A Large-Scale Multi-Dimensional Empirical Study of LLMs for Conversation Summarization

Weixiao Zhou[α]
Gengyao Li[γ] **Junnan Zhu**[β*] **Xianfu Cheng**[α] **Feifei Zhai**[βγ] **Zhoujun Li**[α*]
[α]CCSE, Beihang University [β]MAIS, CASIA [γ]Fanyu AI Laboratory
wxzhou@buaa.edu.cn junnan.zhu@nlpr.ia.ac.cn

Abstract

Despite the significant advancement of LLMs in conversation summarization, their evaluation remains limited by insufficient scenarios, input lengths, and sample sizes. Furthermore, existing benchmarks often omit frontier reasoning systems and efficient small models, or lack fine-grained, multi-dimensional assessments. To bridge these gaps, we propose OMNICSEVAL, a unified benchmark comprising 1,800 diverse conversations across six real-world scenarios, featuring context lengths ranging from 128 to 32k tokens. For fine-grained evaluation, we employ a bidirectional fact-checking framework that integrates key fact matching to assess completeness and conciseness, alongside summary fact verification to evaluate faithfulness. To ensure reliable assessment, we establish a human-LLM collaborative pipeline for key fact extraction and a multi-LLM consensus verifier for summary fact decomposition. Leveraging this framework, we evaluate 28 LLMs across four distinct categories grouped by reasoning capability and model scale. Our extensive empirical study reveals critical insights regarding the cross-scenario challenges current LLMs continue to face, the impacts of reasoning and scale, and the efficiency and adaptability of reasoning models. We also provide guidance for system selection in real-world deployments[1].

1 Introduction

Conversation summarization aims to distill key information from a dialogue into a faithful summary (Kirstein et al., 2025). The advancement of LLMs has significantly improved this field (Zhou et al., 2023; Zhu et al., 2025). However, existing evaluation studies remain constrained by narrow conversation scenarios (Zhu et al., 2023; Gao and Wan, 2022), insufficient sample sizes (Lee et al., 2024; Min et al., 2025), and context length restrictions (Laban et al., 2023; Tang et al., 2024b). In Table 1, we present a comprehensive comparison of current conversation summarization benchmarks.

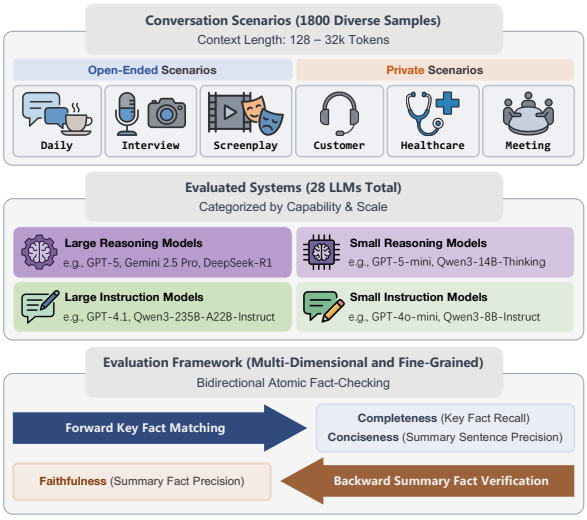

Figure 1: Core components of OMNICSEVAL. We scale the diversity of conversations and evaluate LLMs with a fine-grained and interpretable evaluation framework.

Moreover, reasoning language models (RLMs) demonstrate exceptional capabilities in complex logical tasks by scaling inference compute (Li et al., 2025; Chen et al., 2025a), while small language models[2] (SLMs) (Gupta et al., 2026; Wang et al., 2025) are increasingly vital for cost-effective applications. Nevertheless, the impacts of explicit reasoning and model scale on conversation summarization remain underexplored. While recent works (Jin et al., 2026; Yuan and Zhang, 2026; Fu et al., 2024b; Xu et al., 2025) offer preliminary investigations into these types of models, they lack fine-grained evaluation on multiple dimensions (Song et al., 2024a), relying heavily on coarse similarity-based metrics (Lin, 2004; Zhang et al., 2020).

To address these limitations, we introduce **OMNICSEVAL** (Omni Conversation Summarization Evaluation), aiming to: (1) establish a *comprehensive* and *trustworthy* benchmark for diverse LLM categories in conversation summarization; (2) in-

*Corresponding Author
[1]https://github.com/zhouweixiao/OmniCSEval.

[2]In this study, *small* models are defined as open-weight LLMs with fewer than 15B parameters (e.g., Qwen3-8B, -14B) or lightweight variants of proprietary LLMs (e.g., GPT-5-nano, -mini). See Table 8 (Appendix C) for our detailed taxonomy.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 15690–15710
October 24-29, 2026 ©2026 Association for Computational Linguistics

Benchmark	Dialogue Scenarios	Context Range	Dialogue Samples	Evaluated LLMs	Reasoning LLMs	Small LLMs	Evaluation Dimensions	Evaluation Granularity	Scoring Method
DialSummEval	1 (1 Source)	24 – 492	100	No	No	No	Multiple	Summary	Likert
DiaSumFact	2 (2 Sources)	23 – 599	120	No	No	No	Single	Sentence	Percentage
TofuEval	2 (2 Sources)	746 – 1.3k	100	5 Models	No	3 Models	Multiple	Sentence	Percentage
SummEdits	5 (7 Sources)	41 – 1.8k	280	10 Models	No	1 Model	Single	Summary	Ternary
UniSumEval	4 (4 Sources)	65 – 3.5k	100	7 Models	No	3 Models	Multiple	Fact & Sentence	Percentage
MSumBench	3 (3 Sources)	187 – 5.9k	75	6 Models	No	No	Multiple	Fact & Sentence	Percentage
OMNICSEVAL	**6 (13 Sources)**	**128 – 32k**	**1,800**	**28 Models**	**14 Models**	**15 Models**	**Multiple**	**Bidirectional Fact**	**Percentage**

Table 1: Comparison of our OMNICSEVAL with existing conversation summarization benchmarks.

vestigate the specific *challenges* LLMs face across scenarios and the impacts of reasoning and model scale; and (3) analyze the efficiency of reasoning models in translating thinking length into performance and their adaptability to scenario complexity. Figure 1 presents an overview of our benchmark.

To ensure data breadth, we source 1,800 conversations from 13 high-quality dialogue summarization corpora across six major real-world scenarios: *Daily Life*, *Media Interview*, *Screenplay*, *Customer Service*, *Healthcare*, and *Meeting*. These scenarios span diverse structural complexities, ranging from brief dyadic exchanges in Daily Life to extensive multi-party discussions in Meeting, yielding context lengths that extend from 128 to 32k tokens. Using this dataset, we evaluate 50,400 summaries generated by 28 widely-adopted LLMs, categorized by reasoning capability and scale into four representative groups: *Large Reasoning*, *Small Reasoning*, *Large Instruction*[3], and *Small Instruction* models.

Our evaluation focuses on three core dimensions: *Completeness*, *Conciseness*, and *Faithfulness*. To achieve a fine-grained and interpretable assessment, we employ an automated bidirectional atomic fact-checking framework[4]. This integrates the complementary processes of key fact matching (Song et al., 2024a) and summary fact verification (Wan et al., 2024). Completeness is measured as the recall of aligned key facts, conciseness is quantified by the density of fact-matched summary sentences, and faithfulness as the precision of supported summary facts. To guarantee reliable evaluation, we implement a rigorous three-stage human-LLM collaborative pipeline to extract gold-standard key facts and a robust multi-LLM consensus filtering mechanism for summary fact decomposition. Furthermore, we isolate the LLMs utilized in the evaluation pipeline from the evaluated ones to minimize potential bias (Zhao et al., 2024). We also demonstrate our evaluation's stability, judge-agnostic robustness, and strong alignment with human judgments (see §5).

Through extensive empirical analysis, we reveal several critical findings:

- LLMs face distinct challenges across scenarios. For instance, they exhibit profound vulnerabilities in completeness and conciseness within Meeting and Screenplay, while struggling with faithfulness in Daily Life.
- Reasoning is essential for improving completeness and conciseness, but it also introduces an increased risk of hallucinations.
- While explicit reasoning can effectively offset parameter deficiencies, model scale remains the decisive factor for multi-dimensional performance and cross-scenario stability.
- Models with longer reasoning processes do not necessarily achieve better performance, particularly in completeness and faithfulness.
- A notable number of reasoning models fail to dynamically scale their computational budget to match scenario complexity, exhibiting rigid or even inverse trends in thinking lengths.

2 Related Work

Evaluating LLMs for Dialogue Summarization. Despite recent efforts to evaluate LLM-based conversation summarization, existing studies are still bottlenecked by constrained scenarios (Tang et al., 2023, 2024b), conversation lengths (Laban et al., 2023; Zhou et al., 2025), evaluation scales (Lee et al., 2024; Min et al., 2025) and dimensions (Ramprasad et al., 2024; Bao et al., 2025), potentially limiting the generalizability and robustness of their results. For instance, TofuEval (Tang et al., 2024b) is restricted to two scenarios, UniSumEval (Lee et al., 2024) and MSumBench (Min et al., 2025) assess merely 25 instances per domain, and FaithBench (Bao et al., 2025) solely focuses on the faithfulness dimension. Furthermore, while emerging

[3]Conventionally, *instruction* models are considered equivalent to *non-reasoning* models, both referring to models that do not produce explicit intermediate reasoning processes.

[4]An atomic fact is a short, self-contained information unit conveying a single factual claim (Min et al., 2023).

literature explores the behavior of reasoning (Jin et al., 2026; Yuan and Zhang, 2026) and small language models (Fu et al., 2024b), these works still lack trustworthy evaluation protocols (Song et al., 2024a). Therefore, constructing a comprehensive and reliable benchmark is a critical priority.

Summarization Evaluation Methods. Conventional similarity-based metrics (Yuan et al., 2021; Zhang et al., 2020) correlate poorly with human judgments. NLI-based (Goyal and Durrett, 2020; Laban et al., 2022) and QA-based (Scialom et al., 2021; Fabbri et al., 2022) approaches primarily focus on the faithfulness dimension. LLM-based evaluators (Liu et al., 2023a; Fu et al., 2024a) can easily assess multiple dimensions but suffer from instability and coarse granularity. Recently, several works (Song et al., 2024a; Wan et al., 2024; Yang et al., 2024; Scirè et al., 2024; Jeong et al., 2025; Koupaee et al., 2025) have achieved fine-grained and interpretable evaluation through atomic facts (Liu et al., 2023b; Min et al., 2023) and LLM-based claim verification (Song et al., 2024b; Tang et al., 2024a; Oh et al., 2025; Chen et al., 2025b). Following this paradigm, we integrate complementary key fact alignment (Song et al., 2024a) and summary fact verification (Wan et al., 2024) to deliver a fully fact-level evaluation across all dimensions.

3 OMNICSEVAL

3.1 Conversation Collection

Data Sources. We aggregate dialogues from 13 established datasets across six real-world scenarios. **Daily Life** captures informal exchanges from SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021). **Media Interview** utilizes MediaSum (Zhu et al., 2021) for broadcast discourse. **Screenplay** incorporates SummScreen (Chen et al., 2022) and MovieSum (Saxena and Keller, 2024), characterized by extensive narratives and rich character interactions. For **Customer Service**, TweetSumm (Feigenblat et al., 2021) and TODSum (Zhao et al., 2021) provide task-oriented, problem-solving conversations. **Healthcare** features specialized clinical consultations sourced from ACI-Bench (Yim et al., 2023), PriMock57 (Papadopoulos Korfiatis et al., 2022), and MTS-Dialog (Ben Abacha et al., 2023). Finally, **Meeting** leverages MeetingBank (Hu et al., 2023), QMSum (Zhong et al., 2021), and ECTSum (Mukherjee et al., 2022) to represent multi-party professional discussions. Detailed dataset descriptions are provided in Appendix A.

Scenario	Samples	#Tokens	#Sents.	#Spks.
Open-Ended				
Daily Life	300	214.9	19.8	2.1
Media Interview	300	1,682.5	90.9	5.8
Screenplay	300	12,867.8	1,240.6	19.2
Private				
Customer Service	300	266.4	22.0	2.0
Healthcare	300	1,073.6	78.4	2.0
Meeting	300	10,261.4	644.6	8.1
Total	1,800	4,394.4	349.4	6.5

Table 2: Conversation statistics across the six scenarios. #Tokens, #Sents., and #Spks. denote the average number of tokens, sentences, and speakers, respectively.

Filtering and Sampling. To ensure data quality, we first eliminate near-duplicates using a pairwise Jaccard similarity threshold of 0.70. Next, we apply a bilateral interquartile range filter (scale factor 1.5) to remove length outliers within each dataset, and then constrain conversation lengths to between 128 and 32k tokens[5]. Finally, we conduct stratified random sampling to draw 300 instances per scenario, establishing a final benchmark of 1,800 diverse evaluations. Detailed sampling quotas and statistics are provided in Table 7 (Appendix B).

Data Analysis. As shown in Table 2, the scenarios exhibit diverse intrinsic structural complexities, ranging from concise interactions in Customer Service and Daily Life to lengthy dialogues in Meeting and Screenplay. Beyond length, the conversational dynamics vary significantly, spanning from dyadic exchanges to multi-party environments, highlighting the comprehensiveness of our benchmark.

3.2 Summary Generation

We evaluate a comprehensive suite of 28 LLMs, categorized into four groups based on reasoning capability and model scale[6]. For the **Large Reasoning** models, we select GPT-5, Gemini 2.5 Pro, Qwen3-235B-A22B-Thinking, DeepSeek-R1, GLM-4.6-Thinking, Kimi-K2-Thinking, and MiniMax-M2. The **Small Reasoning** category consists of GPT-5-mini, GPT-5-nano, Gemini 2.5 Flash, Gemini 2.5 Flash Lite, Qwen3-14B-Thinking, Qwen3-8B-Thinking, and GLM-Z1-9B. Regarding **Large Instruction** systems, we evaluate GPT-4.1, GPT-4o, Qwen2.5-72B-Instruct-128K, Qwen3-235B-A22B-Instruct, DeepSeek-V3, and GLM-4.6-Instruct. Finally, the **Small Instruction** group comprises GPT-

[5]We use the `gpt-4o` tokenizer from tiktoken.

[6]Models with a `Thinking` suffix explicitly generate intermediate reasoning, whereas `Instruct` systems bypass explicit thinking processes.

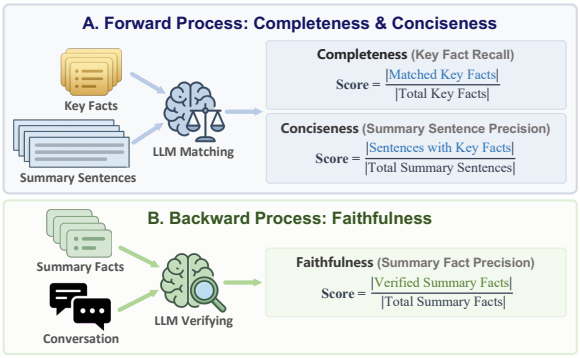

Figure 2: Overview of bidirectional fact-checking.

4.1-mini, GPT-4.1-nano, GPT-4o-mini, Gemini 2.0 Flash, Gemini 2.0 Flash Lite, Qwen3-14B-Instruct, Qwen3-8B-Instruct, and GLM-4-9B-Chat. See Table 8 (Appendix C) for detailed specifications, such as versions, sizes, and reasoning configurations.

To assess the intrinsic capabilities of LLMs, we apply a standardized zero-shot instruction: *Summarize the above conversation*. To minimize randomness, we set the temperature to zero. In total, we obtain 50,400 summaries (1,800 samples $\times$ 28 models) for evaluation.

3.3 Annotation and Evaluation

We focus on three key aspects of summary quality[7]: **Completeness** (whether the summary captures all key content), **Conciseness** (whether the summary avoids non-salient information), and **Faithfulness** (whether the summary is factually consistent with the source dialogue). Figure 2 outlines the bidirectional fact-checking framework. It integrates key fact matching (Song et al., 2024a) and summary fact verification (Wan et al., 2024) to perform an interpretable, fully fact-level assessment. Since the reliability of this framework hinges on the quality of evaluation units, we establish rigorous pipelines to curate key and summary facts.

Key and Summary Fact Collection. We annotate key facts using a three-stage human-LLM collaborative pipeline comprising preliminary extraction, self-review, and human adjudication. First, we utilize the advanced Gemini 3 Pro (high thinking-level) to extract candidate key facts from the source conversation adhering to importance and atomicity. The model then performs a second-round self-reflection to scrutinize these facts and filter out non-essential ones. Candidates consistently retained across both rounds are automatically accepted as high-confidence key facts, while those rejected during self-review undergo human adjudication.

Scenario	Self-Rej.	IAA	Agreed	Final
Open-Ended				
Daily Life	547	0.612	219	2,572
Media Interview	763	0.594	226	4,518
Screenplay	796	0.566	339	5,292
Private				
Customer Service	647	0.454	289	2,979
Healthcare	1,039	0.672	231	6,414
Meeting	967	0.595	355	5,856
Total	4,759	0.601	1,659	27,631

Table 3: Key fact annotation statistics. Self-Rej. counts rejected facts. IAA reports Cohen's κ. Agreed indicates the number of facts unanimously accepted as key, and Final represents the total number of key facts collected.

To ensure annotation quality, we enlist four NLP experts (i.e., the first four authors). They are organized into two decision groups, with each member independently re-evaluating the rejected facts. To mitigate cognitive load, we also implement context mapping (Lee et al., 2024), highlighting up to 25% (capped at 2,048 tokens for lengthy dialogues) of the source sentences that are most semantically similar to the rejected facts[8]. To facilitate accurate judgment, we further provide annotators with the specific rejection reasons generated by Gemini 3 Pro. By synthesizing the focused context with the model's critique, the experts make a final binary determination (key or non-key) for each fact. As shown in Table 3, this annotation achieves substantial inter-annotator agreement, yielding a Cohen's κ of 0.60. To ensure criticality, only facts unanimously agreed upon as key are integrated into the final dataset. Detailed statistics and instructions are provided in Appendices E and H, respectively.

For summary fact processing, we implement a two-stage pipeline comprising atomic decomposition and multi-LLM consensus filtering. Initially, we utilize DeepSeek-V3.2-Instruct to decompose all 50,400 summaries into atomic facts adhering to losslessness and atomicity. To prevent hallucinated artifacts from corrupting downstream faithfulness evaluations, we verify the entailment of each fact against its source summary through a majority vote among GLM-4.7-Instruct, Kimi-K2-Instruct, and Qwen3-Coder-480B-A35B. Facts deemed unsupported by at least two judges are classified as hallucinations and discarded. Notably, only 0.22% (2,387) of the 1.1 million facts are flagged, under-

[7]We exclude *fluency* and *coherence* because current LLMs perform well on these dimensions (Song et al., 2025).

[8]We use `text-embedding-3-small` from OpenAI for sentence and fact embeddings.

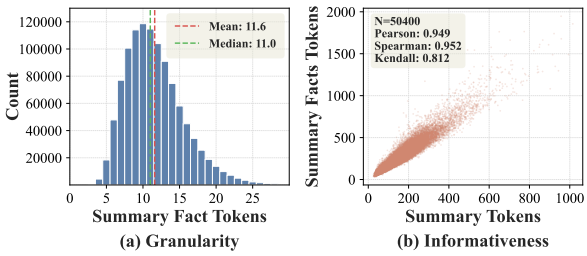

Figure 3: Statistics of collected summary facts.

scoring the high reliability of our decomposition process. Figure 3 further validates the fine granularity and losslessness of the decomposition. Detailed instructions are provided in Appendix H.

Bidirectional Fact-Checking and Scoring. We leverage DeepSeek-V3.2-Instruct as the backbone model to execute this framework. Forward matching determines the semantic entailment of key facts by summary sentences, whereas backward verification identifies whether summary facts are supported by the source conversation. Consequently, completeness is measured as the recall of key facts, conciseness is quantified by the density of informative summary sentences, and faithfulness is calculated as the precision of trustworthy summary facts (see Appendices D and H for detailed formulas and instructions). In meta-evaluation (§5), we show the stability of results, the robustness of leaderboards, and fact-level alignment with human judgments.

4 Results and Analysis

4.1 Main Results

Figure 4 and Table 4 present the overall landscape and evaluation results across 28 LLMs. We highlight several critical findings regarding scenario challenges and the impact of explicit reasoning and model scale on conversation summarization.

Cross-Scenario Challenges. As presented in Table 4, no single scenario yields the highest scores across all dimensions, indicating that *different scenarios present distinct difficulties*. In Meeting and Screenplay, LLMs exhibit a pronounced decline in both completeness and conciseness, with Meeting generally suffering the poorest conciseness. This highlights that current LLMs still struggle to accurately and effectively capture essential information from extensive and entangled conversations. Media Interview demonstrates similar, albeit less severe, challenges in these two dimensions. Intriguingly, while LLMs achieve respectable completeness and conciseness in the seemingly simple Daily Life scenario, they overwhelmingly exhibit the worst faithfulness. This underscores a severe vulnerability to hallucinations when processing highly open-ended and informal interactions. Furthermore, Screenplay and Customer Service also pose moderate-to-high faithfulness challenges. Finally, the moderate hallucination rates observed in Healthcare represent a critical concern, given the high stakes and strict factual requirements of medical documentation.

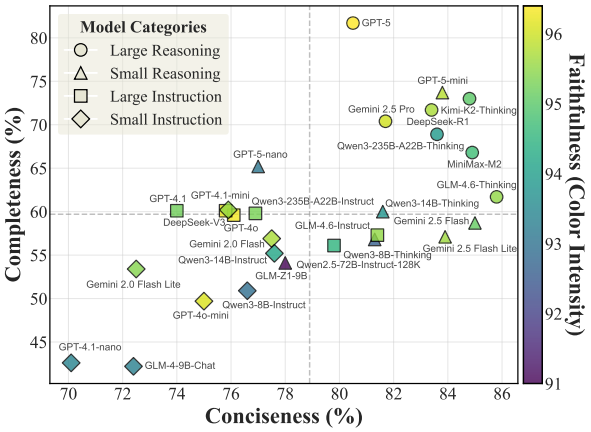

Figure 4: **Overall performance landscape**. Large reasoning models lead in both completeness and conciseness. Small reasoning models match large instruction models in completeness but outperform them in conciseness, whereas small instruction models lag behind.

Reasoning Is the Key to Completeness and Conciseness, but Not Faithfulness. As shown in Figure 4 and Table 4, reasoning systems demonstrate a decisive advantage in completeness and conciseness over instruction models of similar scale. However, they yield no gain, or even regress, in faithfulness. For instance, DeepSeek-R1 significantly outperforms DeepSeek-V3 in completeness (71.7% vs. 60.1%) and conciseness (83.4% vs. 75.8%), yet lags in faithfulness (95.7% vs. 96.2%, $p < 0.01$). This trade-off is mirrored by Qwen3-235B-A22B, whose Thinking variant suffers a notable faithfulness drop compared to its Instruct version (95.3% $\rightarrow$ 93.9%). This tendency is pervasive across the benchmark: Kimi-K2-Thinking and MiniMax-M2 trail the majority of large instruction systems in faithfulness; GLM-4.6-Thinking merely ties its instruction equivalent (95.5%); and even the SOTA GPT-5 exhibits no meaningful improvement over GPT-4o (96.4% vs. 96.3%, $p > 0.05$). Smaller models, such as the Qwen3-8B and -14B series, echo this exact dynamic. These results reveal that *while explicit reasoning is highly effective for key content identification and noise data filtering, it paradoxically increases the risk of hallucinations*.

Model Name	OMNICSEVAL (*Multi-dimensional evaluation results for* **Completeness / Conciseness / Faithfulness**)						
	Open-Ended			*Private*			**Overall Performance**
	Screenplay	Interview	Daily	Meeting	Healthcare	Customer	
Large Reasoning Language Models							
GPT-5 🔒	75.1 / 76.7 / 97.1	70.1 / 81.3 / 97.5	81.8 / 87.8 / 93.1	80.7 / 63.6 / 98.2	91.2 / 90.3 / 96.5	91.1 / 83.5 / 95.8	81.7 / 80.5 / 96.4
Gemini 2.5 Pro 🔒	63.1 / 73.1 / 96.3	60.1 / 78.9 / 96.4	75.6 / 90.2 / 93.4	61.6 / 63.2 / 97.7	80.2 / 95.0 / 96.5	81.6 / 89.6 / 95.6	70.4 / 81.7 / 96.0
Qwen3-235B-A22B-Thinking 🔓	52.8 / 74.4 / 95.4	63.7 / 83.9 / 94.2	79.0 / 93.7 / 90.2	55.0 / 71.5 / 96.3	78.9 / 92.2 / 93.8	84.1 / 85.8 / 93.3	68.9 / 83.6 / 93.9
DeepSeek-R1 🔓	57.8 / 79.1 / 96.4	62.2 / 82.9 / 95.8	86.4 / 87.0 / 93.8	53.9 / 74.4 / 97.0	81.6 / 92.5 / 95.6	88.3 / 84.5 / 95.6	71.7 / 83.4 / 95.7
GLM-4.6-Thinking 🔓	51.2 / 80.0 / 95.4	56.3 / 83.5 / 97.2	67.9 / 91.0 / 92.0	51.0 / 74.2 / 97.3	67.9 / 95.1 / 96.4	76.1 / 90.8 / 94.8	61.7 / 85.8 / 95.5
Kimi-K2-Thinking 🔓	63.0 / 81.8 / 95.6	65.3 / 83.1 / 95.8	79.4 / 90.2 / 92.2	62.8 / 72.9 / 96.8	82.3 / 93.0 / 95.8	85.2 / 88.1 / 94.0	73.0 / 84.8 / 95.0
MiniMax-M2 🔓	51.3 / 84.2 / 93.8	56.8 / 84.2 / 95.8	76.6 / 90.2 / 92.3	57.1 / 71.8 / 96.7	75.4 / 92.7 / 94.9	83.2 / 86.3 / 95.0	66.8 / 84.9 / 94.7
Small Reasoning Language Models							
GPT-5-mini 🔒	51.8 / 77.9 / 95.7	63.7 / 87.0 / 96.4	81.4 / 92.3 / 92.6	67.9 / 70.5 / 97.3	85.8 / 93.0 / 96.4	91.5 / 82.1 / 96.2	73.7 / 83.8 / 95.8
GPT-5-nano 🔒	34.5 / 66.6 / 90.9	58.5 / 79.8 / 94.1	75.7 / 86.1 / 89.4	54.7 / 64.1 / 95.7	80.1 / 87.5 / 94.2	87.4 / 78.1 / 94.5	65.2 / 77.0 / 93.1
Gemini 2.5 Flash 🔒	56.5 / 82.1 / 95.3	48.6 / 81.3 / 95.9	67.0 / 90.2 / 91.9	46.2 / 70.8 / 97.1	65.7 / 95.7 / 96.0	68.5 / 89.5 / 94.4	58.7 / 85.0 / 95.1
Gemini 2.5 Flash Lite 🔒	44.4 / 75.0 / 94.8	51.1 / 82.3 / 97.3	66.2 / 90.6 / 92.7	44.8 / 68.8 / 97.5	66.6 / 93.6 / 96.4	69.8 / 92.9 / 95.1	57.1 / 83.9 / 95.6
Qwen3-14B-Thinking 🔓	36.5 / 69.4 / 91.8	53.0 / 85.1 / 95.2	76.4 / 89.2 / 90.6	43.5 / 69.9 / 95.5	70.0 / 92.6 / 95.0	80.7 / 83.2 / 95.0	60.0 / 81.6 / 93.8
Qwen3-8B-Thinking 🔓	31.6 / 64.7 / 89.5	52.0 / 85.3 / 93.7	70.6 / 90.0 / 89.9	42.3 / 69.9 / 94.3	67.5 / 93.6 / 94.5	76.9 / 84.1 / 93.7	56.8 / 81.3 / 92.6
GLM-Z1-9B 🔓	26.5 / 56.9 / 85.7	50.5 / 81.5 / 92.9	68.9 / 87.8 / 88.7	39.3 / 67.9 / 92.6	63.4 / 91.3 / 93.1	75.9 / 82.5 / 93.1	54.1 / 78.0 / 91.0
Large Instruction Language Models							
GPT-4.1 🔒	44.6 / 63.5 / 96.0	48.2 / 71.1 / 95.9	71.5 / 84.8 / 92.0	42.6 / 54.6 / 97.3	76.3 / 89.0 / 95.5	77.4 / 80.9 / 94.7	60.1 / 74.0 / 95.2
GPT-4o 🔒	42.5 / 66.4 / 96.8	50.4 / 74.0 / 96.7	74.1 / 87.7 / 93.9	36.5 / 56.0 / 97.6	73.7 / 90.3 / 96.5	80.2 / 82.2 / 96.1	59.6 / 76.1 / 96.3
Qwen3-235B-A22B-Instruct 🔓	42.9 / 65.5 / 95.6	51.3 / 76.6 / 96.4	74.1 / 85.0 / 91.7	41.9 / 62.3 / 96.7	72.3 / 90.8 / 95.8	76.4 / 81.1 / 95.7	59.8 / 76.9 / 95.3
Qwen2.5-72B-Instruct-128K 🔓	39.2 / 71.8 / 94.7	47.5 / 82.6 / 95.6	72.3 / 87.9 / 91.9	35.6 / 62.6 / 95.8	64.9 / 91.6 / 94.6	77.1 / 82.1 / 94.4	56.1 / 79.8 / 94.5
DeepSeek-V3 🔓	39.0 / 65.2 / 96.7	50.8 / 75.9 / 96.6	82.1 / 85.7 / 93.5	34.2 / 58.7 / 97.9	71.8 / 91.0 / 96.2	82.5 / 78.3 / 96.3	60.1 / 75.8 / 96.2
GLM-4.6-Instruct 🔓	44.6 / 75.0 / 95.5	47.0 / 80.8 / 95.9	72.6 / 87.6 / 92.8	38.3 / 67.2 / 97.8	64.8 / 93.4 / 95.6	76.4 / 84.2 / 95.2	57.3 / 81.4 / 95.5
Small Instruction Language Models							
GPT-4.1-mini 🔒	36.4 / 60.7 / 95.2	51.8 / 76.1 / 96.0	73.5 / 87.2 / 92.9	46.1 / 58.8 / 97.2	72.6 / 91.8 / 95.9	80.5 / 80.5 / 95.7	60.2 / 75.9 / 95.5
GPT-4.1-nano 🔒	16.8 / 43.5 / 93.5	38.5 / 76.1 / 94.1	56.9 / 83.8 / 89.3	20.9 / 48.2 / 96.6	58.1 / 90.9 / 94.0	64.1 / 77.9 / 92.6	42.6 / 70.1 / 93.4
GPT-4o-mini 🔒	26.3 / 56.5 / 96.3	43.4 / 77.3 / 97.3	68.9 / 88.3 / 93.5	27.6 / 55.8 / 97.3	60.8 / 91.1 / 96.4	71.2 / 81.0 / 95.6	49.7 / 75.0 / 96.1
Gemini 2.0 Flash 🔒	43.0 / 69.8 / 95.0	46.5 / 75.1 / 96.6	70.4 / 88.2 / 92.8	37.2 / 52.7 / 98.0	70.5 / 92.1 / 96.5	73.8 / 86.9 / 95.4	56.9 / 77.5 / 95.7
Gemini 2.0 Flash Lite 🔒	32.0 / 57.7 / 95.0	46.1 / 70.9 / 97.0	69.2 / 83.2 / 91.9	34.2 / 53.6 / 97.8	63.9 / 88.8 / 96.0	74.9 / 80.6 / 94.5	53.4 / 72.5 / 95.4
Qwen3-14B-Instruct 🔓	33.6 / 62.9 / 91.8	45.7 / 81.8 / 95.2	73.1 / 85.1 / 90.7	35.2 / 62.3 / 96.6	65.3 / 92.0 / 95.4	78.2 / 81.6 / 95.5	55.2 / 77.6 / 94.2
Qwen3-8B-Instruct 🔓	29.3 / 56.6 / 89.5	43.0 / 79.6 / 94.3	66.6 / 87.8 / 90.2	34.7 / 59.7 / 94.9	61.6 / 92.5 / 94.2	70.0 / 83.4 / 94.5	50.9 / 76.6 / 92.9
GLM-4-9B-Chat 🔓	20.1 / 47.3 / 91.2	36.7 / 76.0 / 95.1	61.8 / 87.2 / 89.5	19.2 / 49.9 / 96.9	50.7 / 92.1 / 94.1	64.7 / 82.1 / 93.1	42.2 / 72.4 / 93.3

Table 4: Main evaluation results. 🔒 and 🔓 denote *proprietary* and *open-weights* systems, respectively. The values in each cell represent **Completeness**, **Conciseness**, and **Faithfulness** of the model-generated summaries, respectively. All reported scores are *macro-averaged* (%). We highlight the *best* model for each dimension in each LLM category.

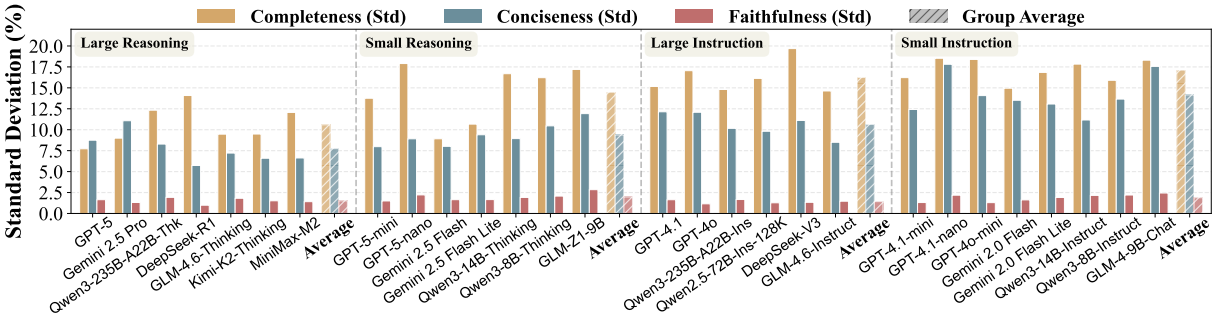

Figure 5: Stability of LLMs across six scenarios. Lower standard deviation indicates better performance consistency.

Furthermore, we find that the advantages of reasoning models in completeness and conciseness stem primarily from their robust performance in complex-context scenarios (i.e., Screenplay and Meeting), granting them superior cross-scenario stability in these two dimensions compared to similarly sized instruction models (see Figure 5).

Reasoning Offsets Scale, but Size Still Matters. As shown in Figure 4, small reasoning models effectively *punch above their weight*, matching the completeness and surpassing the conciseness of large instruction systems. This indicates that explicit reasoning can compensate for parameter deficiencies to a remarkable extent. However, scaling laws remain a fundamental constraint: larger models widely achieve superior multi-dimensional performance, especially in completeness and faithfulness, within both reasoning and instruction settings. Crucially, scale dictates robustness. As evidenced by Figure 5, larger models maintain greater cross-scenario stability in both reasoning and instruction paradigms. These findings suggest that *while reasoning serves as a powerful equalizer, model scale remains an indispensable prerequisite for comprehensive and robust conversation summarization*.

4.2 Further Analysis

Thinking Efficiency of Reasoning Models. As Figure 6 shows, larger reasoning models densely cluster in the upper-left region, demonstrating su-

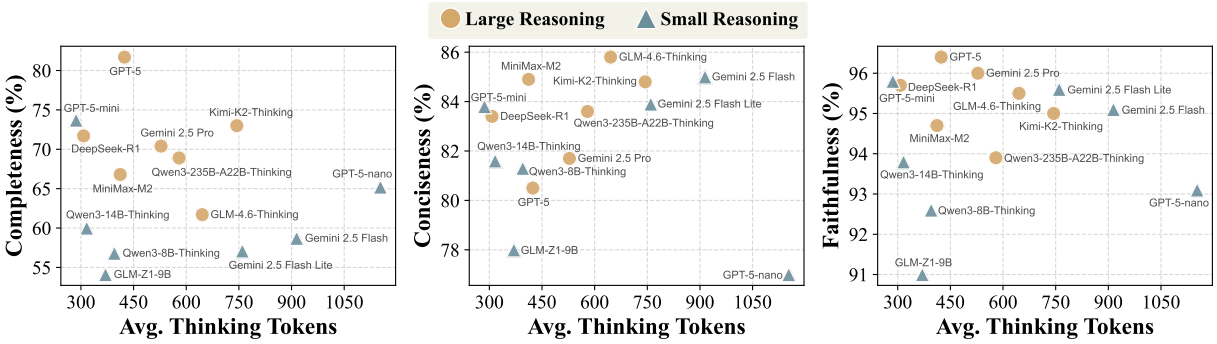

Figure 6: Visualization of average thinking tokens versus multi-dimensional performance. Models positioned closer to the upper-left corner demonstrate superior thinking efficiency, achieving high scores with minimal reasoning cost.

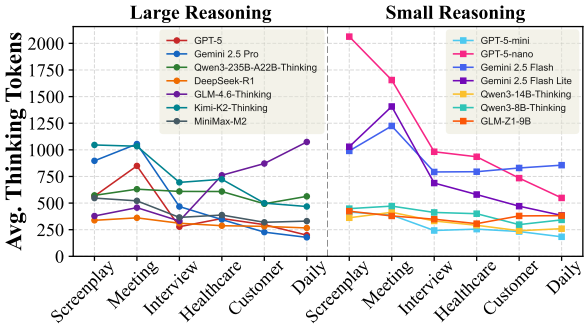

Figure 7: Thinking length distribution of reasoning models across scenarios, sorted by context structural complexity (i.e., dialogue length and number of speakers).

Dimension	System-Level	Summary-Level
Completeness	0.999	0.982 ↑ 0.129
Conciseness	0.995	0.962 ↑ 0.054
Faithfulness	0.977	0.912*

Table 5: Evaluation stability across three runs. System-level is quantified by Kendall's W for ranking consistency, and summary-level by Krippendorff's α for score agreement. ↑ indicates the advantage of our implementation over FineSure (Song et al., 2024a). * represents no comparable results in ACUEval (Wan et al., 2024).

perior and more consistent thinking efficiency. Crucially, models that think longer do not necessarily achieve greater performance, particularly in completeness and faithfulness. For instance, Gemini 2.5 Flash and GLM-4.6-Thinking consume significantly more tokens than highly efficient peers such as GPT-5-mini and GPT-5, yet fail to achieve comparable completeness and faithfulness. A conspicuous outlier is GPT-5-nano, which exhibits the highest token consumption yet remains in the middle-to-lower performance tier for conciseness and faithfulness. These results underscore that *thinking efficiently is far more critical than simply thinking extensively for conversation summarization*.

Reasoning Adaptability to Scenario Complexity. Beyond efficiency, a highly adaptable reasoning model should dynamically scale its reasoning budget to match scenario structural complexity. As Figure 7 illustrates, *models exhibit distinctly different strategies for allocating thinking resources*. Flexible thinkers (e.g., Kimi-K2-Thinking) demonstrate strong context awareness, systematically reducing their thinking duration as input complexity decreases. Conversely, rigid models (e.g., Qwen3-235B-A22B-Thinking and DeepSeek-R1) maintain a nearly constant reasoning budget across all scenarios. Most notably, GLM-4.6-Thinking displays an anomalous inverse trend, expending significantly more thinking tokens on lower-complexity scenarios. This highlights a critical misalignment in cognitive load allocation, suggesting the model struggles to accurately gauge task complexity. A similar inverse pattern appears among small models, where systems like GLM-Z1-9B and Gemini 2.5 Flash devote disproportionately heavy reasoning efforts to straightforward scenarios like Customer Service and Daily Life.

5 Meta Evaluation

Stability of Benchmark Results. Considering the inherent randomness of LLM evaluators (even with the temperature set to zero (Atil et al., 2025)), we assess the stability of our benchmark results to ensure reproducibility. For all 28 evaluated LLMs, we sample 300 summaries per model (50 per scenario). To guarantee representative coverage, this sampling is stratified by conversation length within each scenario, encompassing all sub-datasets and resulting in a total of 8,400 summaries. We then execute the bidirectional fact-checking framework three independent times using the DeepSeek-V3.2-Instruct backbone. As shown in Table 5, the evaluations exhibit exceptional consistency, confirming our results are highly stable and reproducible.

LLM as a Fact-Checker	Matching	Verification
	Macro F1	BAcc
GPT-4o (*Previous SOTA Judge*)	0.935	0.774
Qwen3-Next-80B-A3B-Instruct	0.929	0.749
Llama 4 Maverick	0.931	0.755
Mistral Large 3	0.931	0.762
Kimi-K2-Instruct	0.953	0.805
GLM-4.7-Instruct	0.946	0.853
DeepSeek-V3.2-Instruct	**0.958**	**0.871**

Table 6: Performance of various LLMs as fact-checkers.

Alignment with Human Judgments. To validate evaluation accuracy, we construct a human-annotated test set comprising 120 summaries (20 per scenario from 20 randomly selected LLMs), excluding the baseline fact-checker GPT-4o. We enlist six graduate students, divided into two groups, to perform key fact matching and summary fact verification, respectively. For the matching task, any complete initial disagreement among the three annotators triggers a secondary review until a majority consensus is reached. Finally, we determine the gold labels via majority voting, yielding a total of 1,871 matching and 2,625 verification annotations. Details of the annotation process and inter-annotator agreement are provided in Appendix F.

We evaluate the performance of seven advanced LLMs as fact-checkers, ensuring that all evaluator models are strictly disjoint from the LLMs being evaluated. We employ Macro F1 (Magomere et al., 2026) to assess the matching task and Balanced Accuracy (BAcc) (Tang et al., 2024a) for the verification task. As presented in Table 6, DeepSeek-V3.2-Instruct demonstrates exceptionally strong alignment with human judgments across both tasks, firmly indicating the accuracy of our evaluation.

Leaderboard Robustness. A reliable benchmark leaderboard should be resilient to the choice of the automated LLM judge. To verify this, we re-evaluated the 8,400 summaries from our stability assessment (encompassing all 28 LLMs) using two additional advanced LLMs as backbones: GLM-4.7-Instruct and Kimi-K2-Instruct. As presented in Figure 8, our evaluation demonstrates exceptional cross-judge consensus. Evaluators achieve near-perfect alignment in the highly discriminative completeness and conciseness (system-level average $\sigma = 0.095$ and 0.049, respectively). Even for faithfulness, which exhibits very low variance (average $\sigma = 0.012$), rank correlations remain strong. These results confirm that the OMNICSEVAL leaderboard possesses robust, judge-agnostic reliability.

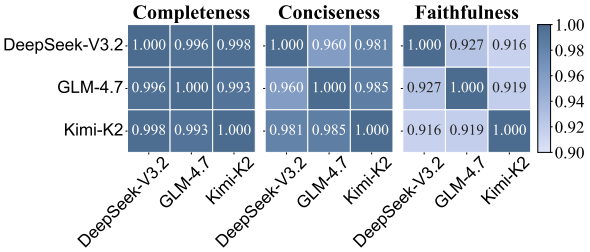

Figure 8: Pairwise system-level rank correlations among three distinct LLM judges (Spearman's ρ, all $p < 0.01$).

6 Discussion

Which Model Should We Choose for Real-World Deployment? For applications demanding superior summarization quality, particularly within complex long-context scenarios, large reasoning models represent the optimal choice. Under budgetary constraints, small reasoning models serve as a powerful alternative, capable of rivaling large instruction models. Furthermore, reasoning systems are indispensable where concise summarization is prioritized. Conversely, for risk-averse tasks requiring hallucination mitigation, large instruction models remain the safest option. Finally, in resource-limited environments demanding low latency, small instruction models constitute the sole viable solution, assuming the quality trade-off is permissible.

Can Open-Weights Models Replace Proprietary Systems? Open-weights models now rival proprietary systems in large-scale reasoning and instruction. They often trade leads in completeness and conciseness. Despite this, a significant gap persists in small reasoning models where open-weights options lag behind SOTA leaders such as GPT-5-mini. Proprietary models also maintain superior faithfulness across all categories. This reliability makes them the safer choice for high-stakes environments.

7 Conclusion

In this study, we introduced OMNICSEVAL, a comprehensive and fine-grained benchmark designed to systematically evaluate LLMs in conversation summarization. Our empirical analysis reveals that LLMs face distinct, scenario-specific challenges, and demonstrates the impacts of explicit reasoning and model scale on multi-dimensional performance and cross-scenario stability. We also highlight the importance of reasoning efficiency and adaptability. We hope this evaluation work provides actionable guidance for real-world system deployment and establishes a rigorous foundation for future advancements in conversation summarization.

Limitations

While OMNICSEVAL represents a significant advancement in the breadth and depth of conversation summarization evaluation, the current study focuses primarily on the English language. This constraint is intentionally adopted to prioritize the precision and reliability of the evaluation framework, thereby minimizing potential bias interference arising from cross-lingual semantic alignment issues. Recognizing that multilingual and cross-lingual capabilities are essential for general-purpose LLMs, we plan to explore broader linguistic environments in future dialogue summarization evaluation work.

Ethics Statement

OMNICSEVAL aggregates 13 existing and publicly available datasets. While our benchmark includes domains that traditionally handle sensitive interactions, such as healthcare and customer service, we strictly utilized established corpora that have been thoroughly de-identified. These datasets were stripped of all personally identifiable information by their original creators. Consequently, this study does not introduce any new private or sensitive data into the public domain. For the meta-evaluation, we enlisted six graduate students. Prior to their participation, all annotators were fully informed of the study's purpose and provided explicit consent. To ensure fair labor practices, annotators were compensated at an hourly rate of $8, which exceeds the local minimum wage requirements.

We acknowledge the substantial computational resources and environmental footprint required to conduct this study. Evaluating 28 LLMs to generate 50,400 summaries and executing a consensus pipeline on approximately 1.1 million facts resulted in a total API expenditure of approximately $2,300. However, by releasing OMNICSEVAL as a comprehensive and reusable benchmark, we aim to offset this initial environmental and financial cost. Our framework is designed to prevent redundant and computationally expensive evaluations, ultimately saving compute for the broader NLP community.

Acknowledgement

This research was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, 62406033, U1636211, 61672081) and the State Key Laboratory of Complex & Critical Software Environment (Grant No. SKLCCSE-2024ZX-18).

References

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" llm settings. *Preprint*, arXiv:2408.04667.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. 2025b. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14976–14995, Vienna, Austria. Association for Computational Linguistics.

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

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024a. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. 2024b. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.

Mingqi Gao and Xiaojun Wan. 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*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Akanksha Gupta, Bijo Thomas, Harshita Asnani, Phanindra Reddy Madduru, Samia Feroze, Shreyas Subramanian, Vikram Elango, and Mecit Gungor. 2026. Small language models (slms) can still pack a punch: A survey (updated 2026). *Preprint*, arXiv:2501.05465.

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

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23591–23608, Suzhou, China. Association for Computational Linguistics.

Keyan Jin, Yapeng Wang, Leonel Santos, Tao Fang, Xu Yang, Sio Kei Im, and Hugo Gonçalo Oliveira. 2026. Reasoning or not? a comprehensive evaluation of reasoning llms for dialogue summarization. *Expert Systems with Applications*, 299:129831.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025. Cads: A systematic literature review on the challenges of abstractive dialogue summarization. *J. Artif. Int. Res.*, 82.

Mahnaz Koupaee, Jake W. Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, Kyu J. Han, and Hang Su. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12209–12246, Albuquerque, New Mexico. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 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*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long

Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Jabez Magomere, Elena Kochkina, Samuel Mensah, Simerjot Kaur, Fernando Acero, Arturo Oncevay, Charese Smiley, Xiaomo Liu, and Manuela Veloso. 2026. Distill and align decomposition for enhanced claim verification. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5887–5912, Rabat, Morocco. Association for Computational Linguistics.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yoen Kim, Taewon Yun, and Hwanjun Song. 2025. Learning to verify summary facts with fine-grained LLM feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 230–242, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 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)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Rohit Saxena and Frank Keller. 2024. MovieSum: An abstractive summarization dataset for movie screenplays. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4043–4050, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

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

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from LLM-generated feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857, Albuquerque, New Mexico. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024b. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. 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)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 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*, pages 56–67, Singapore. Association for Computational Linguistics.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10036–10056, Bangkok, Thailand. Association for Computational Linguistics.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Trans. Intell. Syst. Technol.*, 16(6).

Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu, and Bingsheng He. 2025. Evaluating small language models for news summarization: Implications and factors influencing performance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4909–4922, Albuquerque, New Mexico. Association for Computational Linguistics.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Preprint*, arXiv:2306.02022.

Haohan Yuan and Haopeng Zhang. 2026. Understanding llm reasoning for abstractive summarization. *Preprint*, arXiv:2512.03503.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A comparative study of explicit and implicit gender biases in large language models via self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 186–198, Torino, Italia. ELRA and ICCL.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 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*, pages 5905–5921, Online. Association for Computational Linguistics.

Weixiao Zhou, Gengyao Li, Xianfu Cheng, Xinnian Liang, Junnan Zhu, Feifei Zhai, and Zhoujun Li. 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*, pages 6893–6908, Singapore. Association for Computational Linguistics.

Weixiao Zhou, Junnan Zhu, Gengyao Li, Xianfu Cheng, Xinnian Liang, Feifei Zhai, and Zhoujun Li. 2025. What are they talking about? a benchmark of knowledge-grounded discussion summarization. In

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2172–2191, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. Factual dialogue summarization via learning from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492, Abu Dhabi, UAE. Association for Computational Linguistics.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6825–6845, Toronto, Canada. Association for Computational Linguistics.

A Source Dataset Details

SAMSum (Gliwa et al., 2019). It comprises 16,369 dialogue-summary pairs designed to simulate real-life messenger conversations. Constructed by linguists, it captures informal linguistic features (e.g., slang, typos) across diverse daily topics.

DialogSum (Chen et al., 2021). It consists of 13,460 dialogue-summary pairs covering diverse daily-life scenarios, including schooling, work, and shopping. Collected from spoken dialogue corpora (e.g., DailyDialog, DREAM), it captures face-to-face interactions typically between two speakers.

MediaSum (Zhu et al., 2021). This large-scale dataset comprises 463,596 dialogue-summary pairs derived from media interviews. Sourced from NPR and CNN transcripts, it covers open-domain topics such as politics, economy, and crime.

SummScreen (Chen et al., 2022). This dataset comprises 26,851 dialogue-summary pairs derived from TV series transcripts. It consists of two subsets: *ForeverDreaming* and *TVMegaSite*, covering diverse narrative genres such as drama and comedy. The dialogues feature complex, entity-centric communications with an average of over 20 speakers.

MovieSum (Saxena and Keller, 2024). It comprises 2,200 dialogue-summary pairs derived from movie screenplays. Collected from various script repositories, the screenplays are manually formatted using Celtx to preserve structural elements like scene headings and speaker utterances. It spans diverse genres such as drama, comedy, and action.

TweetSumm (Feigenblat et al., 2021). This dataset comprises 1,100 dialogues focusing on real-world customer service interactions. Sourced from the *Customer Support on Twitter* corpus, it covers diverse domains such as airlines and retail.

TODSum (Zhao et al., 2021). This dataset comprises 9,906 dialogue-summary pairs derived from the MultiWOZ corpus, focusing on task-oriented interactions in customer service. It covers five domains: hotel, restaurant, attraction, taxi, and train.

ACI-Bench (Yim et al., 2023). This dataset comprises 207 dialogue-summary pairs focusing on doctor-patient interactions. Constructed through role-playing sessions involving medical experts, it includes both human and machine-generated transcripts of clinical encounters.

MTS-Dialog (Ben Abacha et al., 2023). It comprises 1,700 dialogue-summary pairs focusing on doctor-patient interactions. The dialogues are manually simulated by healthcare professionals based on authentic clinical notes. It covers diverse specialties including neurology and orthopedics.

PriMock57 (Papadopoulos Korfiatis et al., 2022). It comprises 57 dialogue-summary pairs focusing on mock primary care consultations. The dialogues feature simulated interactions between clinicians and patients, recorded via telemedicine software.

MeetingBank (Hu et al., 2023). This benchmark comprises 6,892 dialogue-summary pairs derived from city council meetings across six U.S. cities. The dialogues are transcripts generated via ASR. It features multi-party interactions with 2 to 20 speakers covering public affairs.

QMSum (Zhong et al., 2021). It comprises 232 dialogue-summary pairs derived from professional meetings across three domains: *Product*, *Academic*, and *Committee*. Sourced from the AMI and ICSI corpora as well as Welsh and Canadian parliamentary proceedings, it features extensive and complex multi-party interactions.

Domain	Scenario	Dataset	Samples	Conversation		Reference	
				#Tokens	#Sentences	#Tokens	#Sentences
Open-Ended	Daily Life	SAMSum	150	212.4	20.6	36.0	2.6
Open-Ended	Daily Life	DialogSum	150	217.4	19.0	40.9	2.3
Open-Ended	Media Interview	MediaSum-NPR	150	864.3	42.4	49.8	2.4
Open-Ended	Media Interview	MediaSum-CNN	150	2,500.6	139.5	44.6	2.1
Open-Ended	Screenplay	SummScreen-FD	100	8,438.3	834.2	133.7	5.4
Open-Ended	Screenplay	SummScreen-TMS	100	6,454.8	650.1	362.0	21.3
Open-Ended	Screenplay	MovieSum	100	23,710.4	2,237.4	749.4	29.8
Private	Customer Service	TweetSumm	150	278.8	19.8	48.0	2.0
Private	Customer Service	TODSum	150	254.1	24.3	64.3	4.1
Private	Healthcare	ACI-Bench	130	1,416.4	90.4	545.5	34.5
Private	Healthcare	MTS-Dialog	130	407.8	35.2	158.4	8.8
Private	Healthcare	PriMock57	40	2,123.0	179.6	211.6	11.4
Private	Meeting	MeetingBank	75	9,905.7	638.0	106.3	2.9
Private	Meeting	QMSum-Product	70	8,897.0	763.4	123.8	5.9
Private	Meeting	QMSum-Academic	50	18,743.4	1,172.2	112.8	5.4
Private	Meeting	QMSum-Committee	30	17,162.5	801.2	156.8	6.4
Private	Meeting	ECTSum	75	3,475.4	125.9	93.7	5.3

Table 7: Detailed per-dataset statistics of our OMNICSEVAL benchmark.

ECTSum (Mukherjee et al., 2022). This dataset comprises 2,425 dialogue-summary pairs derived from earnings call transcripts collected from The Motley Fool. It captures financial discussions between company executives and analysts.

B Sampled Data Details

This section provides a granular breakdown of the 13 source datasets that constitute the OMNICSEVAL benchmark. Table 7 reports the specific sampling quotas and linguistic statistics for each dataset within their respective scenarios and domains.

The sampling strategy primarily aims to maintain a balanced representation of 300 instances per scenario. However, for certain datasets with limited intrinsic sizes—specifically **PriMock57** (57 total), **QMSum-Academic** (59 total), and **QMSum-Committee** (36 total)—we made necessary sampling adjustments by sampling 40, 50, and 30 instances, respectively. These adjustments are essential to ensure that the sampled subsets remain representative of the source datasets while fulfilling the aggregate requirements for each scenario. In Figure 9, we present the length distribution of the resulting conversations across all scenarios.

C Evaluated LLM Details

This section provides detailed specifications for the 28 evaluated LLMs. As outlined in Table 8, our selection encompasses a diverse array of models, comprising 13 proprietary and 15 open-weight models across various parameter scales. For each model, we document key metadata, including the provider, version, parameter size, and maximum context window. We also specify the architectural type and the specific reasoning configuration (Effort / Budget). Additionally, Table 9 details the 8 LLMs employed in our evaluation pipeline. To mitigate self-evaluation bias, these models are strictly isolated from the evaluated systems.

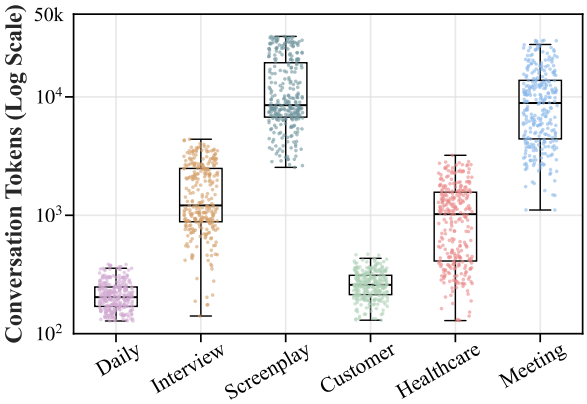

Figure 9: Conversation length distribution across scenarios. Each scatter point denotes an individual sample.

D Evaluation Metric Details

D.1 Forward Key Fact Matching

Let $\mathcal{F}_{key} = \{f_1, f_2, \ldots, f_N\}$ represent the set of key facts extracted from the source conversation, and $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ denote the set of sentences in the model-generated summary. We define an LLM-based matching function $\mathcal{M}(\cdot)$ that determines semantic entailment. The completeness

Model Name	Availability	Provider	Version	Size	Context	Type	Effort / Budget
Large Reasoning Language Models							
GPT-5	Proprietary	OpenAI	2025-08	–	400k	Reasoning	Medium
Gemini 2.5 Pro	Proprietary	Google	2025-06	–	1,024k	Reasoning	8,192
Qwen3-235B-A22B-Thinking	Open-Weights	Qwen	2025-07	235B	256k	Reasoning	8,192
DeepSeek-R1	Open-Weights	DeepSeek	2025-05	671B	160k	Reasoning	8,192
GLM-4.6-Thinking	Open-Weights	Zhipu AI	2025-09	335B	200k	Hybrid	8,192
Kimi-K2-Thinking	Open-Weights	MoonShot	2025-11	1T	256k	Reasoning	8,192
MiniMax-M2	Open-Weights	MiniMax	2025-10	230B	200k	Reasoning	8,192
Small Reasoning Language Models							
GPT-5-mini	Proprietary	OpenAI	2025-08	–	400k	Reasoning	Medium
GPT-5-nano	Proprietary	OpenAI	2025-08	–	400k	Reasoning	Medium
Gemini 2.5 Flash	Proprietary	Google	2025-06	–	1,024k	Hybrid	8,192
Gemini 2.5 Flash Lite	Proprietary	Google	2025-06	–	1,024k	Hybrid	8,192
Qwen3-14B-Thinking	Open-Weights	Qwen	2025-04	14B	128k	Hybrid	8,192
Qwen3-8B-Thinking	Open-Weights	Qwen	2025-04	8B	128k	Hybrid	8,192
GLM-Z1-9B	Open-Weights	Zhipu AI	2025-04	9B	128k	Reasoning	8,192
Large Instruction Language Models							
GPT-4.1	Proprietary	OpenAI	2025-04	–	1,024k	Instruction	–
GPT-4o	Proprietary	OpenAI	2024-11	–	128k	Instruction	–
Qwen3-235B-A22B-Instruct	Open-Weights	Qwen	2025-07	235B	256k	Instruction	–
Qwen2.5-72B-Instruct-128K	Open-Weights	Qwen	2024-09	72B	128k	Instruction	–
DeepSeek-V3	Open-Weights	DeepSeek	2025-03	671B	128k	Instruction	–
GLM-4.6-Instruct	Open-Weights	Zhipu AI	2025-09	335B	200k	Hybrid	Disable
Small Instruction Language Models							
GPT-4.1-mini	Proprietary	OpenAI	2025-04	–	1,024k	Instruction	–
GPT-4.1-nano	Proprietary	OpenAI	2025-04	–	1,024k	Instruction	–
GPT-4o-mini	Proprietary	OpenAI	2024-07	–	128k	Instruction	–
Gemini 2.0 Flash	Proprietary	Google	2025-02	–	1,024k	Instruction	–
Gemini 2.0 Flash Lite	Proprietary	Google	2025-02	–	1,024k	Instruction	–
Qwen3-14B-Instruct	Open-Weights	Qwen	2025-04	14B	128k	Hybrid	Disable
Qwen3-8B-Instruct	Open-Weights	Qwen	2025-04	8B	128k	Hybrid	Disable
GLM-4-9B-Chat	Open-Weights	Zhipu AI	2024-06	9B	128k	Instruction	–

Table 8: Detailed specifications of the evaluated LLMs. The **Type** column classifies the model's native architecture as *Reasoning* (specialized for long chain-of-thought), *Instruction* (standard generation without explicit reasoning), or *Hybrid* (capable of both). **Effort / Budget** denotes the reasoning configuration applied in our benchmark, where `Disable` indicates that explicit reasoning was suppressed for Hybrid models.

score is calculated as the proportion of key facts that are successfully matched by the summary:

$$\text{Completeness} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\exists s, \mathcal{M}(f_i, s) = 1\right]$$

where $\mathbb{I}[\cdot]$ is the indicator function, and $\mathcal{M}(f_i, s) = 1$ if the judge model determines that summary sentence s supports key fact f_i.

Similarly, the conciseness score is calculated as the density of informative summary sentences:

$$\text{Conciseness} = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}\left[\exists f, \mathcal{M}(f, s_j) = 1\right]$$

This metric measures the precision of summary sentences, penalizing sentences that do not contain salient information or cover redundant content.

D.2 Backward Summary Fact Verification

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$ denote the set of atomic facts decomposed from the model-generated summary, and $\mathcal{C}$ represent the source conversation context. We define an LLM-based verification function $\mathcal{V}(\cdot)$ to check factual consistency. The faithfulness score is calculated as the proportion of atomic facts supported by the source conversation:

$$\text{Faithfulness} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\left[\mathcal{V}(a_k, \mathcal{C}) = \text{supported}\right]$$

where $\mathcal{V}(a_k, \mathcal{C})$ returns "supported" if the judge model finds evidence for atomic fact a_k in the conversation context $\mathcal{C}$, and "not supported" otherwise.

D.3 Meta-Evaluation Metrics

To measure the alignment between LLM-based fact-checkers and human annotations in our meta-evaluation (§5), we employ Macro F1 for the key fact matching task and Balanced Accuracy (BAcc) for the summary fact verification task.

For the key fact matching task, the number of summary sentences varies across instances, result-

Model Name	Availability	Provider	Version	Size	Context	Type	Effort / Budget
Gemini 3 Pro	Proprietary	Google	2025-11	–	1,024k	Reasoning	High
DeepSeek-V3.2-Instruct	Open-Weights	DeepSeek	2025-12	671B	160k	Hybrid	Disable
GLM-4.7-Instruct	Open-Weights	Zhipu AI	2025-12	355B	200k	Hybrid	Disable
Kimi-K2-Instruct	Open-Weights	MoonShot	2025-09	1T	256k	Instruction	–
Qwen3-Coder-480B-A35B	Open-Weights	Qwen	2025-07	480B	256k	Instruction	–
Qwen3-Next-80B-A3B-Instruct	Open-Weights	Qwen	2025-09	80B	256k	Instruction	–
Llama 4 Maverick	Open-Weights	Meta	2025-04	400B	1,024k	Instruction	–
Mistral Large 3	Open-Weights	Mistral AI	2025-12	675B	256k	Instruction	–

Table 9: Detailed specifications of the 8 LLMs utilized within the OMNICSEVAL evaluation pipeline. To eliminate self-evaluation bias, these models are strictly isolated from the 28 systems being evaluated.

ing in dynamic dimensions for the binary matching matrices. To address this, we flatten the two-dimensional boolean matching maps of all evaluated key fact-summary sentence pairs into a unified one-dimensional list. Let $y \in \{0, 1\}$ represent the human-annotated ground truth and $\hat{y} \in \{0, 1\}$ denote the LLM prediction. The Macro F1 score is calculated as the unweighted mean of the F_1 scores for both the matched (1) and unmatched (0) classes:

$$\text{Macro F1} = \frac{F_1^{(0)} + F_1^{(1)}}{2}$$

where $F_1^{(c)}$ is the harmonic mean of precision and recall for class $c \in \{0, 1\}$.

For the summary fact verification task, which is formulated as a binary classification problem (supported vs. not supported), we utilize Balanced Accuracy to account for potential class imbalances in the test set. BAcc is defined as the arithmetic mean of the True Positive Rate (TPR) and the True Negative Rate (TNR):

$$\text{BAcc} = \frac{\text{TPR} + \text{TNR}}{2}$$

where TPR represents the proportion of actual supported facts correctly identified, and TNR represents the proportion of actual unsupported (hallucinated) facts correctly identified.

E Key Fact Details

Table 10 reveals a clear tiered distribution of information volume across scenarios. Concise interactions, such as Daily Life and Customer Service, average fewer than 10 key facts, reflecting their focused nature. Conversely, the information load nearly doubles in structurally complex scenarios such as Screenplay and Meeting. Notably, Healthcare yields the highest fact density, containing the most key facts despite its moderate context length. This density stems from the high concentration of critical, discrete units inherent in clinical consultations, such as symptoms and treatment plans. Figure 17 provides concrete examples.

Scenario	#Count (min-max)	#Tokens (min-max)
Open-Ended		
Daily	8.6 (4-16)	102.8 (38-224)
Interview	15.1 (4-29)	264.8 (47-484)
Screenplay	17.6 (10-31)	318.9 (134-613)
Private		
Customer	9.9 (3-23)	135.2 (38-272)
Healthcare	21.4 (4-48)	253.3 (40-496)
Meeting	19.5 (10-47)	392.7 (143-1,098)
Overall	15.4 (3-48)	244.6 (38-1,098)

Table 10: Statistics of key facts across different scenarios. The column #Count denotes the average number of key facts per conversation, while #Tokens represents the average total token count of these facts.

F Human Annotation Details

To reduce cognitive load and enhance annotation quality, we introduced assistive mechanisms. For the key fact matching task, we provided annotators with the summary sentences alongside their specific similarity scores to the key fact, highlighting the top three most similar sentences (or only the top one if the summary contained three or fewer sentences). For the summary fact verification task, we similarly employed context mapping by highlighting the top 25% of source sentences most similar to the summary fact (capped at 2,048 tokens).

Throughout the annotation process, we did not explicitly provide any LLM pre-annotation results. However, when encountering challenging cases, annotators were permitted to access a designated verification platform[9]. When utilizing this feature, annotators were required to consult the judgments of at least three distinct LLMs and synthesize these insights to reach a final, independent decision. Both tasks exhibited substantial inter-annotator agree-

[9] https://openrouter.ai/chat

Human Reference	Comp.	Conci.	Faith.
Screenplay	49.1	48.5	81.4
Media Interview	17.7	71.6	76.9
Daily Life	38.7	83.8	84.9
Meeting	15.1	47.1	80.8
Healthcare	81.8	71.6	84.6
Customer Service	32.3	69.7	80.9
Overall	**39.1**	**65.4**	**81.6**
GPT-4.1-nano (*weakest*)	**42.6**	**70.1**	**93.4**
GPT-5 (*best*)	**81.7**	**80.5**	**96.4**

Table 11: Performance of human-created reference summaries. Human summaries substantially underperform SOTA LLMs and even trail the weakest small model.

ment, achieving Fleiss' Kappa coefficients of 0.86 for key fact matching (computed by flattening the binary mapping matrices of key facts and summary sentences) and 0.63 for summary fact verification.

G Human Reference Analysis

We further re-evaluate the original human-written reference summaries using the same bidirectional fact-checking framework. As shown in Table 11, these references are substantially outperformed by frontier reasoning models, and even lag behind the weakest small instruction model across all three dimensions. This finding suggests that human references in many existing conversation summarization datasets should no longer be treated as reliable upper bounds for evaluating LLM-generated summaries. Many references are sparse, overly compressed, or shaped by dataset-specific annotation conventions, rather than by complete, concise, and faithful coverage of the source conversation. Therefore, reference-based similarity metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) become increasingly inadequate: *they may penalize high-quality LLM summaries that cover more key facts, while rewarding summaries that merely resemble incomplete references*.

The scenario-level results underscore further differences. Human references perform particularly poorly in Meeting and Media Interview, where they miss a large portion of key information. In contrast, Healthcare references remain relatively strong in completeness, highlighting the value of expert annotation and domain-specific summarization protocols. Nevertheless, even in Healthcare, human references still do not consistently surpass modern LLM outputs, reinforcing the need for source-grounded, fact-level evaluation rather than unconditional reliance on human references.

```
[Conversation Context]
<content>
[Conversation End]

**Instruction**: Summarize the above conversation.

Provide your answer in JSON format. Output example:
{"summary": "Your generated summary text here."}
```

Figure 10: Conversation summarization instruction.

```
[Conversation Context]
<content>
[Conversation End]

**Instruction**: Extract key facts from the
conversation provided above.

**Extraction Criteria**: You must extract key facts by
strictly adhering to the following four core
principles:

* **Importance**: Ensure that each key fact represents
the most representative and central information within
the conversation.

* **Atomicity**: Ensure that each key fact is the
smallest indivisible unit of information.

* **Independence**: Ensure that each key fact is self-
contained and fully understandable in isolation.

* **Declarativeness**: Ensure that each key fact is
formatted as a concise, objective declarative sentence
that conveys a core, fundamental truth.

Provide your answer in JSON format. Output example:
{"key_facts": ["fact_1", "fact_2", "...", "fact_n"]}
```

Figure 11: Key fact extraction instruction.

H Instructions

In this section, we detail the complete set of instructions used to interact with the LLMs throughout our methodology. To ensure reproducibility, we provide the specific directives designed for each phase of our evaluation pipeline. These include the zero-shot conversation summarization task, alongside instructions for our human-LLM collaborative framework, such as key fact extraction, self-reflection, atomic fact decomposition, multi-LLM consensus filtering, and bidirectional fact-checking. A comprehensive overview and corresponding figure references are summarized in Table 12.

Instruction	Reference
Conversation Summarization	Figure 10
Key Fact Extraction	Figure 11
Self-Reflection	Figure 12
Summary Fact Decomposition	Figure 13
Hallucinated Fact Filtering	Figure 14
Key Fact Matching	Figure 15
Summary Fact Verification	Figure 16

Table 12: Overview of the instructions.

```
Review the "key facts" you provided in the previous turn. Identify any facts that are actually not critical or essential
(non-key facts).

Provide your answer in JSON format using the following structure:

If non-key facts are found, output: [{"non-key_fact": "original_fact_content", "reason": "concise_explanation"}, ...]

If all facts are still key, output: []
```

Figure 12: Self-reflection instruction.

```
Text:
<content>

**Instruction**: Decompose the above text into atomic facts.

**Decomposition Criteria**: You must perform the decomposition by strictly adhering to the following four core principles:

* **Losslessness**: Ensure that the collection of atomic facts captures all information present in the source text,
without omitting any details.

* **Atomicity**: Ensure that each atomic fact is the smallest indivisible unit of information.

* **Independence**: Ensure that each atomic fact is self-contained and fully understandable in isolation.

* **Declarativeness**: Ensure that each atomic fact is formatted as a concise, objective declarative sentence that conveys
a basic truth.

Provide your answer in JSON format. Output example:
{"atomic_facts": ["fact_1", "fact_2", "...", "fact_n"]}
```

Figure 13: Summary fact decomposition instruction.

```
You are an expert faithfulness checker. Your task is to verify if specific "Facts" are entailed by a provided "Summary".

# Input Data Format:
1. **Summary**: A text string of the summary content.
2. **Facts**: A list of facts to verify, indexed in the format `[index] fact content`.

# Instructions:
1. **Analyze**: Read the summary and the facts carefully.
2. **Verify**: For each fact, determine if it is supported by the summary.
   * **criterion**: base your inference solely on the information provided within the summary.
   * **supported**: the fact is explicitly stated in the summary or can be inferred from it.
   * **not supported**: the fact is missing, contradictory, or unproven in the summary.
3. **Explain**: Provide a concise `reason`.
   * **supported**: briefly explain how the summary proves the fact.
   * **not supported**: briefly explain what information is missing, contradictory, or unproven.

# Output Format:
Provide your answer strictly as a JSON list of dictionaries. Output example:
[
  {
    "fact_idx": <integer>, // the fact index
    "conclusion": "<string>", // "supported" or "not supported"
    "reason": "<string>" // concise explanation
  },
  ...
]

# Input Data:
**Summary**:
<content>

**Facts**:
[0] fact_0
[1] fact_1
...
[n] fact_n
```

Figure 14: Hallucinated fact filtering instruction.

```
You are an expert fact-checker. Your task is to verify if specific "Facts" are entailed by a provided "Summary".

# Input Data Format:
1. **Summary**: A text where each sentence is indexed in the format `[index] sentence content`.
2. **Facts**: A list of facts to verify, indexed in the format `[index] fact content`.

# Instructions:
1. **Analyze**: Read the summary and the facts carefully.
2. **Verify**: For each fact, determine if it is supported by the summary.
   * **supported**: the fact is explicitly stated in the summary or can be inferred from it.
   * **not supported**: the fact is missing, contradictory, or unproven in the summary.
3. **Trace**: If supported, identify the single sentence in the summary that provides the strongest evidence.
   * **supported**: set `align_sentence_idx` to the index of the evidence sentence.
   * **not supported**: set `align_sentence_idx` to `null`.
4. **Explain**: Provide a concise `align_reason`.
   * **supported**: briefly explain how the evidence sentence proves the fact.
   * **not supported**: briefly explain what information is missing, contradictory, or unproven.

# Output Format:
Provide your answer strictly as a JSON list of dictionaries. Output example:
[
  {
    "fact_idx": <integer>, // the fact index
    "align_sentence_idx": <integer or null>, // the evidence sentence index, or null if not found
    "align_reason": "<string>" // concise explanation of the inference or failure
  },
  ...
]

# Input Data:
**Summary**:
[0] sentence_0
[1] sentence_1
...
[m] sentence_m

**Facts**:
[0] fact_0
[1] fact_1
...
[n] fact_n
```

Figure 15: Key fact matching instruction.

```
You are an expert faithfulness checker. Your task is to verify if specific "Facts" are entailed by a provided
"Conversation".

# Input Data Format:
1. **Conversation**: A text enclosed within the tags `[Conversation Context]` and `[Conversation End]`.
2. **Facts**: A list of facts to verify, indexed in the format `[index] fact content`.

# Instructions:
1. **Analyze**: Read the conversation and the facts carefully.
2. **Verify**: For each fact, determine if it is supported by the conversation.
   * **criterion**: base your inference solely on the information provided within the conversation.
   * **supported**: the fact is explicitly stated in the conversation or can be inferred from it.
   * **not supported**: the fact is missing, contradictory, or unproven in the conversation.
3. **Explain**: Provide a concise `reason`.
   * **supported**: briefly explain how the conversation proves the fact.
   * **not supported**: briefly explain what information is missing, contradictory, or unproven.

# Output Format:
Provide your answer strictly as a JSON list of dictionaries. Output example:
[
  {
    "fact_idx": <integer>, // the fact index
    "conclusion": "<string>", // "supported" or "not supported"
    "reason": "<string>" // concise explanation
  },
  ...
]

# Input Data:
[Conversation Context]
<content>
[Conversation End]

**Facts**:
[0] fact_0
[1] fact_1
...
[n] fact_n
```

Figure 16: Summary fact verification instruction.

Daily-Life	[1] Pat has an unexpected work project to finish. [2] Pat will arrive the morning following the conversation. [3] Bart agrees to unlock the door for Pat. [4] Bart can work from home in the morning. [5] Pat is not a painter. [6] Pat is working on a stage project. [7] Pat intends to travel by train.
Media Interview	[1] Jon Hamilton is reporting for NPR on Hurricane Gustav. [2] Jon Hamilton is flying approximately 20 miles from Havana. [3] Jon Hamilton is flying at an altitude of about 10,000 feet. [4] Jon Hamilton is flying in a C-130 cargo plane. [5] The C-130 cargo plane has four engines. [6] The C-130 cargo plane is popular for flying through hurricanes due to its durability. [7] Jon Hamilton has flown through Hurricane Gustav several times during the day of the report. [8] Hurricane Gustav has a very clear eye wall. [9] Jon Hamilton compares the visual structure of Hurricane Gustav to a stadium. [10] Hurricane Gustav has extremely clear architecture. [11] Clear storm architecture is typically found in strong, well-defined storms.
Screenplay	[1] Jesse is a young aspiring model in Los Angeles. [2] Ruby is a make-up artist who befriends Jesse. [3] Ruby introduces Jesse to two models named Sarah and Gigi. [4] Sarah and Gigi are jealous of Jesse's natural beauty and youth. [5] Jesse stays at a motel managed by a man named Hank. [6] A wildcat breaks into Jesse's motel room. [7] Dean is a photographer who assists Jesse after the wildcat incident. [8] Jesse signs with a modeling agency led by Jan Campbell. [9] Jan Campbell sends Jesse to a photographer named Jack for a test shoot. [10] Jack paints Jesse gold during their photo shoot. [11] Sarah sucks blood from a cut on Jesse's hand in a restroom. [12] Jesse books a runway show and becomes the center of attention. [13] Jesse rejects Dean's romantic interest after the runway show. [14] Hank threatens Jesse with a knife at the motel. [15] Jesse flees the motel and seeks refuge at Ruby's house. [16] Ruby attempts to seduce Jesse at Ruby's house. [17] Jesse rejects Ruby's sexual advances. [18] Ruby, Sarah, and Gigi chase Jesse through Ruby's house. [19] Jesse falls into an empty swimming pool while fleeing. [20] Ruby, Sarah, and Gigi kill Jesse. [21] Ruby, Sarah, and Gigi bathe in Jesse's blood. [22] Ruby, Sarah, and Gigi consume Jesse's flesh. [23] Gigi suffers a mental breakdown and physical illness after consuming Jesse. [24] Gigi cuts open her own torso to remove Jesse's remains. [25] Sarah consumes Jesse's eyeball. [26] Sarah believes she has absorbed Jesse's beauty after eating the eyeball.
Customer Service	[1] The customer recently updated their phone. [2] Siri activates randomly on the customer's phone. [3] Music stops playing randomly on the customer's phone. [4] The customer is using the Apple Music app. [5] The customer is using the Spotify app. [6] The customer is connected via Apple headphones. [7] The customer is not connected to any Bluetooth devices. [8] The reported issues only occur when headphones are connected. [9] The customer attempted to restart the iPhone. [10] Restarting the iPhone did not resolve the reported issues.
Healthcare	[1] The patient's mouth has been sore since last week. [2] The patient is experiencing trouble with their eyes. [3] The patient's vision is fine. [4] The patient's diarrhea has stopped. [5] The patient's abdominal pain has stopped. [6] The patient's bowel movements are working well. [7] The patient denies having nausea. [8] The patient denies having vomiting. [9] The patient is eating fine. [10] The patient's appetite has returned. [11] The patient is emptying their bladder properly. [12] The patient denies pain during urination. [13] The patient's cough has not improved. [14] The patient's back pain is still hurting a lot. [15] The patient has pain in the neck. [16] The patient has pain in the right shoulder. [17] The patient's neck and shoulder pain started last week. [18] The patient does not recall injuring themselves last week.
Meeting	[1] Mark is the Project Manager. [2] Dirk Meinfeld is the Marketing Expert. [3] Nick Broer is the User Interface Designer. [4] Xavier Juergens is the Industrial Designer. [5] The project goal is to design a remote control for a television set. [6] The remote control design must be original, trendy, and user-friendly. [7] The product design is inspired by the latest trends in fashion and interior design. [8] The project motto is to put fashion into electronics. [9] The first phase of the project is functional design. [10] Nick Broer draws a dolphin to represent freedom. [11] Dirk Meinfeld draws an elephant to represent being a strong market player. [12] Xavier Juergens draws a tiger to represent speed, efficiency, and security. [13] Mark draws an ant to represent teamwork. [14] Nick Broer suggests the remote control design should be intuitive. [15] Nick Broer suggests considering usability for disabled users. [16] Xavier Juergens suggests the remote control should be made of light material. [17] The target selling price for the remote control is twenty-five Euros. [18] The target production cost for the remote control is twelve and a half Euros. [19] The team is scheduled to reconvene in thirty minutes.

Figure 17: Examples of key facts across six scenarios.

Summary: Patient (Michael) presented after PCP visit for intermittent left-sided back pain (variable sharp/dull, sometimes migratory). Reports increased water intake and nocturia; denies fatigue; appetite unchanged. Active (bikes), in school with good grades. Home exposure: brother smokes. History notable for recurrent strep throat (≈4 episodes in 5 months). Physical exam: normal oropharynx (no erythema/tonsillar swelling), no cervical lymphadenopathy, clear lungs, normal heart sounds, non-tender back and abdomen, no rebound/guarding. Prior CT from PCP shows findings concerning for glomerulonephritis, possibly post-streptococcal, with risk for CKD. Plan: order labs (lipid panel, urinalysis, CMP, CBC, vitamin D, TSH); some require fasting. Refer to ENT to evaluate for recurrent strep and possible tonsillectomy. Counseling on potential salt restriction if kidney issues confirmed. Logistics: patient has class/work today and new job without insurance; physician offered a note to facilitate same-day labs. Follow-up in 1 week for results (phone if normal; in-person if abnormalities). No vaccines/shots today.

Atomic Facts: [1] The patient's name is Michael. [2] Michael presented after a visit to his primary care physician (PCP). [3] Michael's PCP visit was for intermittent left-sided back pain. [4] Michael's back pain is variable. [5] Michael's back pain is sometimes sharp. [6] Michael's back pain is sometimes dull. [7] Michael's back pain is sometimes migratory. [8] Michael reports increased water intake. [9] Michael reports nocturia. [10] Michael denies fatigue. [11] Michael's appetite is unchanged. [12] Michael is active. [13] Michael bikes. [14] Michael is in school. [15] Michael gets good grades in school. [16] Michael has home exposure to a brother who smokes. [17] Michael has a history notable for recurrent strep throat. [18] Michael has had approximately four episodes of strep throat in five months. [19] Michael's physical exam showed a normal oropharynx. [20] Michael's oropharynx showed no erythema. [21] Michael's oropharynx showed no tonsillar swelling. [22] Michael's physical exam showed no cervical lymphadenopathy. [23] Michael's physical exam showed clear lungs. [24] Michael's physical exam showed normal heart sounds. [25] Michael's physical exam showed a non-tender back. [26] Michael's physical exam showed a non-tender abdomen. [27] Michael's physical exam showed no rebound. [28] Michael's physical exam showed no guarding. [29] A prior CT scan was ordered by Michael's PCP. [30] The prior CT scan shows findings concerning for glomerulonephritis. [31] The glomerulonephritis is possibly post-streptococcal. [32] The glomerulonephritis carries a risk for chronic kidney disease (CKD). [33] The plan is to order labs. [34] The labs to be ordered include a lipid panel. [35] The labs to be ordered include a urinalysis. [36] The labs to be ordered include a comprehensive metabolic panel (CMP). [37] The labs to be ordered include a complete blood count (CBC). [38] The labs to be ordered include vitamin D. [39] The labs to be ordered include thyroid-stimulating hormone (TSH). [40] Some of the ordered labs require fasting. [41] The plan includes a referral to an ENT specialist. [42] The ENT referral is to evaluate Michael for recurrent strep. [43] The ENT referral is to evaluate Michael for a possible tonsillectomy. [44] The plan includes counseling on potential salt restriction. [45] The potential salt restriction is conditional on kidney issues being confirmed. [46] Logistically, Michael has class today. [47] Logistically, Michael has work today. [48] Michael has a new job. [49] Michael's new job does not provide insurance. [50] The physician offered Michael a note. [51] The physician's note is to facilitate same-day labs. [52] A follow-up is scheduled in one week for lab results. [53] The follow-up will be via phone if the lab results are normal. [54] The follow-up will be in-person if the lab results show abnormalities. [55] No vaccines were administered today. [56] No shots were administered today.

Summary: Cierra shares a photo of an item she just picked up from the parcel shop; it arrived quickly and looks/feels nice. She offers to bring it to school tomorrow, but Tucker asks to come today. He finishes in about 20 minutes and can arrive in 40–45; Cierra will be home but leaves at 3, and Tucker will try to make it.

Atomic Facts: [1] Cierra shares a photo of an item. [2] Cierra just picked up the item from the parcel shop. [3] The item arrived quickly. [4] The item looks nice. [5] The item feels nice. [6] Cierra offers to bring the item to school tomorrow. [7] Tucker asks to come today. [8] Tucker finishes in about 20 minutes. [9] Tucker can arrive in 40 to 45 minutes. [10] Cierra will be home. [11] Cierra leaves at 3. [12] Tucker will try to make it.

Summary: The user asked for a train from Peterborough to Cambridge on Monday arriving by 18:15 and requested the last train that gets in before that time. After confirming the day, the assistant suggested train TR3021 departing at 05:19. The user agreed and booked five tickets; the assistant confirmed the booking with a total of £82.50 payable at the station and reference CWLKMZER. The user then asked for a multi-sport place in the east and received details: Colville Road, Cherry Hinton, postcode CB1 9EJ, phone 01223 576412. The conversation ended with thanks and well wishes.

Atomic Facts: [1] The user asked for a train from Peterborough to Cambridge on Monday. [2] The user requested the train to arrive by 18:15. [3] The user requested the last train that gets in before 18:15. [4] The assistant confirmed the day. [5] The assistant suggested train TR3021. [6] Train TR3021 departs at 05:19. [7] The user agreed to the suggested train. [8] The user booked five tickets. [9] The assistant confirmed the booking. [10] The total cost for the booking is £82.50. [11] The payment for the booking is to be made at the station. [12] The booking reference is CWLKMZER. [13] The user asked for a multi-sport place in the east. [14] The multi-sport place is located at Colville Road, Cherry Hinton. [15] The postcode of the multi-sport place is CB1 9EJ. [16] The phone number of the multi-sport place is 01223 576412. [17] The conversation ended with thanks. [18] The conversation ended with well wishes.

Figure 18: Examples of decomposed summary facts.