign an integer score from 1 (very poor) to 5 (excellent), based solely on the specified aspect of the reasoning process. The prompt used for evaluation is detailed in Appendix A.1. The scoring was conducted independently by three advanced LLMs—DeepSeek-V3, GPT-4o, and GPT-4.1—serving as automatic judges. This setup enables a granular assessment of the strengths and weaknesses of model-generated reasoning, revealing how well each approach performs across multiple facets of process-level reasoning in dialogue summarization tasks.

Figure 7 shows the distribution of scores assigned by LLM evaluators across the five specific evaluation dimensions (*Relevance*, *Validity*, *Coherence*, *Utility*, and *Depth*) as well as an aggregated *Overall* dimension, for the reasoning processes generated by *DeepSeek-R1*. We observe that the model achieves generally high ratings in *Relevance* and *Validity*, with most scores clustering around 4 or 5, indicating strong factual grounding and logical correctness in reasoning steps. Conversely, scores for *Depth* and *Utility* exhibit broader distributions and notably lower medians, suggesting variability in the model's capacity for higher-level abstraction and efficient information

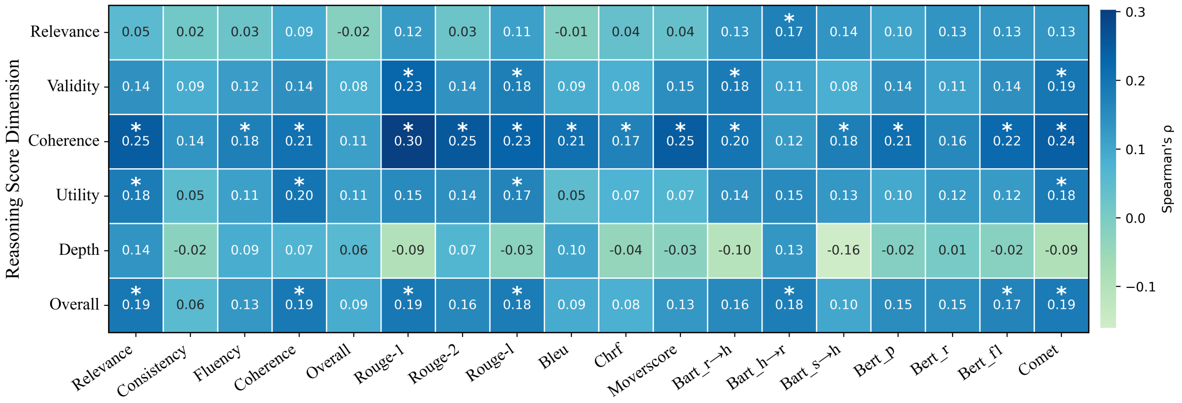


Figure 8: Spearman correlations between reasoning process evaluation dimensions and summarization quality metrics (LLMs evaluation: Relevance–Overall; automatic metrics: ROUGE-1–COMET). Asterisks (*) denote statistically significant correlations ($p < 0.05$).

synthesis. Additionally, the distribution for *Coherence* reveals moderate consistency, implying occasional difficulties in maintaining smooth logical transitions between reasoning steps. The aggregated *Overall* scores demonstrate a relatively centralized distribution, reflecting balanced performance across these dimensions but also highlighting room for improvement, particularly regarding deeper, integrative reasoning abilities.

Notably, the relatively lower scores observed in the *Depth* dimension indicate that the explicit reasoning processes generated by DeepSeek-R1 often lack sufficient abstraction, integration, and higher-level summarization capabilities. This finding suggests a potential reason behind the observed performance gap, where reasoning-oriented models fail to outperform their non-reasoning counterparts in dialogue summarization tasks. Specifically, the limited depth in reasoning processes may restrict these models from effectively synthesizing information and capturing nuanced dialogue content, ultimately constraining their summarization quality.

Figure 8 illustrates the correlation between reasoning process scores across evaluation dimensions and various summarization quality metrics. Contrary to our initial hypothesis—that higher-quality reasoning processes, particularly in dimensions like *Depth* and *Coherence*, would positively correlate with summarization performance—our analysis reveals predominantly weak and statistically insignificant correlations. This indicates a possible disconnect between explicitly assessed reasoning quality and summarization effectiveness.

5.4. Case Study

Table 6 presents a representative example comparing DeepSeek-R1 (reasoning) and DeepSeek-V3 (non-reasoning) on a dialogue summarization task. In this case, the ground truth summary is concise, directly highlighting the key facts: the accident location and the absence of fatalities. DeepSeek-V3 closely matches this ground truth, providing a short and accurate summary.

In contrast, DeepSeek-R1's generated summary is noticeably more verbose and includes details that are not central to the main event, such as identifying all participants and their emotional reactions. Examination of DeepSeek-R1's explicit reasoning process reveals several key phenomena. First, a significant portion of the reasoning steps consists of direct paraphrasing or repetition of the dialogue, rather than abstraction or synthesis. These repetitive segments introduce redundancy and increase the risk of introducing noisy or irrelevant information. Second, the reasoning process introduces hallucinated or factually inconsistent content (e.g., "I tried to get to the Circle Mall" is highlighted as a possible hallucination when not directly corroborated by the context), which is subsequently propagated to the generated summary.

Additionally, the stepwise reasoning process attempts to consider and weigh multiple aspects and reactions in the dialogue, but this breadth of consideration comes at the expense of clarity and conciseness. Instead of distilling the dialogue down to its most salient facts, the model's reasoning leads to summaries that are overly detailed and less focused. This example demonstrates that, in the context of dialogue summarization, the inclusion of explicit multi-step reasoning does not inherently improve—and may even harm—summary quality, by increasing both verbosity and the risk of factual errors or inconsistencies.

Table 6
Case study comparing DeepSeek-R1 and DeepSeek-V3 on a representative dialogue summarization example. In the reasoning process column, red text indicates templated or redundant sentences, yellow highlights mark hallucinated or factually inconsistent content, and blue text identifies the model's actual abstraction or synthesis steps within the reasoning process.

Source	Content
Dialogue	Oli: Theres a car accident Katie: Where? Oli: On circle drive Oli: I tried to get to the Circle Mall Katie: Oh no Pavel: Its on the news now Pavel: Theres no deaths Katie: Thank god *(angel emoji)*
Ground Truth	There has been an accident on Circle Drive, near Circle Mall. There are no fatalities.
DeepSeek-V3	There was a car accident on Circle Drive, but thankfully there were no deaths.
DeepSeek-R1	Oli informs Katie and Pavel about a car accident on Circle Drive near Circle Mall. Pavel confirms it's on the news with no fatalities, and Katie expresses relief.
DeepSeek-R1 Reasoning Process	Okay, let me try to figure out how to summarize this dialogue. First, I need to read through the conversation carefully to understand the main points. So, Oli starts by saying there's a car accident. Katie asks where, and Oli responds it's on Circle Drive, and he tried to get to Circle Mall. Then Pavel mentions it's on the news now and that there are no deaths. Katie is relieved and says "Thank god" with an emoji. Looking at the examples given, summaries capture the key actions and outcomes without extra details. For instance, in the previous summaries, they mention who is involved and the main event or conclusion. Here, the main points are: a car accident occurred on Circle Drive, near Circle Mall. Pavel reports it's on the news with no fatalities, and Katie is relieved.The key elements are the accident location, no deaths, and the reactions. I should make sure the summary includes the accident location, that there were no deaths, and maybe Katie's relief. Also, Oli tried to go to the mall, which might explain why he's mentioning the accident. But the main focus is the accident itself and the outcome. So the summary should be something like: Oli reports a car accident on Circle Drive near Circle Mall. Pavel says there are no deaths, and Katie is relieved.

6. Conclusion

In this work, we present the first systematic, scenario-driven comparison of reasoning LLMs and their conventional counterparts for dialogue summarization. By evaluating strong models—including OpenAI-o1, DeepSeek-R1, QwQ-32B, and matched base models—across generic, role-oriented, and query-oriented paradigms, we provide a detailed assessment of current capabilities and limitations.

Our results demonstrate that, contrary to the gains observed in other reasoning-intensive domains (Chen et al., 2025a), explicit stepwise reasoning does not necessarily confer an advantage for dialogue summarization. Instead, reasoning models often fall short of strong base LLMs like GPT-4o, DeepSeek-V3, and Qwen2.5-32B, with common issues including verbosity, factual inconsistencies, and loss of conciseness—especially in tasks that require high-level abstraction.

Through a multi-perspective evaluation combining automatic metrics, LLM-based judgment, and scenario-specific error analysis, we further reveal that current benchmarks and evaluators still struggle to fully capture the challenges of dialogue summarization. Although LLM-based evaluation offers better alignment with human judgment than traditional metrics, it remains difficult to robustly assess factuality, conciseness, and pragmatic adequacy.

Our analysis underscores the importance of developing more nuanced modeling and evaluation strategies that effectively balance stepwise reasoning, abstraction, and faithfulness in dialogue summarization. We believe that our comprehensive evaluation and findings will provide useful guidance for future research and help inform the design and assessment of LLMs in real-world dialogue summarization tasks.

References

Block, J., Chen, Y.P., Budharapu, A., Anthony, L., Dorr, B., 2023. Summary Cycles: Exploring the Impact of Prompt Engineering on Large Language Models' Interaction with Interaction Log Information, in: Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Association for Computational Linguistics, Bali, Indonesia. pp. 85–99. doi:10.18653/v1/2023.eval4nlp-1.7.

Bommasani, R., Cardie, C., 2020. Intrinsic Evaluation of Summarization Datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 8075–8096. doi:10.18653/v1/2020.emnlp-main.649.

Bradley, R.A., Terry, M.E., 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39, 324–345. doi:10.2307/2334029, arXiv:2334029.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-Shot Learners. doi:10.48550/arXiv.2005.14165. arXiv:2005.14165 [cs].

Chen, A., Song, Y., Zhu, W., Chen, K., Yang, M., Zhao, T., zhang, M., 2025a. Evaluating o1-Like LLMs: Unlocking Reasoning for Translation through Comprehensive Analysis. doi:10.48550/arXiv.2502.11544.

Chen, J., Wei, Z., Ren, Z., Li, Z., Zhang, J., 2025b. LR\$^2\$Bench: Evaluating Long-chain Reflective Reasoning Capabilities of Large Language Models via Constraint Satisfaction Problems. doi:10.48550/arXiv.2502.17848.

Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., Song, L., Liu, Q., Zhou, M., Zhang, Z., Wang, R., Tu, Z., Mi, H., Yu, D., 2025c. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. doi:10.48550/arXiv.2412.21187.

Chen, Y., Liu, Y., Chen, L., Zhang, Y., 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online. pp. 5062–5074. doi:10.18653/v1/2021.findings-acl.449.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z., 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. doi:10.48550/arXiv.2501.12948.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., Sui, Z., 2024. A Survey on In-context Learning, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 1107–1128. doi:10.18653/v1/2024.emnlp-main.64.

Elangovan, A., Xu, L., Ko, J., Elyasi, M., Liu, L., Bodapati, S.B., Roth, D., 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge, in: The Thirteenth International Conference on Learning Representations. URL: https://openreview.net/forum?id=E8gYIrbP00.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., Wan, X., 2025. LLM-based NLG Evaluation: Current Status and Challenges. Computational Linguistics , 1–27doi:10.1162/coli_a_00561.

Gao, M., Wan, X., 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States. doi:10.18653/v1/2022.naacl-main.418.

Gliwa, B., Mochol, I., Biesek, M., Wawer, A., 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Hong Kong, China. pp. 70–79. doi:10.18653/v1/D19-5409.

Grusky, M., Naaman, M., Artzi, Y., 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana. pp. 708–719. doi:10.18653/v1/N18-1065.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., Guo, J., 2025. A Survey on LLM-as-a-Judge. doi:10.48550/arXiv.2411.15594, arXiv:2411.15594.

Hashemi, M., Bamgbose, O., Madhusudhan, S.T., Nair, J.S., Tiwari, A., Yadav, V., 2025. DNR Bench: Benchmarking Over-Reasoning in Reasoning LLMs. doi:10.48550/arXiv.2503.15793.

Hayes, A.F., and Krippendorff, K., 2007. Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures 1, 77–89. doi:10.1080/19312450709336664.

Jin, K., Wang, Y., Yang, X., Im, S.K., 2025. HiSum: Hierarchical Topic-Driven Approach for Role-Oriented Dialogue Summarisation. Expert Systems 42, e70043. doi:10.1111/exsy.70043.

Kryscinski, W., Keskar, N.S., McCann, B., Xiong, C., Socher, R., 2019. Neural Text Summarization: A Critical Evaluation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics. pp. 540–551. doi:10.18653/v1/D19-1051.

Laban, P., Kryscinski, W., Agarwal, D., Fabbri, A., Xiong, C., Joty, S., Wu, C.S., 2023. SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 9662–9676. doi:10.18653/v1/2023.emnlp-main.600.

Lee, J., Hockenmaier, J., 2025. Evaluating Step-by-step Reasoning Traces: A Survey. doi:10.48550/arXiv.2502.12289, arXiv:2502.12289.

Lee, Y., Yun, T., Cai, J., Su, H., Song, H., 2024. UniSumEval: Towards Unified, Fine-Grained, Multi-Dimensional Summarization Evaluation for LLMs. doi:10.48550/arXiv.2409.19898.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. doi:10.48550/arXiv.1910.13461. arXiv:1910.13461 [cs].

Liang, X., Bian, C., Wu, S., Li, Z., 2022. Towards Modeling Role-Aware Centrality for Dialogue Summarization, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online only. pp. 43–50. doi:10.18653/v1/2022.aacl-short.6.

Lin, C.Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain. pp. 74–81.

Lin, H., Ma, L., Zhu, J., Xiang, L., Zhou, Y., Zhang, J., Zong, C., 2021. CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 4436–4451. doi:10.18653/v1/2021.emnlp-main.365.

Liu, Y., Deb, B., Teruel, M., Halfaker, A., Radev, D., Awadallah, A.H., 2023. On Improving Summarization Factual Consistency from Natural Language Feedback, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 15144–15161. doi:10.18653/v1/2023.acl-long.844.

Lu, Y.J., Hu, T.Y., Koppula, H.S., Pouransari, H., Chang, J.H.R., Xia, Y., Kong, X., Zhu, Q., Wang, X.S., Tuzel, O., Vemulapalli, R., 2025. Mutual Reinforcement of LLM Dialogue Synthesis and Summarization Capabilities for Few-Shot Dialogue Summarization, in: Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico. pp. 7237–7256.

Marcu, D., 1997. From discourse structures to text summaries, in: Intelligent Scalable Text Summarization.

Mishra, N., Sahu, G., Calixto, I., Abu-Hanna, A., Laradji, I., 2023. LLM aided semi-supervision for efficient Extractive Dialog Summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 10002–10009. doi:10.18653/v1/2023.findings-emnlp.670.

Murugadoss, B., Poelitz, C., Drosos, I., Le, V., McKenna, N., Negreanu, C.S., Parnin, C., Sarkar, A., 2025. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. Proceedings of the AAAI Conference on Artificial Intelligence 39, 19589–19597. doi:10.1609/aaai.v39i18.34157.

Okadome, Y., Yuguchi, A., Fukui, R., Matsumoto, Y., 2024. Prompt Design Using Past Dialogue Summarization for LLMs to Generate the Current Appropriate Dialogue, in: Artificial Neural Networks and Machine Learning – ICANN 2024, Springer Nature Switzerland, Cham. pp. 33–41. doi:10.1007/978-3-031-72356-8_3.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2001. BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, Association for Computational Linguistics, Philadelphia, Pennsylvania. p. 311. doi:10.3115/1073083.1073135.

Popović, M., 2015. chrF: Character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal. pp. 392–395. doi:10.18653/v1/W15-3049.

Ramprasad, S., Ferracane, E., Lipton, Z., 2024. Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 12549–12561. doi:10.18653/v1/2024.acl-long.677.

Rei, R., Stewart, C., Farinha, A.C., Lavie, A., 2020. COMET: A Neural Framework for MT Evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213.

Reiter, E., 2018. A Structured Review of the Validity of BLEU. Computational Linguistics 44, 393–401. doi:10.1162/coli_a_00322.

Rush, A.M., Chopra, S., Weston, J., 2015. A Neural Attention Model for Abstractive Sentence Summarization. doi:10.48550/arXiv.1509.00685. arXiv:1509.00685 [cs].

See, A., Liu, P.J., Manning, C.D., 2017. Get To The Point: Summarization with Pointer-Generator Networks. doi:10.48550/arXiv.1704.04368. arXiv:1704.04368 [cs].

Song, H., Su, H., Shalyminov, I., Cai, J., Mansour, S., 2024. FineSurE: Fine-grained Summarization Evaluation using LLMs. doi:10.48550/arXiv.2407.00908.

Sun, S., Yuan, R., Cao, Z., Li, W., Liu, P., 2024. Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics. pp. 7551–7558. doi:10.18653/v1/2024.findings-acl.449.

Tang, L., Shalyminov, I., Wong, A., Burnsky, J., Vincent, J., Yang, Y., Singh, S., Feng, S., Song, H., Su, H., Sun, L., Zhang, Y., Mansour, S., McKeown, K., 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 4455–4480. doi:10.18653/v1/2024.naacl-long.251.

Tang, Y., Puduppully, R., Liu, Z., Chen, N., 2023. In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis, in: Proceedings of the 4th New Frontiers in Summarization Workshop, Association for Computational Linguistics, Singapore. pp. 56–67. doi:10.18653/v1/2023.newsum-1.6.

Tian, Y., Xia, F., Song, Y., 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), Association for Computational Linguistics, Bangkok, Thailand. pp. 7143–7155. doi:10.18653/v1/2024.acl-long.385.

Valmeekam, K., Stechly, K., Gundawar, A., Kambhampati, S., 2025. A Systematic Evaluation of the Planning and Scheduling Abilities of the Reasoning Model o1. Transactions on Machine Learning Research 2025.

Wang, J., Liang, Y., Meng, F., Zou, B., Li, Z., Qu, J., Zhou, J., 2023. Zero-Shot Cross-Lingual Summarization via Large Language Models, in: Proceedings of the 4th New Frontiers in Summarization Workshop, Association for Computational Linguistics, Singapore. pp. 12–23. doi:10.18653/v1/2023.newsum-1.2.

Wang, Q., Fu, Y., Cao, Y., Wang, S., Tian, Z., Ding, L., 2025. Recursively summarizing enables long-term dialogue memory in large language models. Neurocomputing 639, 130193. doi:10.1016/j.neucom.2025.130193.

Xiao, J., Chen, Y., Ou, Y., Yu, H., Shu, K., Xiao, Y., 2024. Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for Dialogue Summarization, in: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi:10.1109/IJCNN60899.2024.10650513.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z., 2025. Qwen3 Technical Report. doi:10.48550/arXiv.2505.09388.

Ye, Y., Fu, B., Huang, C., Chen, Y., Shi, X., 2025. How Well Do Large Reasoning Models Translate? A Comprehensive Evaluation for Multi-Domain Machine Translation. doi:10.48550/arXiv.2505.19987.

Yuan, W., Neubig, G., Liu, P., 2021. BARTScore: Evaluating Generated Text as Text Generation, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 27263–27277.

Zeng, Z., Cheng, Q., Yin, Z., Zhou, Y., Qiu, X., 2025. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? doi:10.48550/arXiv.2502.12215.

Zhang, H., Liu, X., Zhang, J., 2023. Extractive Summarization via ChatGPT for Faithful Summary Generation, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 3270–3278. doi:10.18653/v1/2023.findings-emnlp.214.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y., 2019. BERTScore: Evaluating Text Generation with BERT, in: International Conference on Learning Representations.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S., 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 563–578. doi:10.18653/v1/D19-1053.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A.H., Celikyilmaz, A., Liu, Y., Qiu, X., Radev, D., 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 5905–5921. doi:10.18653/v1/2021.naacl-main.472.

Zhou, W., Li, G., Cheng, X., Liang, X., Zhu, J., Zhai, F., Li, Z., 2023. Multi-Stage Pre-training Enhanced by ChatGPT for Multi-Scenario Multi-Domain Dialogue Summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 6893–6908. doi:10.18653/v1/2023.findings-emnlp.460.

Zhou, Y., Ringeval, F., Portet, F., 2025. Can GPT models Follow Human Summarization Guidelines? A Study for Targeted Communication Goals. doi:10.48550/arXiv.2310.16810.

Zhu, R., Lau, J.H., Qi, J., 2025. Factual Dialogue Summarization via Learning from Large Language Models, in: Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE. pp. 4474–4492.

A. Prompt Template

Figure A.1: Prompt Template for LLM-based Reasoning Process Scoring.

Prompt Template for LLM-based Reasoning Process Scoring

Instruction:Your task is to strict assign a score from 1 to 5 for the reasoning chain based only on the following criterion:
Evaluation Criterion {criterion}

Scoring Instructions
1 = Very poor (contains major errors, irrelevant, or unsupported information, or fails to address the task at all)
3 = Moderate (partially meets the criterion but contains some errors, missing information, or only partly relevant/useful)
5 = Excellent (fully meets the criterion, no errors, perfectly relevant and logical)

Output ONLY the score (an integer between 1 and 5). No explanation.

Dialogue {dialogue}

Reasoning Process {reasoning}

Figure A.2: Prompt Template for LLM-based Ranking Evaluation.

Prompt Template for LLM-based Ranking Evaluation

Instruction:You are an expert evaluator for dialogue summarization. You will see a dialogue, the reference summary (the ground truth written by humans), and several candidate summaries generated by different models.

Your task is to rank all candidate summaries from best to worst STRICTLY based on their quality compared to the reference summary, only considering the following criterion:

Evaluation Criterion {criterion}

When ranking, please judge each candidate summary ONLY by how well it matches the content and quality of the reference summary (not your own preference), based on the above criterion.

Output ONLY the final ranking in the exact format as below, and nothing else:
Ranking: [model_key_1] > [model_key_2] = [model_key_3] > [model_key_4] > ...
(Use '>' for better, '=' for tied, all keys must be included. No extra explanation.)

Dialogue {dialogue}

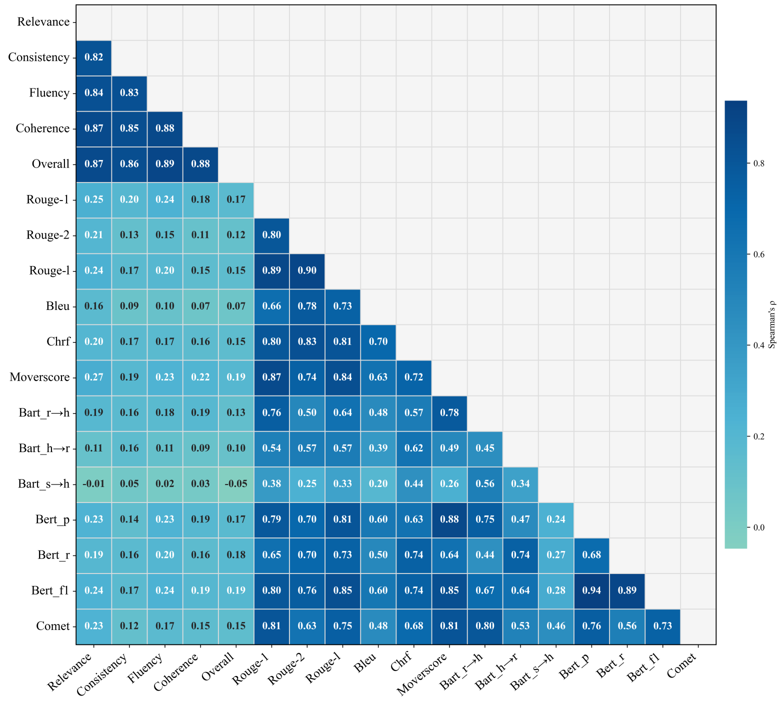
Reference Summary {reference}

Candidate Summary {summary}

B. Correlations between automatic evaluation and LLMs evaluation

Figure B.1 shows that the correlations between automatic evaluation metrics and LLM-based evaluation in dialogue summarization are generally low. This observation is consistent with previous studies in other NLG tasks (Reiter, 2018; Elangovan et al., 2025), Together, these findings highlight the limitations of automatic metrics for evaluating summary quality, reinforcing the need for more human-aligned assessment approaches.

Figure B.1: The correlation (Pearson's r) between different automatic evaluation metrics and LLMs evaluation.



	Relevance	Consistency	Fluency	Coherence	Overall	Rouge-1	Rouge-2	Rouge-l	Bleu	Chrf	Moverscore	Bart_r→h	Bart_h→r	Bart_s→h	Bert_p	Bert_r	Bert_f1	Comet
Relevance																		
Consistency	0.82																	
Fluency	0.84	0.83																
Coherence	0.87	0.85	0.88															
Overall	0.87	0.86	0.89	0.88														
Rouge-1	0.25	0.20	0.24	0.18	0.17													
Rouge-2	0.21	0.13	0.15	0.11	0.12	0.80												
Rouge-l	0.24	0.17	0.20	0.15	0.15	0.89	0.90											
Bleu	0.16	0.09	0.10	0.07	0.07	0.66	0.78	0.73										
Chrf	0.20	0.17	0.17	0.16	0.15	0.80	0.83	0.81	0.70									
Moverscore	0.27	0.19	0.23	0.22	0.19	0.87	0.74	0.84	0.63	0.72								
Bart_r→h	0.19	0.16	0.18	0.19	0.13	0.76	0.50	0.64	0.48	0.57	0.78							
Bart_h→r	0.11	0.16	0.11	0.09	0.10	0.54	0.57	0.57	0.39	0.62	0.49	0.45						
Bart_s→h	-0.01	0.05	0.02	0.03	-0.05	0.38	0.25	0.33	0.20	0.44	0.26	0.56	0.34					
Bert_p	0.23	0.14	0.23	0.19	0.17	0.79	0.70	0.81	0.60	0.63	0.88	0.75	0.47	0.24				
Bert_r	0.19	0.16	0.20	0.16	0.18	0.65	0.70	0.73	0.50	0.74	0.64	0.44	0.74	0.27	0.68			
Bert_f1	0.24	0.17	0.24	0.19	0.19	0.80	0.76	0.85	0.60	0.74	0.85	0.67	0.64	0.28	0.94	0.89		
Comet	0.23	0.12	0.17	0.15	0.15	0.81	0.63	0.75	0.48	0.68	0.81	0.80	0.53	0.46	0.76	0.56	0.73	