What Are They Talking About? A Benchmark of Knowledge-Grounded Discussion Summarization

Weixiao Zhou[α] **Junnan Zhu**[β] **Gengyao Li**[βγ]
Xianfu Cheng[α] **Xinnian Liang**[δ] **Feifei Zhai**[βϵ] **Zhoujun Li**[α*]
[α]State Key Laboratory of Complex & Critical Software Environment, Beihang University
[β]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[γ]University of Chinese Academy of Sciences [δ]ByteDance Inc. [ϵ]Fanyu AI Laboratory
wxzhou@buaa.edu.cn junnan.zhu@nlpr.ia.ac.cn

Abstract

Traditional dialogue summarization primarily focuses on dialogue content, assuming it comprises adequate information for a clear summary. However, this assumption often fails for discussions grounded in shared background, where participants frequently omit context and use implicit references. This results in summaries that are confusing to readers unfamiliar with the background. To address this, we introduce **K**nowledge-**G**rounded **D**iscussion **S**ummarization (**KGDS**), a novel task that produces a supplementary *background summary* for context and a clear *opinion summary* with clarified references. To facilitate research, we construct the first KGDS benchmark, featuring news-discussion pairs and expert-created multi-granularity gold annotations for evaluating sub-summaries. We also propose a novel hierarchical evaluation framework with fine-grained and interpretable metrics. Our extensive evaluation of 12 advanced large language models (LLMs) reveals that KGDS remains a significant challenge. The models frequently miss key facts and retain irrelevant ones in background summarization, and often fail to resolve implicit references in opinion summary integration.[1]

1 Introduction

Dialogue summarization aims to distill key topics and interactions from a dialogue into a concise summary (Kirstein et al., 2024; Rennard et al., 2023; Jia et al., 2023). Conventional paradigm relies primarily on dialogue content, whether in benchmark construction (Gliwa et al., 2019; Chen et al., 2021; Zhu et al., 2021), methodologies (Zhou et al., 2023; Tian et al., 2024; Lu et al., 2025; Zhu et al., 2025), or evaluation (Wang et al., 2022; Gao and Wan, 2022; Zhu et al., 2023; Tang et al., 2023, 2024b; Liu et al., 2024b; Ramprasad et al., 2024). These efforts implicitly assume that "*the dialogue itself contains sufficient information to generate a clearly understandable summary for readers.*"

However, we find this assumption has fundamental limitations and often fails, particularly when *participants discuss shared background knowledge they are already familiar with*. Such discussions exhibit two main traits: (1) Information Omission and Implicit Reference: Participants naturally skip mutually known details and frequently use pronouns or phrases to refer to entities or facts within the contextual background. (2) Personal Opinion: Unlike simple information interactions in general dialogues, these discussions focus on exchanging viewpoints, with participants expressing personal opinions from various perspectives. These characteristics make understanding the discussion heavily reliant on background knowledge. Consequently, traditional dialogue summarization paradigm inherits this dependency, leading to confusion for outside readers unfamiliar with the context, leaving them wondering: "**What are they talking about?**". We present an illustrative example in Figure 1.

To address this problem, we introduce KGDS, a novel task designed to combine shared background knowledge with discussion content to create reader-centered summaries. We argue that a successful KGDS summary must achieve two complementary objectives: (1) to bridge readers' knowledge gaps by providing the necessary background information that supports the discussion; and (2) to present readers with clear participant opinions by clarifying the implicit references within them. Thus, we formalize these requirements by modeling the task output as a **Background Summary** and an **Opinion Summary**. The background summary, which retrieves or condenses relevant background information, can be **either Extractive or Abstractive**, while the opinion summary integrates clarified opinions and is **inherently Abstractive**. In Figure 1, we provide a visual example of our task paradigms.

To advance research in this field, we construct

*Corresponding Author
[1]Our benchmark is available at zhouweixiao/KGDS

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

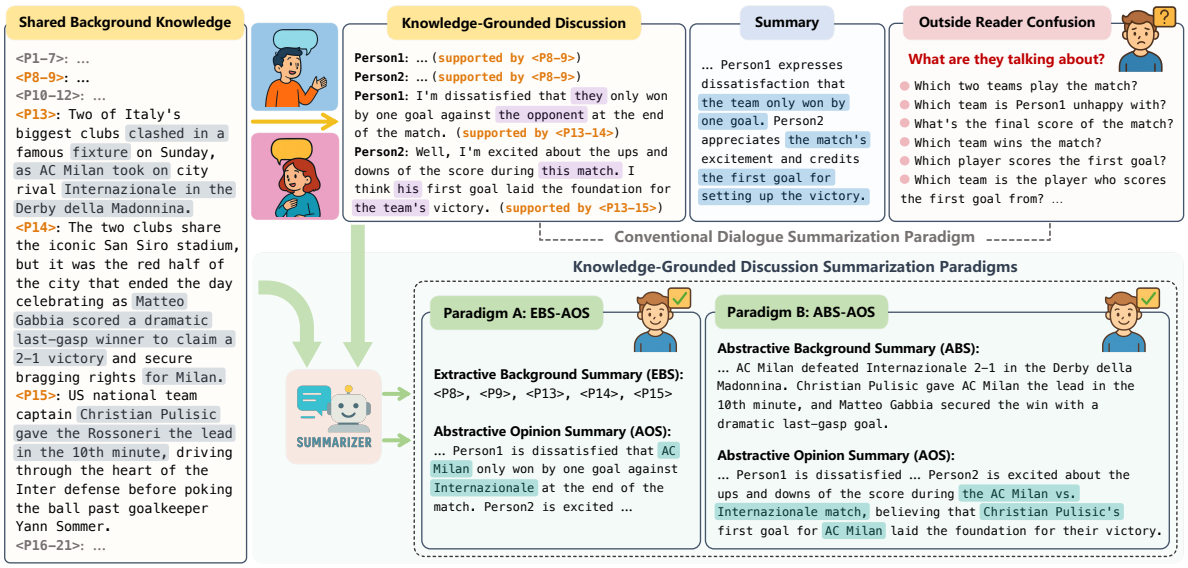

Figure 1: An overview example. **Gray Blocks** in shared background knowledge denote crucial background details omitted by participants during discussion. **Purple Blocks** in discussion content indicate referential pronouns or phrases. **Blue Blocks** in discussion summary represent content that may cause confusion for outside readers. Compared to traditional dialogue summarization, KGDS achieves ***better reader preference*** by providing a supplementary background summary and a clear opinion summary, in which **Cyan Blocks** highlight clarified implicit references.

the first KGDS benchmark, situated in a common and realistic scenario of news discussions. It consists of event-rich news articles, each paired with a human-authored discussion. To enable robust assessment, we develop multi-granularity gold evaluation components, which are annotated by experts under strict consistency controls. For the extractive background summary, we annotate coarse-grained *supporting* and *nonsupporting paragraphs*. For the abstractive background summary, we create two sets of finest-grained atomic facts (Min et al., 2023), including *key supporting* and *nonsupporting facts*. For the opinion summary, we introduce *clear atomic opinions* as basic evaluation units that are minimized and have implicit references clarified.

Furthermore, we propose a novel hierarchical framework for evaluation. For each summarization paradigm, we first evaluate its sub-summaries separately and then aggregate their scores to derive a paradigm-level score. At the sub-summary level, our framework assesses multiple dimensions. For the two types of background summaries, we evaluate their **coverage**, **focus**, and **overall quality** by comparing them against annotated paragraphs or atomic facts, using paragraph index matching or LLM-based fact verification (Wei et al., 2024). For the opinion summary, we assess **overall quality** by checking its coverage of annotated atomic opinions and identify fine-grained **integration errors** by classifying failure types for uncovered opinions. We also conduct human evaluation and show a strong correlation with our automatic metrics.

We evaluate 12 advanced LLMs on our benchmark under the *structured-prompt* (Li et al., 2023) and *self-reflection* (Shinn et al., 2023) settings. Our comprehensive results demonstrate that KGDS remains a substantial challenge, with even the top-performing models achieving an average score below 69%. We also identify several key weaknesses: LLMs struggle with the coverage-focus trade-off in background summary retrieval, frequently miss key facts during background summary generation, and fail to resolve implicit references in opinion summary integration. Furthermore, the limited effectiveness of self-reflection indicates that current LLMs lack sufficient self-correction abilities for this task. These findings highlight key bottlenecks and provide concrete directions for future advancements in coarse-grained retrieval, fine-grained generation, and knowledge integration.

2 Task Formulation

Let K denote the shared background knowledge among participants and D represent the discussion grounded in K. The objective of KGDS is to provide a supplementary background summary and a clear opinion summary by integrating K and D. We define two summarization paradigms based on the type of background summary.

EBS-AOS Paradigm. The output comprises an extractive background summary (EBS) and an ab-

stractive opinion summary (AOS):

$$B_e, O \leftarrow f(K, D),\ B_e \subseteq K \quad (1)$$

Here, f is the summarizer. B_e is the EBS, defined as extractive background supporting chunks for D from K. Chunks are text sequences of a predefined granularity (e.g., sentences, paragraphs, etc.). O is the AOS, defined as clear personal opinions of the participants with clarified implicit references.

ABS-AOS Paradigm. The output contains an abstractive background summary (ABS) and an abstractive opinion summary (AOS):

$$B_a, O \leftarrow f(K, D) \quad (2)$$

The definitions of f and O follow Eq. (1). B_a, the ABS, is defined as abstractive background supporting information for D from K.

3 Benchmark Construction

3.1 Preliminary

Scenario Setting. We establish our benchmark scenario as a two-participant discussion of news content for two reasons. First, news discussions are highly prevalent in daily life, making them more representative than private scenarios such as internal meetings and medical consultations. Second, news summarization (Goyal et al., 2022; Zhang et al., 2024; Liu et al., 2024a) is a well-established subfield of automatic summarization research.

News Collection. We collect 100 multi-domain (i.e., business, sports, and world) event-rich news articles from Google News[2] as shared background knowledge, with a time cutoff of Oct. 2024. We preserve the original news paragraph structure and define ***paragraph-as-chunk*** as the minimum extraction granularity under the EBS-AOS paradigm.

Expert Annotators. To ensure high-quality data, we recruit four PhD candidates specializing in NLP for our annotation tasks. Each pair of experts forms a collaborative group to conduct the full-process annotation, which includes discussion generation and the creation of multi-granularity evaluation components for background and opinion summaries.

3.2 Annotation

Human Discussion. This construction follows the sequence of **read, understand, then discuss**.

Data	Scale	Avg. Tokens
KGDS Task Inputs		
News-Discussion Pair	100	729.5
EBS Evaluation Components		
Supporting Paragraph	432	43.8
Nonsupporting Paragraph	1,005	42.6
ABS Evaluation Components		
Key Supporting Atomic Fact	1,638	11.3
Nonsupporting Atomic Fact	4,996	11.4
AOS Evaluation Components		
Clear Atomic Opinion	873	19.4
Clarified Implicit References	1,113	4.1

Table 1: Benchmark statistics. For three types of subsummaries, we annotate evaluation components at multiple granularities, including coarse-grained paragraphs, fine-grained facts and opinions. The green boxes highlight the average lengths of fine-grained annotations.

For each news article, we require two experts to independently read and thoroughly comprehend its content. This preparatory step aims to ensure background consistency between the participants. Afterward, they engage in a discussion to exchange viewpoints. The entire process is open-ended, meaning the discussion initiator is chosen randomly, and the discussion topics can encompass any events, facts, or detailed information within the news article.

Paragraph-Level Annotation for EBS. An ideal EBS should include all paragraphs that support the discussion and exclude any that do not. To create gold labels for this, two experts independently classify each paragraph from the source news as either *supporting* or *nonsupporting*. We perform a consistency control, retaining only paragraphs where both experts agree on the label. Paragraphs with conflicting annotations are considered *ambiguous* and are removed from the news source to ensure data clarity. Our statistics show that out of a total of 1,696 paragraphs, 1,437 (84.7%) are annotated consistently. Table 1 provides more statistics.

Atomic-Fact-Level Annotation for ABS. Unlike coarse-grained paragraph annotation, a model-generated ABS is condensed and requires fine-grained, key-point-focused labeling. Inspired by the *minimal granularity* of atomic facts (Liu et al., 2023b; Tang et al., 2024a), we argue that an ideal ABS should cover all *key supporting* facts while filtering out *nonsupporting* ones. Motivated by this, we create these two distinct atomic fact sets.

First, we utilize GPT-4o[3] to decompose supporting paragraphs into candidate atomic facts based

[2] https://news.google.com

[3] gpt-4o-2024-08-06

Person1 thinks that it was wise of **Rees-Zammit** (**him**) to sign with **the Kansas City Chiefs** (**this team**).
Person2 thinks that worrying about **Jersey's** (**their**) economic diversification is an overconcern.

Table 2: Examples of clear atomic opinions. Each opinion is an indivisible fine-grained basic unit with clarified implicit references. The content in parentheses indicates the original referential pronouns and phrases.

on principles of indivisibility, independence, and declarativity (see Appendix G.2). Two experts then independently classify each fact as either key or non-key. Through a consistency check, only facts unanimously labeled as key by both experts are considered as the gold standard, resulting in 1,638 key supporting atomic facts.

Second, we use the same decomposition method to nonsupporting paragraphs to generate an initial set of nonsupporting facts. From this set, we then mask[4] any *conflicting* facts, defined as those that could also be inferred from supporting paragraphs. Such conflicts typically arise when identical or similar facts appear in both supporting and nonsupporting paragraphs at the atomic granularity level. For instance, two events have the same timestamp, but one is the background event while the other is not. By masking these facts, we ensure that the key supporting and nonsupporting fact sets remain non-overlapping, thereby avoiding external verification issues during evaluation. This process yields 4,996 nonsupporting atomic facts, with conflicting facts proving to be sparse at just 176 instances (3.39%).

Atomic-Opinion-Level Annotation for AOS. According to the task setup, an effective AOS must clearly present participant opinions by clarifying implicit references. To enable fine-grained evaluation, we introduce the *clear atomic opinion* as the basic unit, which is both *minimal* and has *implicit references clarified*. Table 2 provides two examples. We argue that the quality of an AOS can be assessed by measuring its coverage of these atomic opinions. For annotation, we first require experts to extract the main opinions from their respective utterances in the discussion. They then identify referential pronouns and phrases within these opinions and clarify them through anaphora resolution or information supplementation to produce clear opinions. Finally, the experts decompose these clear opinions into atomic opinion units, following the principles of indivisibility and independence. This annotation process is fully expert-authored. We create 873 clear atomic opinions, among which 800 contain clarified implicit references, while 73 require no clarification. Table 1 provides statistics.

[4]The masking process is consistent with automatic fact verification (Tang et al., 2024a), and the masked conflicting facts are not considered in ABS evaluation.

4 Evaluation Framework

Our framework is comprehensive and hierarchical. For each of the two summarization paradigms (§2), we independently evaluate the sub-summaries and then aggregate their quality to assess overall performance at the paradigm level. The evaluation methods are interpretable, and the metrics are fine-grained. Figure 2 shows an example. Below, we first describe the evaluation dimensions[5] at the sub-summary and paradigm levels (§4.1), and then introduce our automatic methods and metrics (§4.2).

4.1 Dimensions

Background Summary. For the two variants of background summaries, we measure the following dimensions of system responses:

- **Coverage**: Whether the model-extracted or -generated summary fully covers the supporting information (i.e., paragraphs or key facts).
- **Focus**: Whether the summary focuses on supporting information while excluding nonsupporting information.
- **Overall Quality**: A holistic assessment of the summary, considering both coverage and focus.

Opinion Summary. We evaluate the output performance according to the following dimensions[6]:

- **Overall Quality**: Whether the summary successfully integrates clarified opinions, enabling it to cover all clear atomic opinions.
- **Integration Error**: For uncovered clear atomic opinions, identify the errors in the summary.

Summarization Paradigm. We assess the overall quality of each summarization paradigm:

- **Overall Quality**: How well the background and opinion summaries work together.

[5]We do not focus on *fluency* and *coherence* because current LLMs perform well on these dimensions (Song et al., 2025).

[6]Unlike background summaries, opinion summaries are inherently *integrative*. The space of opinions that could be wrongly integrated is effectively unbounded, making it infeasible to construct an exhaustive set of all incorrect atomic opinions to be excluded. Therefore, measuring a **Focus** dimension is impractical. Instead, we identify integration errors in the summary for uncovered clear atomic opinions.

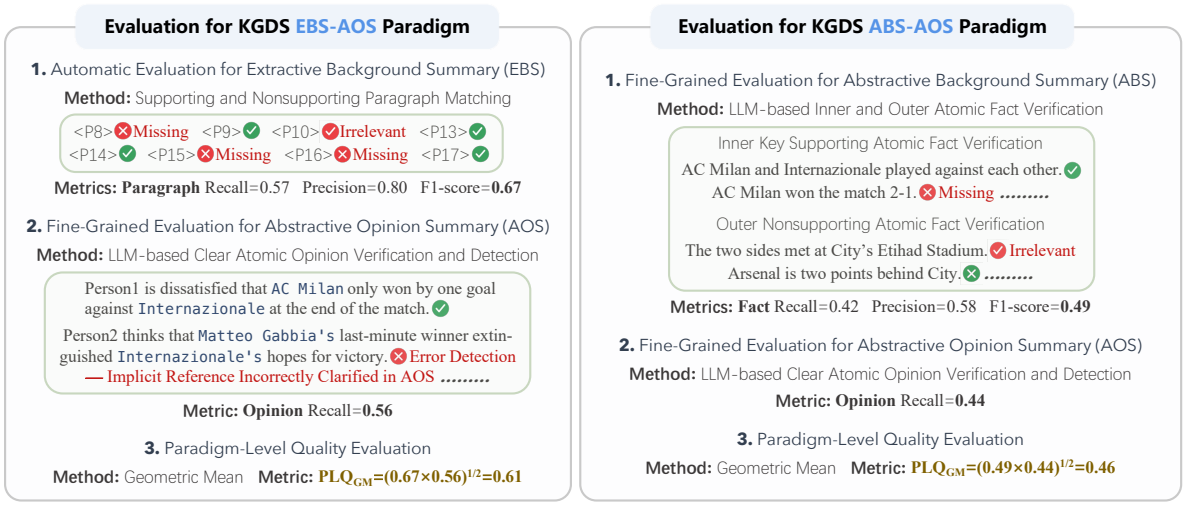

Figure 2: An overview example of our evaluation framework. It comprehensively and accurately evaluates sub-summaries and summarization paradigms through fine-grained, interpretable metrics and hierarchical aggregation.

4.2 Methods and Metrics

EBS Evaluation. We evaluate its quality by matching the paragraph indices from the system output against our annotations of supporting and nonsupporting paragraphs. We utilize **Supporting Paragraph Recall, Precision, and F1-score** to measure the coverage, focus, and overall quality of the summary, respectively.

ABS Evaluation. We employ an LLM-based verifier (Wei et al., 2024) to check if the summary entails our annotated atomic facts. This process determines whether the summary includes key supporting facts while excluding nonsupporting ones. We use three fine-grained metrics, **Key Supporting Atomic Fact Recall, Precision, and F1-score**, to quantify the three dimensions of the summary. We provide detailed formulas in Appendix C.1.

AOS Evaluation. Similar to ABS, we utilize an opinion verifier to measure its coverage of our annotated atomic opinions. We utilize **Clear Atomic Opinion Recall** to assess its overall quality. For each uncovered atomic opinion, we perform LLM-based error detection to identify the specific integration failure. Specifically, we define five fine-grained error types: implicit reference unclarified, implicit reference incorrectly clarified, opinion misattribution, opinion fact inconsistency, and opinion sentiment distortion. Detailed formulas and error definitions are provided in Appendix C.2 and G.6.

Paradigm-Level Quality Evaluation. A successful KGDS output requires high quality in both its background and opinion summaries. To capture this dependency, we evaluate the paradigm-level quality using the **Geometric Mean** of the two sub-summary overall scores. The multiplicative nature of this metric ensures a high score is only achieved when both components are strong, reflecting their joint fulfillment of the task goal. Moreover, root normalization maintains dimensional consistency between the paradigm-level score and its parts.

5 Experimental Setup

System Selection. We select 12 LLMs as summarizers, covering the most advanced and lightweight variants[7]: **GPT-4o**, **GPT-4-turbo**, **GPT-4o-mini**, **Claude 3 Opus**, **Claude 3.5 Sonnet**, **Claude 3.5 Haiku**, **Gemini 1.5 Pro**, **Llama-3.1-405B**, **Mistral Large**, **DeepSeek-V3**, **Qwen-Max**, **GLM-4-Plus**. All model sources are listed in Appendix E.

Prompt Engineering. We benchmark LLMs for KGDS under two interaction patterns: (1) Single-turn **Structured-prompt**: A well-structured standard prompt (Li et al., 2023) that includes input content, input definition, task description, output definition, and return format (Appendix G.3). (2) Multi-turn **Self-reflection**: A second-round self-reflection instruction (Shinn et al., 2023) with step-by-step chain-of-thought reasoning (Wei et al., 2022) following the structured prompt (Appendix G.4).

Verifier and Detector. We use GPT-4o[8] to perform fact and opinion verification as well as error detection due to its excellent consistency with human judgment (Song et al., 2024a,b). All prompts are provided in Appendices G.5 and G.6.

[7] Our evaluation began on *January*, 2025, and all LLMs used the latest official API versions available at that time.

[8] `gpt-4o-2024-11-20`

Model Name	KGDS BenchMark (*single-turn structured-prompt and multi-turn self-reflection*)						
	EBS-AOS Paradigm				ABS-AOS Paradigm		
	SP (R-P-F1)			CAO (R)	PLQ (GM)	KSAF (R-P-F1)			CAO (R)	PLQ (GM)
GPT-4o	73.39$_{\uparrow 1.56}$	88.10$_{\uparrow 0.34}$	**78.12**$_{\uparrow 1.27}$	**76.18**$_{\uparrow 0.54}$	**76.24**$_{\uparrow 1.04}$	63.57$_{\downarrow 1.08}$	61.73$_{\uparrow 1.62}$	**58.34**$_{\uparrow 0.60}$	**69.42**$_{\downarrow 0.21}$	**61.09**$_{\uparrow 0.47}$
GPT-4-turbo	73.03$_{\downarrow 7.72}$	87.42$_{\uparrow 1.67}$	77.14$_{\downarrow 3.80}$	72.73$_{\uparrow 0.45}$	73.94$_{\downarrow 1.47}$	46.28$_{\downarrow 5.78}$	54.42$_{\uparrow 5.33}$	47.26$_{\downarrow 1.64}$	58.32$_{\downarrow 1.90}$	48.28$_{\downarrow 0.74}$
GPT-4o-mini	76.23$_{\downarrow 0.30}$	67.71$_{\uparrow 0.31}$	67.92$_{\downarrow 0.03}$	29.08$_{\downarrow 1.00}$	34.24$_{\downarrow 0.97}$	33.75$_{\uparrow 0.48}$	45.07$_{\uparrow 0.52}$	36.51$_{\uparrow 0.52}$	27.74$_{\downarrow 0.41}$	24.64$_{\downarrow 0.21}$
Claude 3 Opus	65.60$_{\uparrow 2.21}$	85.01$_{\downarrow 4.57}$	72.07$_{\downarrow 1.31}$	**75.20**$_{\downarrow 2.19}$	72.09$_{\downarrow 2.45}$	51.10$_{\uparrow 1.03}$	74.59$_{\downarrow 1.13}$	**58.03**$_{\downarrow 0.91}$	**69.95**$_{\downarrow 2.41}$	**60.72**$_{\downarrow 1.29}$
Claude 3.5 Sonnet	80.36$_{\uparrow 2.42}$	87.93$_{\downarrow 0.62}$	**82.33**$_{\uparrow 0.46}$	74.93$_{\downarrow 5.30}$	**77.12**$_{\downarrow 4.45}$	40.74$_{\uparrow 5.84}$	65.18$_{\downarrow 0.79}$	47.75$_{\uparrow 3.51}$	58.55$_{\uparrow 1.29}$	48.83$_{\uparrow 2.60}$
Claude 3.5 Haiku	62.69$_{\downarrow 8.39}$	76.01$_{\uparrow 0.37}$	66.02$_{\downarrow 5.66}$	40.35$_{\downarrow 9.13}$	46.02$_{\downarrow 10.1}$	37.42$_{\downarrow 6.90}$	48.72$_{\downarrow 4.65}$	39.91$_{\downarrow 6.63}$	29.64$_{\downarrow 5.71}$	27.22$_{\downarrow 4.48}$
Gemini 1.5 Pro	84.64$_{\downarrow 4.38}$	82.25$_{\uparrow 4.43}$	**79.73**$_{\uparrow 1.36}$	**76.86**$_{\downarrow 0.53}$	**76.71**$_{\uparrow 0.49}$	50.40$_{\downarrow 4.13}$	52.33$_{\uparrow 0.28}$	48.42$_{\downarrow 2.11}$	**69.09**$_{\downarrow 6.08}$	**54.83**$_{\downarrow 4.45}$
Llama-3.1-405B	79.09$_{\uparrow 2.72}$	77.96$_{\downarrow 1.60}$	75.16$_{\uparrow 0.36}$	64.25$_{\downarrow 0.01}$	67.58$_{\uparrow 0.01}$	38.19$_{\downarrow 1.08}$	58.47$_{\uparrow 7.38}$	43.39$_{\uparrow 1.12}$	52.98$_{\downarrow 2.44}$	43.79$_{\uparrow 0.11}$
Mistral Large	68.81$_{\uparrow 1.40}$	78.55$_{\downarrow 0.18}$	71.07$_{\uparrow 0.89}$	60.82$_{\downarrow 2.62}$	63.63$_{\downarrow 1.59}$	53.91$_{\downarrow 3.05}$	56.78$_{\downarrow 0.32}$	**52.57**$_{\downarrow 1.94}$	46.24$_{\downarrow 1.71}$	46.36$_{\downarrow 2.26}$
DeepSeek-V3	86.98$_{\downarrow 1.19}$	73.83$_{\uparrow 1.37}$	75.66$_{\uparrow 1.07}$	64.98$_{\uparrow 1.84}$	66.50$_{\uparrow 1.61}$	47.64$_{\uparrow 0.65}$	42.12$_{\uparrow 0.60}$	42.17$_{\uparrow 0.65}$	56.02$_{\uparrow 0.77}$	44.09$_{\uparrow 0.72}$
Qwen-Max	74.59$_{\downarrow 5.53}$	79.86$_{\downarrow 1.30}$	73.79$_{\downarrow 2.78}$	60.93$_{\downarrow 1.24}$	64.34$_{\downarrow 1.59}$	46.49$_{\downarrow 0.02}$	53.59$_{\downarrow 0.35}$	46.87$_{\downarrow 0.17}$	45.29$_{\downarrow 1.54}$	40.87$_{\downarrow 0.23}$
GLM-4-Plus	80.28$_{\downarrow 1.03}$	72.81$_{\uparrow 0.61}$	71.17$_{\uparrow 0.34}$	69.34$_{\downarrow 1.73}$	67.55$_{\downarrow 0.43}$	43.50$_{\downarrow 1.77}$	46.93$_{\uparrow 0.43}$	41.64$_{\downarrow 0.61}$	55.80$_{\downarrow 3.17}$	43.54$_{\downarrow 1.26}$

Table 3: Main evaluation results. **SP**, **KSAF**, **CAO**, and **PLQ** represent Supporting Paragraph, Key Supporting Atomic Fact, Clear Atomic Opinion, and Paradigm-Level Quality, respectively. **R**, **P**, **F1**, and **GM** denote Recall, Precision, F1-score, and Geometric Mean, respectively. All reported metrics are macro-averaged (%). ↑ and ↓ indicate performance **increases** and **decreases**, respectively, after self-reflection following the structured-prompt. For each overall metric (i.e., $\mathrm{SP}_{\mathrm{F}_1}$, $\mathrm{KSAF}_{\mathrm{F}_1}$, $\mathrm{CAO}_{\mathrm{R}}$, and $\mathrm{PLQ}_{\mathrm{GM}}$), we highlight the top-3 performing models.

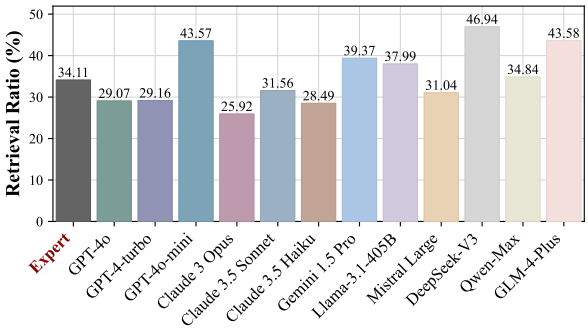

Figure 3: Paragraph retrieval ratios (%) of LLMs. The majority of models can be classified as either conservative ($ratio < 30\%$) or open retrievers ($ratio > 38\%$).

6 Analysis

In this section, we primarily reveal the challenges LLMs face in KGDS and analyze the commonalities and differences among them. Sections §6.1, §6.2, and §6.3 respectively discuss the performance in background summary, opinion summary, and both paradigms under the structured-prompt setting. §6.4 explores the impact of LLM self-reflection.

6.1 Background Summary

LLMs are moderate but imbalanced retrievers for EBS. From Table 3, we find that most LLMs achieve moderate retrieval performance ($\mathrm{SP}_{\mathrm{F}_1} \in [71, 82]$) and lightweight models (i.e., GPT-4o-mini and Claude 3.5 Haiku) perform poorly. However, from SP_{R} and SP_{P}, we observe significant polarization among LLMs, which indicates distinct retrieval strategies: some prioritize precision at the cost of recall (e.g., GPT-4-turbo), while others do the exact opposite (e.g., DeepSeek-V3). Such imbalance reveals that the current LLMs struggle with the *precision-recall trade-off*. Moreover, we investigate the paragraph retrieval ratio[9] (Figure 3) and identify that most LLMs exhibit either under- or over-retrieval, which is consistent with imbalance.

LLMs are inadequate generators for ABS. As presented in Table 3, all LLMs exhibit unsatisfactory performance ($\mathrm{KSAF}_{\mathrm{F}_1} \in [37, 58]$). The low $\mathrm{KSAF}_{\mathrm{R}}$ (average of 46.08%) reveals that LLMs often omit key facts, while the weak $\mathrm{KSAF}_{\mathrm{P}}$ (average of 54.99%) indicates persistent inclusion of irrelevant facts. This dual-failure reflects the fundamental deficiencies of LLMs in meeting the requirements of *coverage* and *focus*. We also observe polarization among LLMs: some prioritize precision at the cost of recall (e.g., Claude 3.5 Sonnet), while others attempt to balance both (e.g., GPT-4o). Unlike EBS, we do not find any extreme recall-oriented models, indicating that LLMs tend to be either conservative or balanced in ABS generation.

6.2 Opinion Summary

Investigating correlation variables influencing AOS quality. From Table 3, we find that $\mathrm{CAO}_{\mathrm{R}}$ decreases as the background summary quality declines (i.e., $\mathrm{SP}_{\mathrm{F}_1} \rightarrow \mathrm{KSAF}_{\mathrm{F}_1}$) across all LLMs when switching paradigms. Furthermore, as shown in Figures 4 and 5, $\mathrm{CAO}_{\mathrm{R}}$ is highly positively correlated with $\mathrm{SP}_{\mathrm{F}_1}$ and $\mathrm{KSAF}_{\mathrm{F}_1}$ among LLMs in both independent paradigms. These findings indicate that *a high-quality background summary facilitates opinion integration*. Meanwhile, individual differences suggest that the *model-inherent integration ability is also a key factor* affecting AOS

[9] Defined as the ratio of the number of retrieved paragraphs to the total number of paragraphs.

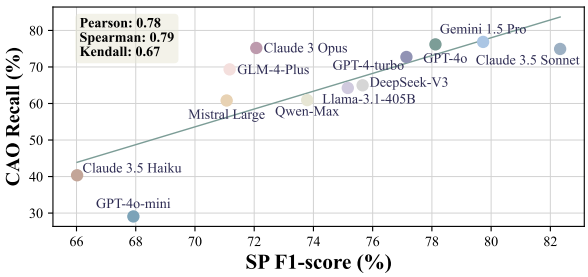

Figure 4: Visualization and metrics of the correlation between $\mathrm{SP}_{\mathrm{F_1}}$ and $\mathrm{CAO}_{\mathrm{R}}$ under the EBS-AOS paradigm.

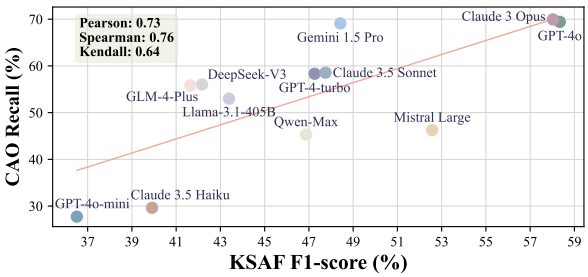

Figure 5: Vis. and metrics of the correlation between $\mathrm{KSAF}_{\mathrm{F_1}}$ and $\mathrm{CAO}_{\mathrm{R}}$ under the ABS-AOS paradigm.

performance. For instance, Claude 3.5 Sonnet and Gemini 1.5 Pro show relatively weaker (i.e., below the line in Figure 4) and stronger (i.e., above the line in Figure 5) integration capabilities in EBS-AOS and ABS-AOS paradigms, respectively.

Nevertheless, even the most advanced models achieve only an average of 72% $\mathrm{CAO}_{\mathrm{R}}$ across both paradigms, indicating that numerous opinions are being incorrectly integrated. Moreover, lightweight models (i.e., GPT-4o-mini and Claude 3.5 Haiku) exhibit significant performance flaws, suggesting that opinion integration is sensitive to model scale.

AOS error distribution analysis. As illustrated in Table 5, a significant ratio of errors is concentrated in IRU and IRIC across all LLMs under both paradigms, indicating that LLMs *struggle to clarify implicit references effectively and correctly* during opinion integration. Moreover, although the ratios of OFI, OSD, and OM errors are relatively lower, their presence still highlights inherent deficiencies such as factual inconsistency, sentiment distortion, and incorrect participant assignment in LLMs.

6.3 Summarization Paradigm

Performance stratification among LLMs. We observe that the overall metric $\mathrm{PLQ}_{\mathrm{GM}}$ exhibits distinct hierarchies in independent paradigms and cross-paradigm (Figures 6, 8, and 9 in Appendix A). This suggests that the performance of LLMs improves in a stepwise manner rather than continuously as their intelligence advances in KGDS. Nevertheless, even the best-performing models achieve less than 69% average performance across both paradigms (see Figure 9), highlighting that KGDS remains a challenging task for current LLMs.

Our evaluation framework aligns well with human judgment. To validate this, we conduct a human evaluation of overall paradigm-level quality[10]. We randomly sample 120 summary pairs (10 per model and 5 per paradigm) from all outputs, and two human evaluators rate their overall quality on a 1-5 Likert scale (Joshi et al., 2015), using the average as the final human score. For comparison, we select three LLMs (i.e., GPT-4o, Llama-3.1-405B, DeepSeek-V3) as LLM-as-a-judge baselines to directly score the summaries, with prompt settings following G-Eval (Liu et al., 2023a). Table 4 presents the Pearson, Spearman, and Kendall correlation coefficients. Our framework achieves significantly higher correlation with human judgments than the baselines, demonstrating the effectiveness of our fine-grained, hierarchical approach.

Metric	Pearson	Spearman	Kendall
LLM-as-a-judge			
GPT-4o	.5273	.5162	.4193
Llama-3.1-405B	.3506	.3760	.2881
DeepSeek-V3	.4282	.4417	.3479
Our Framework	**.6717**	**.6602**	**.5829**

Table 4: Correlation coefficients between different evaluation methods and human judgments, where our framework demonstrates the best alignment ($p < 0.01$).

6.4 Self-Reflection Impact

Self-reflection does not essentially affect the performance of LLMs on KGDS. From Table 3, we observe that the performance fluctuations (i.e., ↑ or ↓) after self-reflection are limited (the maximum average fluctuation across all metrics is 2.30 for $\mathrm{CAO}_{\mathrm{R:ABS\text{-}AOS}}$). This means that self-reflection does not fundamentally influence the capacities of LLMs in KGDS, revealing two key limitations: (1) LLMs lack sufficient self-evaluation abilities for KGDS. (2) The reflection strategies of LLMs struggle to provide excellent reasoning paths for KGDS.

Self-reflection makes LLMs more risk-averse for opinion summary. As presented in Table 5,

[10]We focus on paradigm-level validation because the KGDS task's success depends on the complementarity of both sub-summaries. Furthermore, the individual sub-summary evaluations are already grounded in expert annotations and interpretable, fine-grained verification methods that are known to correlate well with human assessment (Song et al., 2024a).

Model Name	AOS Error Detection (*single-turn structured-prompt and multi-turn self-reflection*)									
	EBS-AOS Paradigm					ABS-AOS Paradigm				
	OFI	**OSD**	**IRU**	**IRIC**	**OM**	**OFI**	**OSD**	**IRU**	**IRIC**	**OM**
GPT-4o	8.57$_{\uparrow 1.09}$	6.19$_{\downarrow 0.88}$	26.67$_{\uparrow 2.80}$	55.71$_{\downarrow 2.09}$	2.86$_{\downarrow 0.92}$	4.07$_{\downarrow 1.12}$	2.96$_{\uparrow 1.10}$	50.74$_{\downarrow 1.66}$	40.37$_{\uparrow 1.33}$	1.86$_{\uparrow 0.35}$
GPT-4-turbo	7.23$_{\downarrow 0.38}$	9.24$_{\downarrow 1.18}$	40.96$_{\uparrow 1.38}$	39.76$_{\downarrow 0.65}$	2.81$_{\uparrow 0.83}$	6.74$_{\uparrow 0.40}$	3.23$_{\uparrow 2.33}$	56.33$_{\downarrow 0.51}$	31.00$_{\downarrow 1.11}$	2.70$_{\downarrow 1.11}$
GPT-4o-mini	0.79$_{\uparrow 0.30}$	0.32$_{\uparrow 0.15}$	81.80$_{\downarrow 0.08}$	16.93$_{\downarrow 0.37}$	0.16$_{\downarrow 0.00}$	1.74$_{\downarrow 0.79}$	0.79$_{\uparrow 0.15}$	73.69$_{\downarrow 0.30}$	23.61$_{\uparrow 1.11}$	0.17$_{\downarrow 0.17}$
Claude 3 Opus	6.96$_{\uparrow 0.54}$	2.17$_{\downarrow 0.09}$	32.17$_{\uparrow 0.75}$	56.96$_{\downarrow 1.96}$	1.74$_{\uparrow 0.76}$	6.61$_{\uparrow 0.64}$	1.18$_{\uparrow 0.99}$	46.69$_{\uparrow 0.05}$	43.19$_{\downarrow 2.61}$	2.33$_{\uparrow 0.93}$
Claude 3.5 Sonnet	6.67$_{\uparrow 0.24}$	5.78$_{\downarrow 1.78}$	50.22$_{\uparrow 2.51}$	35.11$_{\downarrow 1.29}$	2.22$_{\uparrow 0.32}$	3.00$_{\uparrow 3.13}$	3.54$_{\downarrow 0.20}$	61.85$_{\downarrow 8.65}$	29.97$_{\uparrow 5.13}$	1.64$_{\uparrow 0.59}$
Claude 3.5 Haiku	6.81$_{\downarrow 2.01}$	4.09$_{\downarrow 0.83}$	55.45$_{\uparrow 4.07}$	32.10$_{\downarrow 1.23}$	1.55$_{\downarrow 0.00}$	3.91$_{\downarrow 0.10}$	1.47$_{\downarrow 0.41}$	70.52$_{\downarrow 2.64}$	23.94$_{\uparrow 3.31}$	0.16$_{\downarrow 0.16}$
Gemini 1.5 Pro	6.28$_{\downarrow 0.91}$	7.73$_{\uparrow 2.03}$	37.68$_{\uparrow 3.78}$	46.86$_{\downarrow 3.45}$	1.45$_{\downarrow 1.45}$	3.62$_{\uparrow 0.05}$	6.52$_{\downarrow 0.40}$	63.41$_{\uparrow 4.17}$	26.09$_{\downarrow 4.38}$	0.36$_{\uparrow 0.56}$
Llama-3.1-405B	6.92$_{\downarrow 1.00}$	1.89$_{\uparrow 1.23}$	55.35$_{\uparrow 0.10}$	34.59$_{\downarrow 0.95}$	1.25$_{\uparrow 0.62}$	3.83$_{\uparrow 1.18}$	0.47$_{\uparrow 1.81}$	69.62$_{\uparrow 2.13}$	25.36$_{\downarrow 5.77}$	0.72$_{\uparrow 0.65}$
Mistral Large	6.67$_{\uparrow 1.35}$	3.33$_{\downarrow 0.39}$	52.78$_{\downarrow 0.91}$	36.94$_{\downarrow 1.91}$	0.28$_{\uparrow 1.86}$	3.73$_{\downarrow 0.12}$	2.28$_{\downarrow 0.48}$	64.52$_{\uparrow 5.82}$	29.05$_{\downarrow 5.60}$	0.42$_{\uparrow 0.38}$
DeepSeek-V3	4.79$_{\uparrow 1.61}$	2.24$_{\downarrow 0.22}$	53.35$_{\downarrow 5.54}$	38.98$_{\uparrow 4.12}$	0.64$_{\uparrow 0.03}$	5.99$_{\downarrow 0.67}$	2.08$_{\uparrow 0.05}$	61.98$_{\downarrow 2.94}$	28.91$_{\uparrow 3.80}$	1.04$_{\downarrow 0.24}$
Qwen-Max	7.54$_{\uparrow 0.13}$	2.90$_{\uparrow 0.79}$	43.48$_{\uparrow 1.12}$	45.22$_{\downarrow 2.32}$	0.86$_{\uparrow 0.28}$	4.47$_{\uparrow 0.05}$	3.25$_{\downarrow 0.30}$	65.45$_{\uparrow 1.94}$	26.22$_{\downarrow 1.66}$	0.61$_{\downarrow 0.03}$
GLM-4-Plus	10.94$_{\downarrow 1.09}$	3.02$_{\uparrow 1.72}$	40.38$_{\downarrow 5.71}$	44.53$_{\uparrow 4.74}$	1.13$_{\uparrow 0.34}$	5.23$_{\uparrow 2.60}$	1.93$_{\uparrow 1.99}$	55.65$_{\uparrow 0.22}$	35.81$_{\downarrow 4.74}$	1.38$_{\downarrow 0.07}$

Table 5: Fine-grained error detection results. **OFI**, **OSD**, **IRU**, **IRIC**, and **OM** represent the five error types: Opinion Fact Inconsistency, Opinion Sentiment Distortion, Implicit Reference Unclarified, Implicit Reference Incorrectly Clarified, and Opinion Misattribution. All reported metrics are error proportions (%) when clear atomic opinions are not covered by AOS. ↓ and ↑ respectively indicate **decreases** and **increases** in error ratios after self-reflection.

most LLMs exhibit a decrease in IRIC and an increase in IRU (e.g., GPT-4o with IRIC ↓ 2.09, IRU ↑ 2.80). This suggests that self-reflection encourages LLMs to adopt more cautious strategies for clarifying implicit references — *opting to leave them unclarified rather than risk incorrect clarification*. However, such conservatism ultimately harms the quality of opinion summaries, as evidenced by the widespread drop in $\mathrm{CAO_R}$ in Table 3.

7 Related Work

Dialogue Summarization. This task spans diverse scenarios, including online chats (Gliwa et al., 2019), meetings (Hu et al., 2023), and customer service (Lin et al., 2021). Existing methods, whether based on feature modeling (Chen and Yang, 2021; Fang et al., 2022; Lin et al., 2022), pre-training (Zou et al., 2021; Zhong et al., 2022), or LLMs (Wang et al., 2023a; Zhu et al., 2025), focus heavily on single-source dialogue inputs while neglecting shared background knowledge among participants. Moreover, current evaluation systems (Ramprasad et al., 2024; Tang et al., 2024b) also assume single-source inputs. In contrast, our KGDS focuses on multi-source inputs and complementary outputs.

Knowledge-Intensive Dialogue Response. This task aims to enhance the dialogue process by incorporating external knowledge sources (Lewis et al., 2020; Chen et al., 2024). It differs from our KGDS in two key aspects. First, knowledge-intensive dialogue focuses on improving the conversational experience for participants (Wang et al., 2023b, 2024), whereas KGDS is designed to generate clear summaries for external readers. Second, the former addresses knowledge gaps between participants, making the dialogue inherently *knowledge-intensive* (Zhang et al., 2020). In contrast, KGDS centers on discussions among participants with shared knowledge, resulting in *background-sparse* conversations that rely heavily on implicit context.

Summarization Evaluation. Earlier similarity-based metrics like Rouge (Lin, 2004), BERTScore (Zhang et al., 2019), and MoverScore (Zhao et al., 2019) are often misaligned with human judgment. Recently, LLM-based evaluators (Chen et al., 2023; Shen et al., 2023; Fu et al., 2024) have been used, yet they still lack interpretability. To address this, some studies (Song et al., 2024a; Lee et al., 2024; Yang et al., 2024; Scirè et al., 2024) utilize atomic facts (Min et al., 2023) and LLM-based claim verification (Song et al., 2024b), achieving more fine-grained evaluation. Among these, FineSurE (Song et al., 2024a) is most relevant to our work. However, its metrics lack consistency in evaluation granularity. In contrast, our coverage and focus metrics are both measured at the atomic fact level, ensuring consistent granularity across dimensions.

8 Conclusion

In this study, we introduce the KGDS task, which aims to create observer-centric summaries by integrating shared background knowledge with discussions. We establish the first benchmark for KGDS and propose a task-aligned hierarchical evaluation framework with fine-grained, interpretable metrics. Our evaluation of 12 leading LLMs reveals significant challenges in background summary retrieval, generation, and opinion summary integration, with even the most advanced models achieving less than 69% average performance across both paradigms.

Limitations

Our study presents several limitations that merit discussion and future exploration:

Domain Generalization. Knowledge-grounded discussions and their confusing summaries are prevalent in both open-domain and private scenarios. Our benchmark is centered on open-domain news discussions, which may limit its applicability to private scenarios (*e.g.,* internal meetings, medical consultations, legal debates, etc.). However, shared background knowledge and the related discussions in private contexts are challenging to obtain and are less common and representative than our open benchmark. Future work will focus on validating the performance of LLMs in handling KGDS across diverse private scenarios.

Language Diversity. Our current research is focused on English, creating a monolingual benchmark that may not fully address the challenges of multilingual KGDS. This limitation could impact model performance in languages with different linguistic structures or sociocultural contexts. Nevertheless, English is the most prevalent language in both academic research and real-world LLM applications, making it a reasonable starting point. We plan to explore multilingual and cross-lingual KGDS in future research.

Ethics Statement

Our KGDS benchmark is developed with careful consideration of ethical implications. The news articles used as shared background knowledge are publicly accessible through Google News and have been carefully reviewed to ensure they do not contain sensitive or harmful information. Explicit consent was obtained from all expert participants involved in discussion construction and data annotation, with a clear explanation of the research purpose and data usage protocols. Our evaluation methodology prioritizes factual accuracy and opinion fidelity to minimize the risk of hallucination propagation in real-world applications. Potential misuse risks, such as generating misleading summaries through improper background-discussion combinations, are mitigated through technical safeguards in our released code and explicit usage guidelines. Researchers using our KGDS benchmark should adhere to responsible AI principles, especially when applying similar techniques to sensitive domains like healthcare or legal discussions.

Annotation and Evaluation Cost. Our annotation process is multi-step, fine-grained, and highly complex. For each sample in our benchmark, the average annotation time is 2.6 hours (including news reading and understanding, discussion construction, supporting paragraphs, key supporting atomic facts, nonsupporting atomic facts, and clear atomic opinions annotation). We pay each annotation expert a wage of $9 per hour (above the minimum wage standard), and the total annotation cost for the KGDS benchmark is $4680 (2.6 hours per sample $\times$ $9 per hour $\times$ 2 experts per sample $\times$ 100 samples $=$ $4680). The evaluation cost of API calls is detailed in Table 6.

Project	Cost ($)
Atomic Fact Decomposition	21
Conflicting Fact Masking	16
Structured Prompt	74
Reflection Instruction	126
Atomic Fact Verification	173
Atomic Opinion Verification	158
Fine-Grained Error Detection	89
Total	657

Table 6: API call costs for the KGDS benchmark.

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). We thank the anonymous reviewers for their feedback that helped improve this work.

References

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. 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.

Yue Fang, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Bo Long, Yanyan Lan, and Yanquan Zhou. 2022. From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3859–3869, Seattle, United States. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. 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.

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, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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.

Qi Jia, Yizhu Liu, Siyu Ren, and Kenny Q Zhu. 2023. Taxonomy of abstractive dialogue summarization: Scenarios, approaches, and future directions. *ACM Computing Surveys*, 56(3):1–38.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2024. Cads: A systematic literature review on the challenges of abstractive dialogue summarization. *arXiv preprint arXiv:2406.07494*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023. Structured chain-of-thought prompting for code generation. *Preprint*, arXiv:2305.06599.

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

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

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.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024a. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Yong Liu, Shenggen Ju, and Junfeng Wang. 2024b. Exploring the potential of chatgpt in medical dialogue summarization: a study on consistency with human preferences. *BMC Medical Informatics and Decision Making*, 24(1):75.

Yen-Ju Lu, Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Jen-Hao Rick Chang, Yin Xia, Xiang Kong, Qi Zhu, Simon Wang, Oncel Tuzel, and 1 others. 2025. Mutual reinforcement of llm dialogue synthesis and summarization capabilities for few-shot dialogue summarization. *arXiv preprint arXiv:2502.17328*.

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.

Sanjana Ramprasad, Elisa Ferracane, and Zachary C Lipton. 2024. Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends. *arXiv preprint arXiv:2406.03487*.

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

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.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

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. *Preprint*, arXiv:2410.13116.

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 Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, and 1 others. 2024b. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

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.

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

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

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023b. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore. Association for Computational Linguistics.

Jiaan Wang, Jianfeng Qu, Kexin Wang, Zhixu Li, Wen Hua, Ximing Li, and An Liu. 2024. Improving the robustness of knowledge-grounded dialogue via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19135–19143.

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

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

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.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

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.

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.

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

A Paradigm-Level Performance Visualization

Figure 6 and Figure 8 illustrate the overall performance of each model under the EBS-AOS and ABS-AOS paradigms, respectively. Figure 9 presents the average performance of each model across both paradigms. Figure 7 reveals the cross-paradigm stability of each model by quantifying the performance gap between the two paradigms.

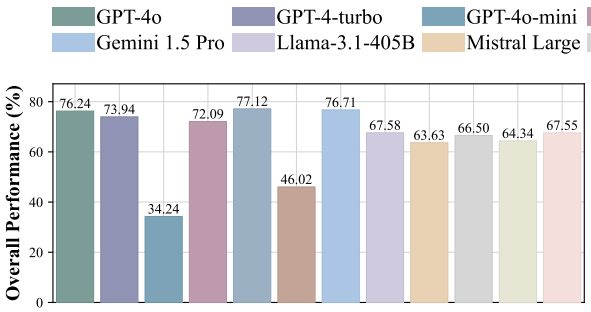

Figure 6: Overall performance under the EBS-AOS paradigm. Models are stratified into three tiers: TIER-1 ([72, 77]), TIER-2 ([64, 68]), and TIER-3 ([34, 46]).

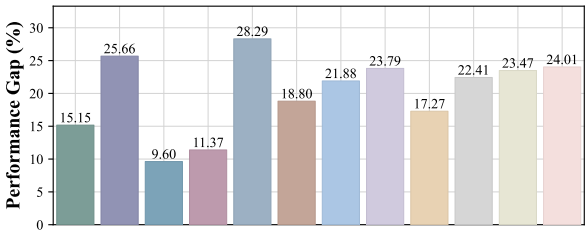

Figure 7: Overall performance gap between the two paradigms. A larger gap indicates lower stability.

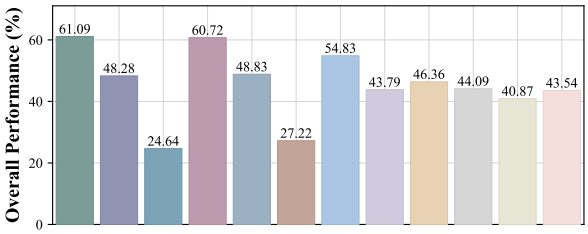

Figure 8: Overall performance under the ABS-AOS paradigm. Models are stratified into three tiers: TIER-1 ([55, 61]), TIER-2 ([41, 49]), and TIER-3 ([25, 27]).

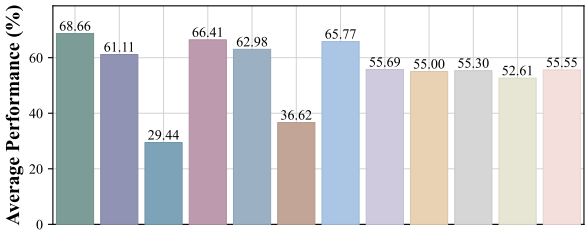

Figure 9: Average overall performance across both paradigms. Models are stratified into three tiers: TIER-1 ([61, 69]), TIER-2 ([53, 56]), and TIER-3 ([29, 37]).

B Further Analysis

LLMs are better retrievers than generators for background summary. As presented in Table 3, all LLMs exhibit higher SP metrics compared to their KSAF counterparts. This gap between retrieval and generation indicates that LLMs possess stronger in-context recognition capabilities for coarse-grained paragraphs than fine-grained facts.

Cross-paradigm stability of LLMs. By analyzing the performance gap between the two summarization paradigms (see Figures 6, 7, and 8 in Appendix A), we find that different LLMs excel in different paradigms. For example, Claude 3.5 Sonnet exhibits a significant gap (28.29%), while Claude 3 Opus demonstrates a relatively smaller gap (11.37%), indicating stronger stability.

LLMs perform better with EBS-AOS than ABS-AOS for KGDS. As presented in Table 3, all LLMs achieve superior $\mathrm{PLQ}_{\mathrm{GM}}$ in the EBS-AOS paradigm compared to ABS-AOS, due to the more complete and accurate background summaries and higher-quality opinion summaries. Therefore, we suggest prioritizing EBS-AOS in real-world KGDS for more effective implementation.

Self-reflection makes LLMs more conservative or open for background summary. As shown in Table 3, most LLMs demonstrate polarization in the increases and decreases between SP_{R} and SP_{P}, as well as $\mathrm{KSAF}_{\mathrm{R}}$ and $\mathrm{KSAF}_{\mathrm{P}}$. This suggests different reflection strategies: some prioritize precision (e.g., DeepSeek-V3 with $\mathrm{SP}_{\mathrm{R}} \downarrow 1.19$, $\mathrm{SP}_{\mathrm{P}} \uparrow 1.37$), while others prioritize recall (e.g., Claude 3 Opus). A few models exhibit simultaneous increases (e.g., GPT-4o with $\mathrm{SP}_{\mathrm{R}} \uparrow 1.56$, $\mathrm{SP}_{\mathrm{P}} \uparrow 0.34$) or decreases (e.g., Qwen-Max), indicating more balanced or weaker reflection abilities.

C Detailed Formulas of Evaluation Metrics

C.1 ABS Evaluation Metrics

Let $\mathcal{B}$ denote the LLM-generated ABS, and let $\mathcal{K} = \{k_i\}_{i=1}^{m}$ and $\mathcal{N} = \{n_j\}_{j=1}^{p}$ represent the sets of m key supporting and p nonsupporting atomic facts, respectively. We define the fact verification function ϕ as:

$$\phi : (\mathcal{B}, f) \rightarrow \{0, 1\}, \ f \in \mathcal{K} \cup \mathcal{N} \tag{3}$$

where $\phi(\mathcal{B}, f) = 1$ if the fact f can be inferred from $\mathcal{B}$, and 0 otherwise.

KSAF Recall. This metric quantifies the **coverage** of key supporting atomic facts in the ABS:

$$\mathrm{KSAF}_{\mathrm{R}} = \frac{1}{m} \sum_{i=1}^{m} \phi(\mathcal{B}, k_i) \tag{4}$$

KSAF Precision. This metric measures the **focus** of the ABS on key supporting atomic facts:

$$\mathrm{KSAF_P} = \frac{\sum_{k \in \mathcal{K}} \phi(\mathcal{B}, k)}{\sum_{f \in \mathcal{K} \cup \mathcal{N}} \phi(\mathcal{B}, f)} \tag{5}$$

KSAF F1-score. This metric evaluates the **overall quality** of the ABS by comprehensively considering both coverage and focus:

$$\mathrm{KSAF_{F_1}} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{6}$$

C.2 AOS Evaluation Metrics

Let $\mathcal{O}$ denote the LLM-generated AOS, and let $\mathcal{C} = \{c_i\}_{i=1}^{n}$ represent the set of n clear atomic opinions. We define the opinion verification function ψ as:

$$\psi : (\mathcal{O}, c) \rightarrow \{0, 1\},\ c \in \mathcal{C} \tag{7}$$

where $\psi(\mathcal{O}, c) = 1$ if the opinion c can be inferred from $\mathcal{O}$, and 0 otherwise.

CAO Recall. This metric quantifies the **coverage** of clear atomic opinions in the AOS:

$$\mathrm{CAO_R} = \frac{1}{n} \sum_{i=1}^{n} \psi(\mathcal{O}, c_i) \tag{8}$$

D Detailed Benchmark Statistics

News Statistics. Refer to Table 7.

Num.	$\mathrm{Paras_{avg}}$	$\mathrm{Tokens_{avg}}$
100	14.4	617.5

Table 7: Statistics for news. Num. indicates the number of news articles. $\mathrm{Paras_{avg}}$ and $\mathrm{Tokens_{avg}}$ represent the average number of paragraphs and tokens, respectively.

Discussion Statistics. Refer to Table 8.

Num.	$\mathrm{Pts._{avg}}$	$\mathrm{Uttrs._{avg}}$	$\mathrm{Tokens_{avg}}$
100	2.0	4.1	112.0

Table 8: Statistics for discussion. Num. indicates the number of discussions. $\mathrm{Pts._{avg}}$, $\mathrm{Uttrs._{avg}}$, and $\mathrm{Tokens_{avg}}$ represent the average number of participants, utterances, and tokens, respectively.

Annotation Statistics for EBS. Among the 1696 paragraphs from 100 original articles, the expert annotation consistency rate is 84.7%. A total of 1437 paragraphs are retained for the final news articles, while 208 are identified as ambiguous paragraphs and removed. Of the 1437 retained, 432 are annotated as supporting and 1005 as nonsupporting.

Annotation Statistics for ABS. Among the 2428 atomic fact units decomposed from the 432 supporting paragraphs, 1638 are consistently annotated as key supporting atomic facts, 780 as non-key atomic facts, and 10 as intra-repetitive atomic facts.

Among the 5187 atomic fact units from the 1005 nonsupporting paragraphs, 4996 are automatically annotated as nonsupporting atomic facts, 176 as masked conflicting facts, and 15 as intra-repetitive atomic facts.

Annotation Statistics for AOS. A total of 873 clear atomic opinions are expert manually annotated. Of these, 800 contain clarified implicit references, while 73 do not require clarification. Within the 800 opinions, there are a total of 1113 written clarified implicit references.

E LLM Sources

OpenAI[11]: GPT-4o, GPT-4-turbo, GPT-4o-mini

Anthropic[12]: Claude 3 Opus, Claude 3.5 Sonnet, Claude 3.5 Haiku

Google[13]: Gemini 1.5 Pro

Meta[14]: Llama-3.1-405B

Mistral AI[15]: Mistral Large

DeepSeek[16]: DeepSeek-V3

Alibaba[17]: Qwen-Max

ZhipuAI[18]: GLM-4-Plus

F Human Evaluation Instruction

> *You will be given a pair of summaries (a Background Summary and an Opinion Summary).*
>
> *Please evaluate the overall quality of this pair of summaries based on the following three criteria. Provide a single, holistic score from 1 to 5 using the Likert scale (5 = Best, 1 = Worst).*

[11] https://platform.openai.com/docs/api-reference/introduction

[12] https://docs.anthropic.com/en/api/getting-started

[13] https://ai.google.dev/gemini-api/docs

[14] https://www.llmapi.com

[15] https://docs.mistral.ai/api/

[16] https://api-docs.deepseek.com/zh-cn/

[17] https://bailian.console.aliyun.com/

[18] https://www.bigmodel.cn/dev/api/normal-model/glm-4

1. Background Summary: Does the summary provide the necessary context to understand the discussion?

2. Opinion Summary: Does the summary clearly provide the participants' viewpoints with implicit references successfully clarified?

3. Overall Performance: Do the two summaries work together effectively for the readers?

Our human evaluation is conducted by the same PhD candidates who constructed the benchmark. We choose these experts as they possess a deep and consistent understanding of the task's goals, which is already demonstrated by their high agreement during the benchmark creation phase. This evaluation process yields a Cohen's Kappa coefficient of 0.74, indicating substantial inter-annotator agreement.

G Prompts and Instructions

G.1 Unified Parameter Settings

In this work, all 12 evaluated LLMs use consistent parameter settings for all prompts and instructions: `max_token`=4096 and `temperature`=0. No other default parameters are modified.

G.2 Atomic Fact Decomposition Instruction

This instruction is used during the benchmark construction phase to guide a large language model in decomposing news paragraphs into fine-grained, atomic fact units. The instruction specifies three core principles for decomposition: indivisibility, independence, and declarativity. This ensures that the resulting fact units are minimal in granularity and informationally complete (see Figure 10).

G.3 Structured Prompts

These prompts serves as the core prompt for evaluating the performance of different LLMs on the KGDS task. It provides the models with standard inputs, including the shared background knowledge and the knowledge-grounded discussion, and clearly defines the objectives and requirements for the two sub-summaries under both summarization paradigms. See Figures 11 and 12 for EBS-AOS and ABS-AOS paradigms, respectively.

G.4 Self-Reflection Instructions

These instructions are used in the multi-turn self-reflection evaluation setting. After the model generates its initial summary, these instructions guide it to critically review its output, checking whether the generated background and opinion summaries strictly adhere to the task definitions. The instructions require the model to provide a detailed chain of thought in a step-by-step manner and, based on this reflection, generate a refined summary. This process is designed to explore and evaluate the model's self-correction capabilities on the KGDS task. See Figures 13 and 14 for the EBS-AOS and ABS-AOS paradigms, respectively.

G.5 Verification Prompts

These prompts are a core component of our evaluation framework, used to automatically verify the quality of the generated summaries. We utilize a powerful verifier LLM (GPT-4o in our work) to execute these prompts, determining whether the model-generated summary contains the key information from the gold standard. Specifically, the fact verification prompt (see Figure 15) is used to assess the recall and precision of the abstractive background summary against key supporting atomic facts, while the opinion verification prompt (see Figure 16) evaluates the coverage of the abstractive opinion summary against clear atomic opinions.

G.6 Error Detection Instruction

This instruction is used for the fine-grained error analysis of the abstractive opinion summary. When the verifier determines that the summary has failed to cover a gold-standard clear atomic opinion, this instruction is invoked to attribute the failure to one of five predefined, specific error types (e.g., opinion misattribution, implicit reference incorrectly clarified). This approach allows us to gain deep insights into the specific weaknesses of each model in integrating opinions and clarifying references. We provide the instruction in Figure 17.

Please carefully read, deeply understand, and strictly follow the instructions below to decompose the "paragraph" into "fine-grained atomic fact units":

Paragraph:
<Paragraph_x>: {content}

Principles of Fine-Grained Atomic Fact Units:
(1). Indivisibility: Ensure that each "fine-grained atomic fact unit" is minimal and cannot be further decomposed into smaller "fine-grained atomic fact units."
(2). Independence: Ensure that each "fine-grained atomic fact unit" can be understood independently, without relying on other "fine-grained atomic fact units."
(3). Declarativity: Ensure that each "fine-grained atomic fact unit" is a concise declarative sentence that clearly conveys a single basic fact.

Output Return Format:
<Paragraph_x>:
<Fact_1>: ...
...
<Fact_n>: ...

Figure 10: Atomic fact decomposition instruction.

Shared Background Knowledge (SBK):
<Paragraph_1>: ...
...
<Paragraph_n>: ...

Knowledge-Grounded Discussion (KGD):
Person1: ...
Person2: ...
...

The above provides SBK and KGD, where SBK represents the shared background knowledge that participants are familiar with prior to the discussion, KGD denotes the discussion by the participants grounded in either partial or complete content of SBK.

Task:
Please combine SBK and KGD to summarize, the result includes two summaries:
* **Extractive Background Summary**: The definition of this summary is the **extractive background-supporting paragraphs for KGD from SBK**.
* **Abstractive Opinion Summary**: The definition of this summary is the **clear personal opinions of the participants with clarified implicit references**. Here, implicit references represent the utilize of pronouns or phrases in KGD to refer to entities, facts, events, or other types of sub-information within SBK.

JSON Result Return Format:
```json
{
  "Extractive_Background_Summary": [
    "<Paragraph_#>",
    ...
    "<Paragraph_#>"
  ],
  "Abstractive_Opinion_Summary": "Person1... Person2... Person1... Person2..."
}
```

Figure 11: Structured prompt for KGDS EBS-AOS paradigm.

Shared Background Knowledge (SBK):
<Paragraph_1>: ...
...
<Paragraph_n>: ...

Knowledge-Grounded Discussion (KGD):
Person1: ...
Person2: ...
...

The above provides SBK and KGD, where SBK represents the shared background knowledge that participants are familiar with prior to the discussion, KGD denotes the discussion by the participants grounded in either partial or complete content of SBK.

Task:
Please combine SBK and KGD to summarize, the result includes two summaries:
* **Abstractive Background Summary**: The definition of this summary is the **abstractive background-supporting information for KGD from SBK**.
* **Abstractive Opinion Summary**: The definition of this summary is the **clear personal opinions of the participants with clarified implicit references**. Here, implicit references represent the utilize of pronouns or phrases in KGD to refer to entities, facts, events, or other types of sub-information within SBK.

JSON Result Return Format:
```json
{
  "Abstractive_Background_Summary": "...",
  "Abstractive_Opinion_Summary": "Person1... Person2... Person1... Person2..."
}
```

Figure 12: Structured prompt for KGDS ABS-AOS paradigm.

Instruction:
* **Please self-reflect: Do the Extractive_Background_Summary and Abstractive_Opinion_Summary you provided align with their respective definitions?**
* **Please think step by step, provide the two summaries again after self-reflection. Additionally, you need to provide a detailed chain-of-thought for self-reflection.**

JSON Result Return Format:
```json
{
  "Extractive_Background_Summary": [
    "<Paragraph_#>",
    ...
    "<Paragraph_#>"
  ],
  "Abstractive_Opinion_Summary": "Person1... Person2... Person1... Person2...",
  "Chain-of-Thought_for_Self-Reflection": "..."
}
```

Figure 13: Self-reflection instruction for KGDS EBS-AOS paradigm.

Instruction:
* **Please self-reflect: Do the Abstractive_Background_Summary and Abstractive_Opinion_Summary you provided align with their respective definitions?**
* **Please think step by step, provide the two summaries again after self-reflection. Additionally, you need to provide a detailed chain-of-thought for self-reflection.**

JSON Result Return Format:
```json
{
   "Abstractive_Background_Summary": "...",
   "Abstractive_Opinion_Summary": "Person1... Person2... Person1... Person2...",
   "Chain-of-Thought_for_Self-Reflection": "..."
}
```

Figure 14: Self-reflection instruction for KGDS ABS-AOS paradigm.

Abstractive Background Summary:
{summary content}

Question:
After carefully reading and deeply understanding the "Abstractive Background Summary" above, can you know each of the following "Atomic Facts"?

Atomic Facts:
<Fact_1>: ...
...
<Fact_n>: ...

Inference Principle:
You may only use information from the "Abstractive Background Summary" to **infer the knowability of each of the "Atomic Facts".**

Result Return Format:
* Provide your verification results in JSON format.
* The returned JSON is a list that provides the verification result of each atomic fact in order.
* Each verification result includes two keys: "Fact_Index" and "Inference_Conclusion".
* The value of "Fact_Index" is a string that provides the index label of the atomic fact.
* The value of "Inference_Conclusion" is a string that provides a binary conclusion: "knowable" or "unknowable".

JSON Format Example:
```json
[
   {
      "Fact_Index": "<Fact_1>",
      "Inference_Conclusion": "knowable or unknowable"
   },
   ...
   {
      "Fact_Index": "<Fact_n>",
      "Inference_Conclusion": "knowable or unknowable"
   }
]
```

Figure 15: Atomic fact verification prompt. The number of facts (n) dynamically changes at the paragraph-level.

Abstractive Opinion Summary:
{summary content}

Question:
After carefully reading and deeply understanding the "Abstractive Opinion Summary" above, can you know the following "Atomic Opinion"?

Atomic Opinion:
<Opinion>: ...

Inference Principle:
You may only use information from the "Abstractive Opinion Summary" to **infer the knowability of the "Atomic Opinion".**

Result Return Format:
* Provide your verification result in JSON format.
* The returned JSON includes two keys: "Inference_Conclusion" and "Analysis_Reasoning".
* The value of "Inference_Conclusion" is a string that provides a binary conclusion: "knowable" or "unknowable".
* The value of "Analysis_Reasoning" is a string that provides a detailed explanation of why the "Inference_Conclusion" is "knowable" or "unknowable".

JSON Format Example:
```json
{
   "Inference_Conclusion": "knowable or unknowable",
   "Analysis_Reasoning": "..."
}
```

Figure 16: Atomic opinion verification prompt.

Instruction:
Please perform fine-grained opinion error detection: For the "Atomic Opinion" that you have verified as "unknowable" above, conduct error classification on the reasons why it cannot be inferred from the "Abstractive Opinion Summary". For an "Atomic Opinion," you only need to provide one most matching error type from the following five error types:

Error Types:
* **Error Type1: Opinion Misattribution**
Definition of Error Type1: The "Abstractive Opinion Summary" incorrectly attributes the opinion of one participant (Person1 or Person2) in the "Atomic Opinion" to another participant, or mistakenly labels it as a group opinion, making it impossible to establish the correct correspondence between the participant and the opinion, and thus impossible to infer the "Atomic Opinion" from the "Abstractive Opinion Summary".

* **Error Type2: Implicit Reference Incorrectly Clarified**
Definition of Error Type2: When the "Atomic Opinion" contains clarified implicit references (the content highlighted between a pair of double asterisks [format: **xxx or **xxx**] in the "Atomic Opinion" is the clarified implicit reference), if the "Abstractive Opinion Summary" contains corresponding incorrect clarifications or explicit explanations of the implicit references, making it impossible to infer the "Atomic Opinion" from the "Abstractive Opinion Summary".

* **Error Type3: Implicit Reference Unclarified**
Definition of Error Type3: When the "Atomic Opinion" contains clarified implicit references (the content highlighted between a pair of double asterisks [format: **xxx or **xxx**] in the "Atomic Opinion" is the clarified implicit reference), if the "Abstractive Opinion Summary" contains corresponding implicit references (such as pronouns/reference phrases/eclipses, etc.) but fails to clarify or explicitly explain them, making it impossible to infer the "Atomic Opinion" from the "Abstractive Opinion Summary".

* **Error Type4: Opinion Sentiment Distortion**
Definition of Error Type4: The "Abstractive Opinion Summary" is inconsistent with the subjective sentiment expressed by the participant in the "Atomic Opinion", making it impossible to infer the "Atomic Opinion" from the "Abstractive Opinion Summary".

* **Error Type5: Opinion Fact Inconsistency**
Definition of Error Type5: The "Abstractive Opinion Summary" is inconsistent with the objective facts expressed by the participant in the "Atomic Opinion", making it impossible to infer the "Atomic Opinion" from the "Abstractive Opinion Summary".

Please think step by step, provide the detection conclusion after error classification. Additionally, you need to provide a detailed chain-of-thought for error detection.

JSON Result Return Format:
```json
{
   "Detection_Conclusion": "Error Type1 or Error Type2 or Error Type3 or Error Type4 or Error Type5",
   "Chain-of-Thought_for_Error-Detection": "..."
}
```

Figure 17: AOS error detection instruction.