UniSumEval: Towards Unified, Fine-Grained, Multi-Dimensional Summarization Evaluation for LLMs

Yuho Lee[1,*], **Taewon Yun**[1,*], **Jason Cai**[2,†], **Hang Su**[2,†], **Hwanjun Song**[1‡]
[1]Korea Advanced Institute of Science and Technology
[2]AWS AI Labs
songhwanjun@kaist.ac.kr

Abstract

Existing benchmarks for summarization quality evaluation often lack diverse input scenarios, focus on narrowly defined dimensions (*e.g.*, faithfulness), and struggle with subjective and coarse-grained annotation schemes. To address these shortcomings, we create UNISUMEVAL benchmark, which extends the range of input context (*e.g.*, domain, length) and provides fine-grained, multi-dimensional annotations. We use AI assistance in data creation, identifying potentially hallucinogenic input texts, and also helping human annotators reduce the difficulty of fine-grained annotation tasks.With UNISUMEVAL, we benchmark nine latest language models as summarizers, offering insights into their performance across varying input contexts and evaluation dimensions. Furthermore, we conduct a thorough comparison of SOTA automated summary evaluators. Our benchmark data will be available at https://github.com/DISL-Lab/UniSumEval-v1.0.

1 Introduction

Despite the enhanced quality of text summarization by large language models (LLMs), they still face persistent challenges like hallucination, information omission, and verbosity (Fabbri et al., 2022; Laban et al., 2023). This multifaceted nature of text summaries inevitably demands manual evaluation by human experts, a labor-intensive and costly process. To streamline this evaluation process, recent efforts aim to design *human-like* automatic evaluators, such as G-Eval (Liu et al., 2023a) and FineSurE (Song et al., 2024), which achieve a satisfactory correlation with human judgments.

Such evaluators are typically validated by examining their consistency with human judgments on established benchmark datasets, such as FRANK (Pagnoni et al., 2021) and TofuEval (Tang et al., 2024b). Yet, these benchmark datasets have limitations in terms of input diversity, granularity of human annotations, and evaluation dimensions.

Firstly, most existing benchmarks are restricted solely to a *single domain*. The predominant focus is often on the news domain such as SummEval (Fabbri et al., 2021) and AggreFact (Tang et al., 2023). This deficiency constrains the accurate evaluation of automated evaluators by failing to capture diverse input contexts across various domains.

Secondly, there is a lack of datasets that consider varying *input types* and *lengths* simultaneously. While these two factors have a significant impact on the summary quality, existing datasets are often limited to short, non-dialogue texts (Bhandari et al., 2020; Pagnoni et al., 2021; Laban et al., 2022). Without considering these factors, they cannot adequately assess distinct perspectives across different input types and lengths. This includes correctly attributing statements to speakers in dialogue, preventing false information in articles with personally identifiable information (PII) redacted, and pinpointing key information in long texts.

Thirdly, no comprehensive datasets exist for *fine-grained, multi-dimensional* summarization evaluation. Some benchmarks offer fine-grained annotations, such as fact verification at the sentence-level (Pagnoni et al., 2021; Laban et al., 2022; Zhu et al., 2023) and alignment at the key-fact[1] level (Bhandari et al., 2020; Tang et al., 2024b), yet they suffer from either a limited evaluation dimension or coarse-grained human labels.

In this paper, we create UNISUMEVAL in Figure 1, the first one-size-fits-all benchmark for fine-grained, multi-dimensional evaluation of automated evaluators. It includes: **Text Inputs** encompassing nine distinct domains (*e.g.*, news, report,

[*] Equal Contribution.
[†] This work is conducted independently and is not related to the author(s)' position at Amazon.
[‡] Corresponding Author.

[1]A key-fact refers a concise sentence conveying a single key piece of information, with at most 2-3 entities, also known as a semantic content unit (Bhandari et al., 2020).

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3941–3960
November 12-16, 2024

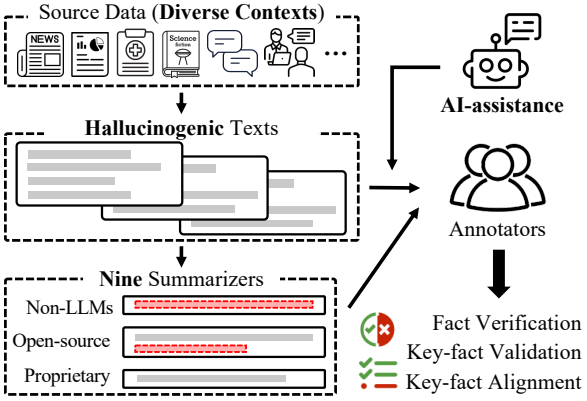


Figure 1: UNISUMEVAL contains fine-grained and multi-dimensional human annotations with high IAA on various input domains, types, and lengths. We conduct AI-assisted manual evaluation on 2,025 hallucinogenic text-summary pairs with 2,509 human key-facts.

booking, meeting) that span from non-dialogue to dialogue, short texts to long texts containing up to 10,462 words, and even with redacted PII to simulate real-world scenarios; **Summaries** generated from nine latest summarizers across three categories, namely non-LLMs, open-source and proprietary LLMs; **Evaluation Dimension** covering three distinct evaluation aspects – assessing faithfulness, information omission (completeness), and verbosity (conciseness) of the generated summaries; **AI-Assisted Manual Evaluation** collecting fine-grained, multi-dimensional human annotations with high IAA[2] – fact verification at the sentence level for faithfulness, key-fact validation and alignment of validated key-facts to each summary sentence for completeness and conciseness.

Unlike existing benchmarks, we identify *hallucinogenic* texts, which can potentially trigger hallucination even for the latest LLMs. We then add these texts into our dataset, ensuring that our benchmark includes challenging text-summary pairs where the latest models actually generate and incur hallucinations. Also, the *wide-ranging* text diversity within UNISUMEVAL enables a comprehensive evaluation of modern automated evaluators. This helps ascertain if the evaluators perform consistently across diverse input scenarios, an aspect overlooked in prior works.

Using UNISUMEVAL's fine-grained human labels, we further benchmark nine latest language models on summarization across *multi-dimensional* aspects. We group them into three distinct categories: non-LLMs, open-source and proprietary LLMs. We then compare their performance across five key evaluation dimensions, including faithfulness, completeness, conciseness, abstractiveness, and domain stability. Next, we unveil the current progress of SOTA automated evaluators, including QA-, NLI-, and LLM-based methods, by comparing their evaluation scores with human judgments in UNISUMEVAL.

Our main contributions are as follows: (1) we are the first to create and release the fine-grained, multi-dimensional benchmark UNISUMEVAL, which covers diverse input contexts; (2) we develop an AI-assisted human evaluation protocol using Amazon Mechanical Turk (MTurk), achieving high IAA comparable to using expert linguistics, even for long input texts; (3) we systematically evaluate latest summarizers on faithfulness, completeness, conciseness, abstractiveness, and domain stability. The performance superiority among them varies depending on input domain, type, and length. PII redaction exacerbates the hallucination issue for all summarization models; (4) we conduct a thorough comparison of SOTA automated evaluators. Non-LLM evaluators perform poorly at verifying LLM-generated hallucinations. Evaluating conciseness (identifying unnecessary summary sentences) is harder than checking for faithfulness and completeness; (5) we release UNISUMEVAL to enable future research on automated evaluation.

2 Related Work

Evaluation Benchmarks. Existing benchmarks in summary quality evaluation have predominantly concentrated on assessing the performance of automated metrics in evaluating faithfulness (Bhandari et al., 2020; Pagnoni et al., 2021; Laban et al., 2022; Tang et al., 2023). These benchmarks have generally been limited to short and non-dialogue texts. This limitation has spurred the development of new benchmarks that specifically address either longer texts (Krishna et al., 2023) or dialogue-based texts (Gao and Wan, 2022; Zhu et al., 2023; Laban et al., 2023; Tang et al., 2024b). Additionally, another segment of benchmarks has expanded beyond the dimension of faithfulness to include relevance and coherence (Fabbri et al., 2021; Gao and Wan, 2022; Tang et al., 2024b), also enhancing the granularity of annotations (Zhu et al., 2023). Recently, more advanced benchmarks have been developed, utilizing the power of LLMs. SummEd-

[2]We obtained the inter-annotator agreement (IAA) of 0.60 (Krippendorff's α for fact verification, 0.88 (Gwet's AC1) for key-fact validation, and 0.58 (Krippendorff's α) for key-fact alignment on average across nine distinct domains.

	Input Text Diversity		Human Annotation Scheme			Summary Generation	
	# of Domains / Input Type	# of Words Avg. (Min – Max)	Granularity	Eval. Dim	Measurement	Summaries from LLMs	Error Source
SummEval	1 / Non-dialogue	408 (106 – 587)	Summary level	Multiple	Likert-scale	No	Realistic
FRANK	1 / Non-dialogue	528 (108 – 1258)	Sentence level	Single	Percentage	No	Realistic
REALSumm	1 / Non-dialogue	745 (227 – 1911)	Key-fact level	Single	Percentage	No	Realistic
SummaC	1 / Non-dialogue	583 (8 – 11,667)	Summary level	Single	Binary-scale	No	Realistic
AggreFACT	1 / Non-dialogue	496 (8 – 11,667)	Summary level	Single	Binary-scale	No	Realistic
LongEval	2 / Non-dialogue	4,917 (1,009 – 12,319)	Key-fact level	Single	Percentage	No	Realistic
DialSummEval	1 / Dialogue	130 (24 – 488)	Summary level	Multiple	Likert-scale	No	Realistic
DiaSumFact	2 / Dialogue	247 (24 – 585)	Sentence level	Single	Percentage	No	Realistic
TofuEval	2 / Dialogue	950 (710 – 1,199)	Mixed level	Multiple	Mixed-scale	Yes	Realistic
SummEdits	9 / Mixed	705 (39 – 2,569)	Summary level	Single	Ternary-scale	Yes	Synthetic
UniSumEval	**9 / Mixed**	**2,092 (21 – 10,462)**	**Sentence & Key-fact**	**Multiple**	**Percentage**	**Yes**	**Realistic**

Table 1: Comparison of UniSumEval with the ten existing summarization evaluation benchmarks. The mixed level of granularity indicates that the evaluation dimensions have annotations at either the sentence or summary level. The mixed level of measurement indicates that a different scale is used for each dimension.

its (Laban et al., 2023) expands seed summaries to non-factual ones by synthetic editing using LLMs but focuses solely on faithfulness. TofuEval (Tang et al., 2024b) generates topic-based summaries using LLMs but limits its scope only to dialogues.

Automated Evaluation. Conventional similarity-based metrics, such as ROUGE-1/2/L (Lin, 2004), BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021) have shown poor correlation with human judgments. In response, natural language inference (NLI)-based methods have emerged to verify the faithfulness of summaries by retrieving relevant evidence in their input texts (Laban et al., 2022; Tang et al., 2024a; Zha et al., 2023). Similarly, Question Answering (QA)-based methods involve generating relevant questions from the reference text and answering them based on the generated content (Fabbri et al., 2022; Zhong et al., 2022). While both directions have shown improved performance, they are generally limited to faithfulness evaluation and also require training specialized models. Recently, LLM-based evaluators have been proposed as reference-free, automated evaluators usable in various contexts (Liu et al., 2023a; Song et al., 2024). While they show promise with short news articles, they still struggle with fine-grained evaluations, and their performance across various domains and input types has not been properly investigated.

3 UniSumEval Pipeline

Our data creation pipeline consists of four consecutive steps in the following sections. Table 1 contrasts UniSumEval with existing benchmarks across various aspects, including input diversity, annotation schemes, and data generation. The statistics of UniSumEval are provided in Appendix A.

3.1 Input Text Sourcing

We use nine source datasets to construct our benchmark dataset: Wikihow (lifestyle) (Koupaee and Wang, 2018), CNN/DM (news) (Nallapati et al., 2016), GovReport (report) (Huang et al., 2021), PubMed (medical literature) (Cohan et al., 2018), SQuALITY (science fiction) (Wang et al., 2022), MultiWOZ (booking conversation) (Zang et al., 2020), DialogSum (daily life conversation) (Chen et al., 2021), MediaSum (interview) (Zhu et al., 2021), and MeetingBank (meeting) (Hu et al., 2023). This selection ensures that each source dataset covers nine distinct domains and maintains a balanced distribution of text types (dialogue, non-dialogue) and lengths (short text, long text).

3.2 Summary Generation and Selection

Summary Generation. We randomly sample 200 input texts from the test set of each source dataset. Then, we generate summaries using the nine latest language models as summarizers, chosen for their widespread usage. These models are classified into three categories: *non-LLMs*, including fine-tuned BART-Large (Lewis et al., 2020) and Flan-T5-Large (Chung et al., 2024), each with fewer than 700M parameters; *open-source LLMs*, including Phi-2 (Javaheripi et al., 2023), Llama2-13B-Chat (Touvron et al., 2023), instruction-tuned Mistral-7B (Jiang et al., 2023) and Mixtral-8x7B (Jiang et al., 2024); and *proprietary LLMs*, including GPT-3.5-turbo, GPT-4-turbo (Achiam et al., 2023), and Claude2.1. See Appendix B.1 for model details and prompts used to generate summaries.

Hallucinogenic Text Selection. The latest models, such as Claude2.1 and GPT-4-turbo, generate hallucinations that are subtle and less common. Hence, to create a more challenging benchmark, we identify *hallucinogenic* texts, which have the potential

to induce hallucination even with the latest models. Specifically, an input text is classified as hallucinogenic if at least one of the nine models we used generates a hallucination from it. We perform an LLM-based automatic evaluation of all text-summary pairs to generate sentence-level binary labels for faithfulness, accompanied by a one-sentence rationale for each label (see Appendix B.2 for details including the prompt). Based on the LLM-based automatic evaluation, we re-sample 25 hallucinogenic texts from the 200 sampled texts for each source. As a result, we finally obtain a total of 2,025 text-summary pairs (= 9 datasets $\times$ 25 hallucinogenic texts $\times$ 9 summarizers).

3.3 Fine-Grained Annotation Tasks

We collect fine-grained human labels for multi-dimensional aspects of summary evaluation. The conventional dimensions like coherence and relevance is not adequate for fine-grained evaluation, due to the ambiguity in their definitions. Thus, we follow the three fine-grained dimensions suggested in the recent work (Song et al., 2024), namely faithfulness, completeness, and conciseness.

Faithfulness is assessed at the sentence level by *fact verification*, a task of assigning a binary label (Yes/No) indicating whether a sentence has factual errors across four predefined categories. These include Out-of-Article Error as an extrinsic error, and Entity Error, Sentence Error, Relation Error as the three subcategories of intrinsic errors (see Appendix C for more details on the error taxonomy). If the response is Yes, we then ask the respondents to identify error types using a multichoice form.

In contrast, completeness and conciseness are annotated at the key-fact level using two different tasks. More specifically, we carefully generate the list of potential key-facts using multiple LLMs[3] and then conduct *key-fact validation*, a human annotation task for verifying if the information in each generated key-fact is significant, factually correct, and relevant with respect to its source text. It enables the identification of human verified key-facts from the generated ones. Next, we perform another annotation task of *key-fact alignment*, matching each human key-fact to all summary sentences from which they can be inferred.

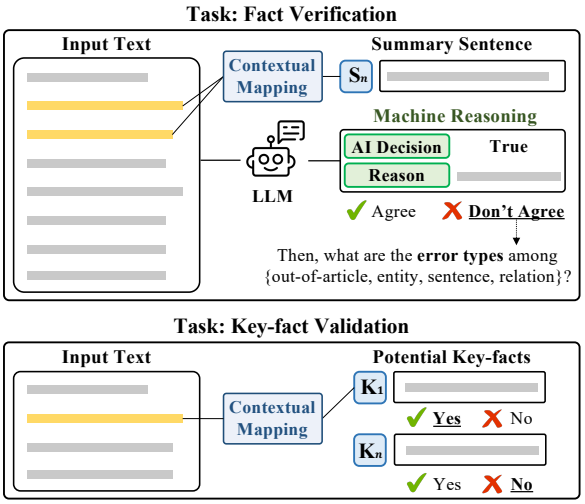


Figure 2: AI-assisted fine-grained manual evaluation.

Across the three annotation tasks, we can compute the percentage (%) scores for three evaluation dimensions at the summary level: (1) *faithfulness*, the proportion of factually correct summary sentences; (2) *completeness*, the proportion of key-facts inferable from the summary; and (3) *conciseness*, the proportion of summary sentences aligned with the key-facts. See Appendix D.1 for the detailed formulas.

3.4 AI-Assisted Manual Evaluation

Unlike key-fact alignment, which only requires annotators to align key-facts and summary sentences, fact verification and key-fact validation tasks necessitate a thorough understanding of the input text. This issue evidently leads to a significant drop in IAA for manual evaluation when input texts are lengthy, as humans struggle to manage large volumes of information simultaneously (Krishna et al., 2023). Hence, we devise AI-assisted manual evaluation in Figure 2, which helps achieve high IAA among non-expert human annotators for long texts.

We apply *contextual mapping*, a task of highlighting sentences in an input text that rank in the top 30% for similarity[3] with a target summary sentence in fact verification (or a key-fact in key-fact validation) being evaluated. This aids annotators by providing contextual cues and narrowing down the relevant sections of the input text. Additionally, for fact verification, we plug in *machine-based reasoning*, the assistance of providing the decision by an automatic evaluator. This is based on the inferred

[3]We generate initial key-facts using GPT-4-turbo with a tuned prompt. These key-facts are then reduced using other LLMs, including GPT-3.5-turbo, Claude2.1, and Mixtral-8x7B, by eliminating those the majority disagree with. This results in an average of 8 and 14 potential key-facts for short and long texts, respectively. See Appendix B.3 for the prompts for the key-fact generation and reduction with multiple LLMs.

[3]We use the pre-trained BERT (Kenton and Toutanova, 2019) to compute the similarity. The choice of a 30% threshold is intentionally set as a conservative value, ensuring high recall to encompass all relevant input sentences.

Evaluation Dim.	Annotation Task	Granularity	Short				Long					Avg. IAA
			News	LifeStyle	Booking	Daily Life	Report	Med Lit	Sci-fi	Interview	Meeting	
Faithfulness	Fact Verification	Sentence level	0.63	0.78	0.66	0.78	0.49	0.42	0.48	0.43	0.65	0.60
Completeness & Conciseness	Key-fact Validation	Key-fact level	0.90	0.96	0.95	0.91	0.98	0.96	0.67	0.84	0.77	0.88
	Key-fact Alignment	Key-fact level	0.61	0.80	0.69	0.71	0.43	0.41	0.47	0.54	0.56	0.58

Table 2: IAA scores of human labels in UNISUMEVAL for each data domain. The subset is categorized as "short" if the average number of words is less than 900, otherwise "long".

sentence-level factuality labels and reasoning obtained in the hallucinogenic text selection. To identify annotators who blindly endorse AI decisions, we replace 20% of the tasks with attention checks, where the correct answers are predetermined. It helps us detect unfaithful responses based on annotators' (dis)agreement with the AI decision. See Appendix G.3 for details on the attention check.

This systematic integration of AI into manual evaluation enables not only to achieve high IAA even for long input texts but also to promote a more cost-effective assessment.

3.5 Annotation Procedure

We conduct AI-assisted manual evaluation using MTruk on our three tasks. We select annotators who pass an English qualification test, with an approval rating above 95% and at least 1,000 accepted HITs. We collect human annotations from three independent qualified annotators for 8,133 summary sentences in fact verification, and for 2,673 key-facts in key-fact validation. The 2,673 potential key-facts are reduced to 2,509 human key-facts. Consequently, annotations for key-fact alignment are collected for all possible pairs between human key-facts and summary sentences of the same input text, *i.e.*, 101,013 pairs in total. Annotators are paid 50% above the U.S minimum wage and receive $25 bonuses for every 500 HITs. See Appendix G for more details on our annotation tasks.

4 UNISUMEVAL Quality Assessment

Evaluation Metric. We report IAA for three fine-grained annotation tasks in UNISUMEVAL: fact verification, key-fact validation and alignment. We use Krippendorff's α (Krippendorff, 2011) by default, but for key-fact validation, where there is significant label imbalance, we use Gwet's AC1 (Wongpakaran et al., 2013) due to its enhanced robustness.

4.1 Inter-Annotator Agreement

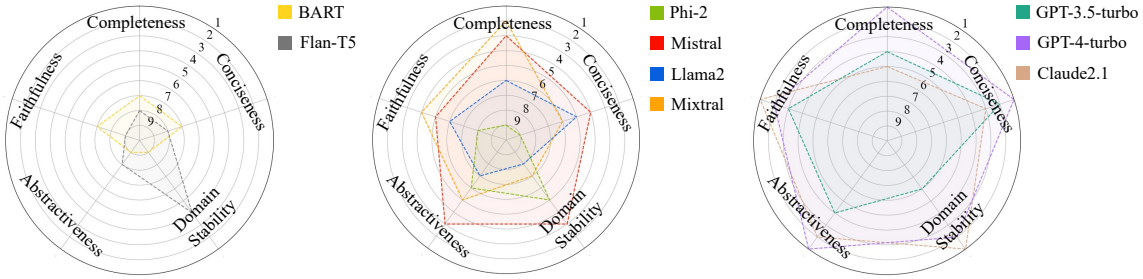
Overall Assessment. Table 2 shows the IAA scores of UNISUMEVAL across nine data domains. In Table 3, we compare UNISUMEVAL with existing benchmarks annotated with fine-grained labels. **UNISUMEVAL stands out as the only benchmark supporting multi-dimensional evaluation of automated evaluators**, while others only focus on either faithfulness (via fact verification) or completeness (via key-fact alignment). Additionally, our benchmark exhibits IAA better than or comparable to others, even with the comprehensive inclusion of varying input contexts, as evidenced by Table 1. Particularly, TofuEval got an IAA of 0.40 with linguistic experts in the meeting domain (long text), while we get a higher IAA of 0.65 in the same domain even with non-expert labels through AI-assisted manual evaluation (see Table 2).

Benchmark	Annotation Task	IAA
FRANK	Sentence level fact verification	0.63
DiaSumFact†	Sentence level fact verification	0.49
TofuEval	Sentence level fact verification	0.40
LongEval†	Key-fact level fact verification	0.64
REALSumm	Key-fact alignment	0.66
UNISUMEVAL	Sentence level fact verification	0.60
	Key-fact alignment	0.58

Table 3: IAA comparison across the existing benchmarks with fine-grained labels. We report Krppendorff's α by default. †: we copy the Fleiss' κ value in the original paper due to the absence of annotator-level labels.

Impact of Input Context. The UNISUMEVAL's input diversity enables a thorough analysis of how IAA varies across different domains in manual evaluation. In Table 2, we note a high Pearson correlation of 0.89 in IAA between fact verification and key-fact alignment tasks across various data domains. Also, input characteristics, such as long-form texts and professional texts including reports, medical literature, science fiction, and interviews, have a negative impact on IAA. Therefore, **input contexts significantly influence the IAA score of fine-grained human annotation tasks.**

Efficacy of AI Assistance. We use LLM-based evaluation for two purposes – selecting hallucinogenic input texts; and assisting annotators in fact verification. For the former, 93.3% of the selected input texts are confirmed to produce real hallucinations in at least one summarizer. For the latter, we



(a) Non-LLMs. (b) Open-source LLMs. (c) Proprietary LLMs.

Figure 3: Performance ranking (1-9) of the nine recent summarizers across the five evaluation dimensions. The summarizers are categorized into three distinct groups: non-LLMs, open-source LLMs, and proprietary LLMs.

Model Type	Summ. Model	Non-Dialogue						Dialogue						Avg. Score (Model-wise)		
		Short			Long			Short			Long					
		Faith	Comp	Conc	Faith	Comp	Conc	Faith	Comp	Conc	Faith	Comp	Conc	Faith	Comp	Conc
Non LLM	$BART_{large}$	87.6	67.9	86.6	89.5	16.3	60.1	81.5	47.8	79.6	77.9	30.9	72.2	84.1	40.7	74.6
	$Flan\text{-}T5_{large}$	90.4	43.5	83.8	83.3	16.3	69.0	72.3	44.5	76.3	75.4	32.2	67.4	80.4	34.1	74.1
Open Source LLM	Phi-2	72.3	56.6	71.1	84.5	16.9	43.6	86.8	36.9	60.8	81.2	25.6	49.5	81.2	34.0	56.3
	$Mistral_{7B\text{-}Inst}$	95.9	77.7	89.3	96.9	43.1	70.4	94.7	66.2	79.6	92.9	59.4	75.1	95.1	61.6	78.6
	$Llama2_{13B\text{-}Chat}$	96.3	61.6	92.6	92.4	29.1	63.6	91.7	40.3	79.9	86.7	35.8	73.1	91.8	41.7	77.3
	$Mixtral_{8x7B\text{-}Inst}$	97.7	80.9	87.7	97.0	44.2	63.6	96.1	68.4	77.1	95.6	60.5	74.0	96.6	63.5	75.6
Prop. LLM	$GPT\text{-}3.5_{turbo}$	95.5	71.7	91.5	97.2	42.4	81.5	98.5	62.8	83.2	93.8	49.3	79.2	96.2	56.6	83.8
	$GPT\text{-}4_{turbo}$	97.3	79.3	90.5	98.1	45.0	82.6	99.0	76.3	86.9	92.6	64.5	81.9	96.8	66.3	85.5
	Claude2.1	96.3	64.2	84.6	99.3	32.0	74.7	98.4	60.4	76.0	95.5	44.8	85.8	97.4	50.4	80.3
Avg. Score (Context-wise)		92.1	67.0	86.4	93.1	31.7	67.7	91.0	56.0	77.7	88.0	44.8	73.1	91.1	50.4	76.2

Table 4: Faithfulness (Faith), completeness (Comp), and conciseness (Cons) scores (%) based on human annotations across different input contexts. Green, orange and red indicate the performance ranking intervals of the model for each dimension, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-9) tiers, respectively.

conduct an ablation study on the fact verification task with three variants: (1) highlighting relevant input sentences through contextual mapping; (2) providing factuality labels estimated by the LLM; and (3) providing the labels with reasoning by the LLM. **The IAA for fact verification improves significantly with AI assistance** from 0.28 to 0.55 by adding (2); and further to 0.63 by adding (3).

In particular, we highlight that human annotators do not indiscriminately accept machine labels. They reveal that 19.31% of the factuality errors flagged by the LLM are inaccurately identified, while 2.59% of sentences deemed error-free by the LLM actually contains factual inaccuracy. See Appendix F.1 for further analysis.

5 Benchmarking Summarizers

Evaluation Dimension. We evaluate the nine summarizers across five crucial evaluation dimensions for text summarization. In addition to *faithfulness*, *completeness*, and *conciseness*, we add two more dimensions: *domain stability*, the consistency of a summarizer's performance across the nine domains; *abstractiveness*, the extent to which a summary generates novel sentences or phrases, leading to a more coherent and condensed summary.

Evaluation Metric. We report percentage scores (in Section 3.3) of faithfulness, completeness, and conciseness, computed by using fine-grained human annotations. For domain stability, we calculate the average of the three percentage scores to obtain a composite score, and then measure domain inconsistency by computing the gap between the highest and lowest composite ones. For abstractiveness, we use the average of novel 1/3/5-grams following Song et al. (2023). See Appendix D.1 for their detailed definitions.

5.1 Comparison over Nine Summarizers

Figure 3 shows the overall performance rankings aggregated across the all text domains, types, and lengths in UNISUMEVAL. The proprietary LLMs notably outperform the non-LLMs and open-source LLMs across the all aspects. GPT-4-turbo is the best summarizer in general, while Claude2.1 has the best faithfulness and domain stability. Proprietary LLMs also exhibit an interesting behavior – **no statistical relationship between faithfulness and abstractiveness** – contradicting recent findings (Maynez et al., 2020; Ladhak et al., 2022) that summaries with higher abstractiveness are more prone to trigger hallucination. Statistical analysis on this can be found in Appendix F.3.

Type	Model	Unredaction	Redaction
Non LLM	$BART_{large}$	82.8	79.0 (-3.8)
	$Flan\text{-}T5_{large}$	78.6	75.7 (-2.9)
Open Source	$Mistral_{7B\text{-}Inst}$	97.9	95.7 (-2.2)
	$Mixtral_{8x7B\text{-}Inst}$	93.9	92.6 (-1.3)
Prop. LLM	$GPT\text{-}3.5_{turbo}$	98.0	97.0 (-1.0)
	$GPT\text{-}4_{turbo}$	100.0	98.0 (-2.0)

Table 5: Impact of PII redaction on faithfulness on booking conversation (source: MultiWOZ).

5.2 Impact of Domain, Type, and Length

Table 4 breaks down the performance in Figure 3, highlighting the impact of input contexts – input domain, type, and length – on each summarizer.

Firstly, **the superiority in summarization quality among the summarizers varies depending on input domain, type, and length.** The general tendency among the three summarizer categories is consistent, while within each category, there are considerable changes in the summarizers' rankings (see the color changes for each column).

Secondly, **most summarizers experience a significant performance drop in completeness and conciseness with lengthy input texts.** Particularly, such a drop is more noticeable when dealing with non-dialogue than dialogue. This confirms that identifying key-facts in summary generation is more challenging with longer input texts.

Appendix F.2 provides the detailed results without the aggregation for the faithfulness, completeness, conciseness, and composite scores.

5.3 Impact of PII Redaction.

We investigate how PII redaction impacts summarization quality using recent summarizers. This is a crucial aspect, because it is very common in industrial use cases, such as call centers and legal service. To construct a redacted dataset, we select a subset of UNISUMEVAL – 25 hallucinogenic texts from the MultiWOZ (booking) data which contain significant amounts of PII-related entities, including phone numbers and addresses. We manually redact the entities by replacing them with their corresponding category name, *i.e.*, <PHONE-NUMBER-1>. On average, eight entities are redacted per input dialogue. See Appendix E for the detailed protocol.

Table 5 shows the faithfulness scores before and after PII redaction into input texts, where the top-2 summarizers are selected from each category. PII redaction negatively affects the faithfulness of the all summarizers. **The non-LLMs are more susceptible to PII redaction than the open-source and proprietary LLMs.** This drop is attributed to filling in the masked entity with an entity either not present in the input text or incorrectly presented.

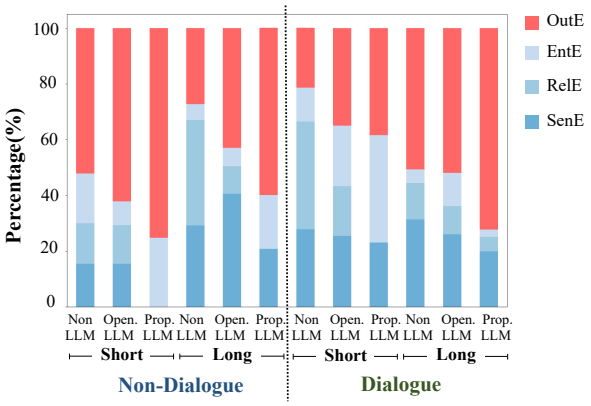


Figure 4: Error distribution by varying input contexts for each summarizer category, showing OutE (out-of-article error), EntE (entity-error), RelE (relation-error), and SenE (sentence-error). Red color indicates extrinsic errors, while blue tones denotes intrinsic errors.

5.4 Factuality Error Analysis

In Figure 4, we examine how the distribution of error types varies across different input contexts for each summarizer group. **The proprietary LLMs exhibit a lower rate of intrinsic errors, while the non-LLMs exhibit a higher rate of them in all input contexts.** Notably, the proprietary LLMs show no relation errors across most context types except in long dialogue texts. This suggests that the higher faithfulness scores of the proprietary LLMs in Table 4 are likely due to their much lower intrinsic error rates compared with the others.

6 Benchmarking Auto-Evaluators

Evaluator Selection. We benchmark SOTA automated evaluators on UNISUMEVAL. The set of compared evaluators varies according to the target evaluation dimension. For faithfulness, we include *QA-based* models: UniEval (Zhong et al., 2022) and QAFactEval (Fabbri et al., 2022), *NLI-based* models: Summac-Conv (Laban et al., 2022), AlignScore (Zha et al., 2023), and MiniCheck (Tang et al., 2024a), and *LLM-based* models at various levels of granularity: G-Eval (Liu et al., 2023a) for summary level, G-Eval+[4] for sentence level and FactScore (Min et al., 2023) for atomic level. For completeness, we include NLI-based models: $Lite^3Pyramid$ (Zhang and Bansal, 2021), A^3CU

[4]We adjusted G-Eval's prompts to align with the granularities and dimensions of UNISUMEVAL, renaming it G-Eval+. See Appendix B.4 for the tuned prompts.

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-based	$\text{UniEval}_{\text{faith}}$	0.11	0.54*	-0.22*	-0.09	-0.26*	0.12	-	0.17*	0.06
	QAFactEval	0.14*	0.45*	-0.07	0.19*	-0.15*	0.22*	0.06	-0.04	0.03
NLI-based	$\text{SummaC}_{\text{Conv}}$	0.07	0.13*	-0.15*	-0.13	-0.08	-0.11	-0.03	-0.21*	-0.08
	AlignScore	0.18*	0.32*	0.17*	0.09	0.26*	0.33*	0.12	0.09	0.38*
	MiniCheck	0.24*	0.69*	-0.02	0.13	-0.10	0.22*	0.30*	0.15*	0.05
LLM-based	$\text{G-Eval}_{\text{faith}}$	**0.65***	**0.72***	0.41*	-0.14	**0.41***	**0.68***	**0.62***	0.48*	**0.60***
	$\text{G-Eval+}_{\text{faith}}$	0.63*	0.57*	**0.46***	**0.55***	0.38*	0.46*	0.59*	**0.52***	0.53*
	$\text{FactScore}_{\text{faith}}$	0.07	-0.02	-0.11	0.10	0.00	-0.06	-0.13	0.10	0.05

Table 6: Agreement with human scores in **faithfulness** evaluation across nine domains (*: p-value < 0.05). For LLM-based methods, G-Eval, G-Eval+, and FactScore are the summary, sentence, and atomic level evaluators.

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-based	$\text{UniEval}_{\text{coh}}$	0.18*	0.04	0.06	0.17*	0.15*	0.05	-	0.19*	0.24*
NLI-based	$\text{Lite}^3\text{Pyramid}$	0.35*	-0.11	0.36*	0.38*	0.57*	0.36*	-	0.25*	0.14*
	A^3CU	0.43*	0.06	0.29*	0.13	0.42*	0.31*	-	0.24*	0.08
LLM-based	$\text{G-Eval}_{\text{coh}}$	0.35*	0.45*	0.47*	0.57*	0.56*	0.41*	0.31*	0.55*	0.63*
	$\text{G-Eval+}_{\text{com}}$	**0.57***	**0.61***	**0.59***	**0.65***	**0.68***	**0.56***	**0.32***	**0.63***	**0.66***

Table 7: Agreement with human scores in **completeness** evaluation across nine domains (*: p-value < 0.05). For UniEval and G-Eval, we use their coherence scores since completeness dimension is not directly supported.

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-based	$\text{UniEval}_{\text{rel}}$	0.06	0.05	0.02	-0.01	0.30*	0.08	-	0.23*	0.24*
LLM-based	$\text{G-Eval}_{\text{rel}}$	0.02	**0.39***	**0.18***	0.09	**0.51***	0.33*	**0.2***	0.46*	0.45*
	$\text{G-Eval+}_{\text{con}}$	0.11	**0.39***	0.17*	**0.24***	0.44*	**0.36***	0.02	**0.49***	**0.45***

Table 8: Agreement with human scores in **conciseness** evaluation across nine domains (*: p-value < 0.05). For UniEval and G-Eval, we use their relevance scores since conciseness dimension is not directly supported.

(Liu et al., 2023b), along with UniEval and G-Eval+. For conciseness, we include the models supporting the evaluation of relevance, such as UniEval and G-Eval, and our G-Eval+ tailored for conciseness. We use GPT-4-turbo for all LLM-based models.

Evaluation Metric. Following prior works (Liu et al., 2023a; Fu et al., 2023), we compare the estimated scores with the ground-truth human scores in UNISUMEVAL. We report their Pearson correlation based on the *summary-level* percentage score of each evaluation dimension. See Appendix D.2 for the details of measurements and Appendix F.4 more results using system-level measurement.

6.1 Alignment with Human Score

Tables 6–8 report the agreement between automated evaluators and humans in faithfulness, completeness, and conciseness. In the booking domain, evaluators that require reference summaries do not report results since these summaries are unavailable. In general, the LLM-based evaluators, G-Eval+, show the highest agreement in all dimensions. The agreement with the human scores appears to vary across different data domains.

In the faithfulness evaluation (Table 6), **the increasing granularity over the three LLM-based methods do not guarantee improved performance.** Specifically, the atomic level evaluator, FactScore, does not perform well due to the difficulty in atomic fact generation – often producing numerous overlapping or factually incorrect atomic facts, which can cascade into further errors by LLMs. Moreover, contrary to prior findings that NLI-based evaluators perform well (Tang et al., 2024b), they show much lower agreement compared to G-Eval when evaluating faithfulness on our hallucinogenic texts. Hence, **non-LLM evaluators tend to perform poorly at verifying LLM-generated hallucinations.**

Additionally, **the agreement with human faithfulness evaluation is significantly affected by domain characteristics.** In particular, QA- and NLI-based methods, which require training on specific data, lack domain generalization in automatic evaluation. They exhibit negative correlations with human judgements on more than half of the do-

mains in faithfulness evaluation. In contrast, the LLM-based evaluators (G-Eval and G-Eval+) show a significantly higher positive correlation compared to the other evaluators.

In the completeness and conciseness evaluation (Tables 7–8), for the LLM-based methods, **employing prompts fine-tuned for target dimensions leads to higher agreements in general**, *i.e.*, G-Eval → G-Eval+. Lastly, we observe a considerable performance discrepancy of SOTA automated evaluators in the conciseness dimension compared to the faithfulness and completeness dimensions. While the LLM-based evaluators show a fairly high agreement of up to 0.68 in evaluating faithfulness and completeness, they exhibit considerably lower agreement of up to 0.51 for conciseness evaluation. This highlights that **evaluating conciseness could be harder than other dimensions** in the text summarization task.

In general, LLM-based evaluators (except for FactScore in faithfulness) achieve higher agreement with human judgments compared to QA- and NLI-based evaluators across all three dimensions. However, they still fall short in certain domains as automated evaluators. If a threshold of 0.50 is set for satisfactory correlation with human judgments, G-Eval+ fails to meet this standard in the Report and Sci-fi domains for faithfulness evaluation, the Booking domain for completeness evaluation, and most domains for conciseness evaluation.

7 Conclusion

We introduce UNISUMEVAL, a benchmark dataset featuring hallucinogenic texts from nine domains, spanning non-dialogue to dialogue and short to long texts, paired with their summaries generated by nine recent summarizers. Built using a data creation pipeline with AI assistance, UNISUMEVAL includes high-quality, fine-grained human annotations that enable in-depth studies on the multi-dimensional performance of summarizers. Additionally, based on our benchmark, we provide a thorough assessment of automated evaluators for text summarization, revealing weaknesses related to specific domains and evaluation dimensions.

Limitations

Our work has some limitations. First, although UNISUMEVAL covers three comprehensive evaluation dimensions for summarization quality, additional dimensions like harmfulness or bias could enhance the nuanced assessment of summaries. Second, the generation of key-facts could be refined by developing tailored strategies for different domains to better extract domain-specific key-facts. Third, since the importance of key-facts can vary, key-fact alignment could measure the completeness and conciseness of summaries more precisely by considering the relative importance of each key-fact. Finally, although we achieve high IAA for human annotations, the IAA for long texts remains lower than for short texts. Future research is required to refine annotation strategies for long texts. Despite these challenges, we hope our work will give valuable insights into the field of text summarization and foster the development of a more advanced automated evaluators.

Ethics Statement

We actively addressed annotators' queries during the annotation process, ensuring faithful communication. Annotators were compensated at a rate 50% above the average American minimum wage and received bonuses for consistent, high-quality work. Our dataset excludes any information that could potentially disclose the annotators' personal details.

Scientific Artifacts

We utilized nine language models to generate summaries on UNISUMEVAL. Apart from the paid APIs like OpenAI and AWS Bedrock, we used readily available checkpoints on Huggingface. All the details are summarized in Table 10.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea goverment (MSIT) (No. RS-2024-00445087, Enhancing AI Model Reliability Through Domain-Specific Automated Value Alignment Assessment). Additionally, this work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334343) and the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2022M3J6A1063021).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *EMNLP*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *ACL*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved qa-based factual consistency evaluation for summarization. In *NAACL*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao and Xiaojun Wan. 2022. Dialsummeval: Revisiting summarization evaluation for dialogues. In *NAACL*.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. In *ACL*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *ACL*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *EACL*.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring llm ability at factual reasoning through the lens of summarization. In *EMNLP*.

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.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen Mckeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

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 *EMNLP*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Towards interpretable and efficient automatic reference-based summarization evaluation. In *EMNLP*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*.

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 *EMNLP*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint 2401.06855*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *NAACL*.

Hwanjun Song, Igor Shalyminov, Hang Su, Singh Siffi, Kaisheng Yao, and Saab Mansour. 2023. Enhancing abstractiveness of summarization models through calibrated distillation. In *EMNLP*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using llms. In *ACL*.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *ACL*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024b. TofuEval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *NAACL*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *EMNLP*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:1–7.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *NeurIPS*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *NLP4ConvAI*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *ACL*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *EMNLP*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *ICLR*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In *NAACL*.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *ACL*.

Dataset	Type	Text Length	Domain	Text Word count (Min – Max)	Summary Word count (Min – Max)	Key-fact Count (Min – Max)
CNNDM	Non-Dialogue	Short	News	674 (227 – 1,231)	105.7 (19 – 420)	11.4 (5 – 17)
WikiHow			Lifestyle	65.6 (21 – 151)	43.6 (5 – 152)	4.2 (1 – 13)
GovReport		Long	Report	6,263.7 (2,429 – 10,462)	155.9 (4 – 764)	17.8 (11 – 20)
PubMed			Medical	3,220.4 (1,204 – 5,586)	165.1 (4 – 2,349)	17.7 (11 – 20)
SQuALITY			Sci-fiction	6,084.8 (4,782 – 6,720)	110.1 (2 – 312)	12.4 (7 – 18)
DialogSum	Dialogue	Short	Daily Life	154.2 (63 – 287)	43.4 (7 – 128)	6.8 (3 – 14)
MultiWOZ			Booking	252 (118 – 349)	64.1 (10 – 125)	8.2 (4 – 13)
MediaSum		Long	Interview	1,635.2 (631 – 2,978)	113.1 (17 – 572)	12.6 (6 – 19)
MeetingBank			Meeting	978.3 (293 – 3,389)	89.3 (7 – 713)	9.2 (5 – 13)
UniSummEval				2,092 (21-10,462)	133 (2 – 2,349)	11.1 (1-20)

Table 9: Summary of the nine datasets in UNISUMEVAL: the average word count of texts and summaries, and the number of key facts, with their respective minimum and maximum ranges in parentheses. UNISUMEVAL sampled 25 hallucinogenic texts from each dataset.

A Summary of the Source datasets

Table 9 provides detailed statistics on the nine source datasets and their generated summaries/key-facts, where datasets with an average word count of more than 900 are classified as long texts. The selection of the source datasets is to cover various domains, with a balanced distribution of text types (dialogue, non-dialogue) and lengths (long, short). Our benchmark contains a total of 225 source documents, with each domain equally containing 25 documents.

B Model Settings and Prompts

B.1 Summary Generation Details

Text: {input text}

Instruction: Summarize the Text.

Provide your answer in JSON format. The answer should be a dictionary with the key "summary" containing a generated summary as a string:
{"summary": "your summary"}

JSON Output:

Figure 5: The prompt to generate a summary.

We briefly describe the settings of the summarization models in our benchmark. For the two non-LLMs, BART-large and Flan-T5-large, we choose the pre-trained models in HuggingFace model hub according to whether the domain is for dialogue or non-dialogue. We use instruction-tuned model checkpoints for the open-source LLMs and the official APIs for the proprietary LLMs. The checkpoints used for each model can be found in Table 10. We set the temperature to 1 and use the prompt shown in Figure 5 for generating summaries across the summarizers.

Model Name	HuggingFace Checkpoints
$BART_{large}$	facebook/bart-large-cnn linydub/bart-large-samsum
$Flan\text{-}T5_{large}$	spacemanidol/flan-t5-large-cnndm oguuzhansahin/flan-t5-large-samsum
Phi-2	microsoft/phi-2
$Mistral_{7B\text{-}Inst}$	mistralai/Mistral-7B-Instruct-v0.2
$Llama2_{13B\text{-}chat}$	meta-llama/Llama-2-13b-chat-hf
$Mixtral_{8x7B\text{-}Inst}$	mistralai/Mixtral-8x7B-Instruct-v0.1
$GPT\text{-}3.5_{turbo}$	gpt-3.5-turbo-0125*
$GPT\text{-}4_{turbo}$	gpt-4-0125-preview*
Claude2.1	claude-2.1*

Table 10: The checkpoints of the summarization models. *Their official APIs are used.

B.2 AI-assistant Details

We conduct LLM-based summary faithfulness evaluation to (1) select hallucinogenic texts and (2) provide machine-based reasoning to aid in the human annotation of fact verification. We modify the factual error types originally used in the prompt of FineSurE (Song et al., 2024), reducing them from nine to five. This makes the annotation task more intuitive and feasible for human annotators. The prompt to generate AI faithfulness evaluations is provided in Figure 6.

Reliability of AI Evaluation We use Claude 2.1 to generate AI faithfulness evaluations. To ensure the reliability of these evaluations, we conduct an automated faithfulness evaluation using an additional SOTA LLM, Llama3-70B-Instruct, on the entire pool of hallucinogenic texts. As a result, we find that 98.4% (1,180 out of 1,199) of the texts are confirmed as hallucinogenic by both models, suggesting that the bias may be insignificant.

You will receive an article followed by a corresponding summary. Your task is to assess the factuality of each summary sentence across five categories:
* out-of-article error: this error occurs when a summary statement introduces facts, subjective opinions, or new information not found in or verifiable by the article.
* entity error: this error occurs when there is an incorrect reference to a key subject or object in a summary statement, such as using a wrong name, number, or pronoun.
* relation error: this error occurs when there is a mistake in semantic relationships within a summary statement, including but not limited to incorrect use of verbs, prepositions, and adjectives.
* sentence error: this error occurs when an entire summary statement contradicts the information provided in the article.
* no error: the summary statement aligns explicitly with the content of the article and is factually consistent with it.

Instruction:
First, compare each summary sentence with the article.
Second, provide a single sentence explaining which factuality error the sentence has.
Third, answer the classified error category for each sentence in the summary.

Please do not change the order of sentences in your answer.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence", "reason", and "category":
[{"sentence": "first sentence", "reason": "your reason", "category": "no error"}, {"sentence": "second sentence", "reason": "your reason", "category": "out-of-article error"}, {"sentence": "third sentence", "reason": "your reason", "category": "entity error"},]

Article:
{input text}

Summary with N sentences:
1. {summary sentence 1}
2. {summary sentence 2}
...
N. {summary sentence N}

Figure 6: The prompt to generate AI evaluations on faithfulness.

B.3 Key-fact Generation Details

We extract an initial set of key-facts from each source text using GPT-4-turbo with the prompt described in Figure 7. To ensure the quality of the initial key-facts, we cross-validate them using GPT-3.5-turbo, Claude 2.1, and Mixtral-8x7B with the prompt described in Figure 8.

Your task is to identify 'key facts' within the Text, which are essential pieces of information for a high-quality summary.
The following is a set of detailed instructions for identifying key facts.

Instruction:
1. Identify up to 20 key facts.
1. A key fact should be brief and clear.
3. A key fact should encompass at most 2-3 entities.
4. Each key fact should deliver distinctive information.

Here are 8 examples of key facts to illustrate the desired level of granularity.
* Bulgaria's Black Sea has resorts.
* Black Sea resorts are cheaper than hotspots in Italy.
* Black Sea resorts are cheaper than hotspots in Spain.
* Cheap prices in Bulgaria are driven by low exchange rates.
* Alexandra Harra has become an Instagram star.
* Alexandra Harra is a model.
* Alexandra Harra posts selfies on Instagram.
* Lindsay Sandiford was convicted for attempting to smuggle cocaine.

Provide your answer in JSON format. The answer should be a dictionary with the key 'key facts' containing the key facts as a list:
{'key_facts': ['first key fact', 'second key facts', 'third key facts']}

Text: {input text}

Figure 7: The prompt to generate key-facts.

B.4 Auto-Evaluators Details

For the non-LLM evaluators such as QA- and NLI-based ones, we choose their default models or load model checkpoints that demonstrated the best performance in the paper. All other settings, including hyperparameters and prompts, are kept as provided in the original papers. In the case of FactScore, reference retrieval from the knowledge base is unnecessary; the evaluator assesses the faithfulness of the summary based on the input text alone.

We also use customized G-Eval prompts, referred to as G-Eval+, tailored to our three evaluation dimensions: faithfulness, completeness, and conciseness. The original G-Eval prompt alone does not perfectly align with these dimensions or their granularities. Therefore, we develop specific prompts for each dimension. For faithfulness, we evaluate at the sentence level, while for completeness and conciseness, we use prompts aligned with the evaluator's criteria. The detailed prompts for faithfulness, completeness, and conciseness are shown in Figures 9 –11, respectively.

You will receive Reference Article and a set of Candidate Statement that contain some information from the Reference Article.
Your job is to identify if each Candidate Statement is useful for making a summary of the Reference Article.

Instruction:
1. Read a Reference Article and a set of Candidate Statement carefully.
2. If the Candidate Statement is useful for making a summary of the Reference Article, response "Yes", otherwise response "No"
3. Provide a single sentence explaining why the Candidate Statement is useful for making a summary.

Reference Article:
{input text}

N Candidate Statement:
1. {key-fact 1}
2. {key-fact 2}
...
N. {key-fact N}

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "statement_id", "statement", "response", "reason". you should provide a response and a reason for all Candidate Statements.

For example:
[{"statement_id": "1", "statement": "first statement", "response": "your response", "reason": "your reason"}, {"statement_id": "2", "statement": "second statement", "response": "your response", "reason": "your reason"}, ... , {"statement_id": "N", "statement" :"N-th statement", "response": "your response", "reason": "your reason"}]

Figure 8: The prompt to cross-validate key-facts.

Is the summary sentence supported by the document?
Response with "Yes" or "No" for each sentence in the summary.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence" and "response":
[{"sentence": "first sentence", "response": "yes or no"}, {"sentence": "second sentence", "response": "yes or no"},{"sentence": "third sentence", "response": "yes or no"},]

Document:
{input text}

Summary with N sentences:
1. {summary sentence 1}
2. {summary sentence 2}
...
N. {summary sentence N}

Figure 9: G-Eval+ prompt tailored for sentence level in faithfulness evaluation.

You will be given a article. You will then be given one summary written for this article. Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Completeness (1-5) - the degree to which the summary includes all key information present in the source document. A complete summary accurately captures the main points, ideas, and relevant details without omitting crucial elements.

Evaluation Steps:
1. Read the article carefully and identify the main points, key information, and relevant details.
2. Read the summary and compare it to the article. Check if the summary captures all essential facts, main ideas, and pertinent details presented in the original article.
3. Assign a score from 1 to 5 for completeness based on the Evaluation Criteria.

Source Text:
{input text}

Summary:
{summary}

Evaluation Form (scores ONLY):

- Completeness:

Figure 10: G-Eval+ prompt tailored for completeness.

You will be given a article. You will then be given one summary written for this article. Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Conciseness (1-5) - the extent to which the summary presents information succinctly and without unnecessary elaboration. A concise summary effectively conveys the essential content of the source document using clear and concise language, avoiding redundant or superfluous information.

Evaluation Steps:
1. Read the article carefully and identify the main points, key information, and relevant details.
2. Read the summary and compare it to the article. Check if the summary effectively conveys the essential content of the document in a concise manner, without unnecessary elaboration or redundancy.
3. Assign a score for conciseness based on the Evaluation Criteria.

Source Text:
{input text}

Summary:
{summary}

Evaluation Form (scores ONLY):

- Conciseness:

Figure 11: G-Eval+ prompt tailored for conciseness.

C Factual Error Types

Table 11 presents detailed descriptions and examples of the factual error types used for the fact verification annotation. The taxonomy is based on a modified version of the taxonomy suggested by Mishra et al. (2024).

D Detailed Calculation of the Benchmark Scores

D.1 Summarization Performance Calculation

After collecting annotations at a fine-grained level, the scores can be aggregated at a summary-level percentage score for all three dimensions, following the recent work by Song et al. (2024).

For a document D, let $S = \{s_1, ..., s_N\}$ be the list of generated summaries with N sentences. Based on the result of faithfulness annotation, we can identify $S_{fact} \subseteq S$, a subset of S that are annotated as having no factual error. Consequently, the percentage score of faithfulness of summary S is calculated by:

$$Faithfulness(D, S) = |S_{fact}|/|S|. \quad (1)$$

Let $K = \{k_1, \ldots, k_M\}$ be the set of key facts, where M is the total number of key facts. Based on the result of the key fact alignment, we can define a bipartite graph $M = (K, S, E)$, where E consists of edges $\{(k, s) : k \rightarrow s \mid k \in K \wedge s \in S\}$ with $k \rightarrow s$ signifying that key fact k is labelled as being present in summary sentence s. The completeness and conciseness score for summary S are then calculated as a percentage score by:

$$\begin{aligned} Completeness(K, S) &= \big|\{k|(k, s) \in E\}\big|/|K| \\ Conciseness(K, S) &= \big|\{s|(k, s) \in E\}\big|/|S|. \end{aligned} \quad (2)$$

Here, the operator $|\cdot|$ denotes the cardinality of a set. With these scores, we can quantify a summary-level score for completeness, which reflects the extent to which the key facts are incorporated into the summary. Additionally, the conciseness score measures the extent to which the summary incorporates the key facts.

Domain Stability Score. The domain stability score quantifies how consistent a model's performance is across the nine given domains. We first calculate the instability score by taking the difference between the maximum and minimum performance scores across these domains. The domain stability score is then determined by subtracting the instability score from a fixed upper bound of 100. The domain stability score can be calculated in terms of the four score types - faithfulness, completeness, conciseness, and the composite score.

Let S_i represent the score of the model in the i-th domain, where $i = 1, \ldots, 9$. The instability score $Instability$ is computed as:

$$Instability = \max_i S_i - \min_i S_i. \quad (3)$$

Then, the domain stability score DoS is given by:

$$DoS = 100 - Instability. \quad (4)$$

Abstractiveness Score. Abstractiveness is quantified by calculating the ratio of novel n-grams in the summary that do not appear in the input text (Liu and Lapata, 2019; Song et al., 2023). For a summary S, let $n\text{-gram}_{\text{copied}}$ be the set of

Example Source Text		
In the heart of the bustling city, nestled inside a park, stands the historic Jefferson Library. Built in 1910, this architectural marvel houses a vast collection of rare books, manuscripts, and artifacts, attracting scholars and history enthusiasts from around the world. Its grand facade and ornate interiors make it a beloved landmark, reflecting the city's rich cultural heritage and commitment to education.		
Error Type	Description	Example Summary Sentence
Out-of-Article Error	This error occurs when a summary sentence introduces facts, subjective opinions, or biases that cannot be verified or confirmed by the source text.	The Jefferson Library was **the first library to offer online book lending services.**
Entity Error	This error involves incorrect or misrepresented entities (such as names, numbers, or main subjects) within the summary sentence.	The Jefferson **School** houses a vast collection of rare books.
Relation Error	This error arises from incorrect semantic relationships within a summary sentence, such as wrong verbs, prepositions, or adjectives, which misrepresent the relationship between entities.	The Jefferson Library is located **beside** a park.
Sentence Error	This error occurs when a summary sentence entirely contradicts the information in the source text, requiring significant revision or removal.	The Jefferson Library is **a modern structure with minimalist architecture.**

Table 11: Descriptions and examples of factual error types. The parts of each summary sentence that are relevant to the specific error type are highlighted in bold.

n-grams that are copied from the document, and let $n\text{-gram}_{\text{total}}$ be the set of all n-grams in the summary. Then, the ratio of novel n-grams N_{n} can be defined as:

$$N_{\text{n}} = 1 - \frac{n\text{-gram}_{\text{copied}}}{n\text{-gram}_{\text{total}}}. \quad (5)$$

Following Song et al. (2023), the abstractiveness score for a summary S is calculated as the average of the novel 1/3/5-gram ratios:

$$Abstractiveness(D, S) = \frac{N_1 + N_3 + N_5}{3}. \quad (6)$$

D.2 Evaluator Performance Calculation

We calculate summary- and system-level correlation to verify the agreement between automated evaluation and human evaluation, following the recent work (Song et al., 2024; Liu et al., 2023a).

For summary-level evaluation, we analyze the correlations between the scores based on the human annotations and those calculated by automatic evaluators. Let x_i be the percentage score of a evaluation based on human annotation and y_i be the score generated by automated evaluators on the i-th data. Then, the summary-level correlation is calculated as follows:

$$\text{Corr}\left([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]\right). \quad (7)$$

where Corr is a function calculating a pearson correlation coefficient.

For system-level evaluation, we aggregate the percentage scores for each summarizer across all input texts, and then calculate the rank correlation between the ranks based on human annotation results. Let $X_{i,j}$ represent the score calculated by an automated evaluator, on an input text i for a summarizer j. The aggregated score $\overline{X}_j$ for a summarizer j is given by:

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}. \quad (8)$$

Similarly, let $Y_{i,j}$ represent the human score for a summarizer j for an input text i. The aggregated score $\overline{Y}_j$ for a summarizer j based on human evaluation is given by:

$$\overline{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{i,j}. \quad (9)$$

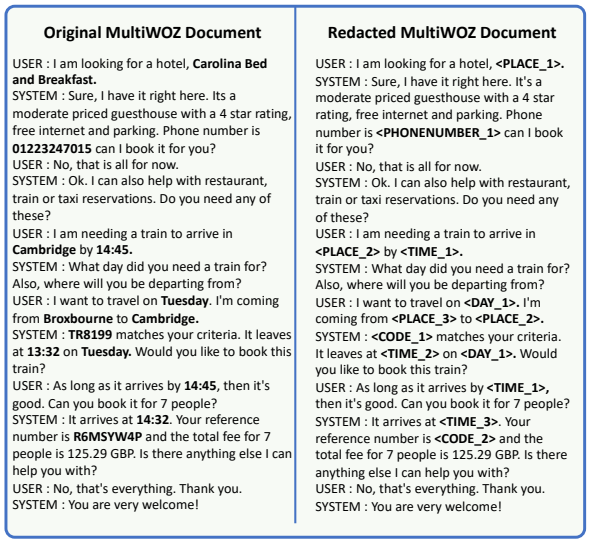
Then, we make a ranking of the average percentage scores for all summarization systems. Let $R(X_j)$ represent the rank of $\overline{X}_j$ and $R(Y_j)$ represent the rank of $\overline{Y}_j$. we compute the rank correlation by:

$$\text{Spearman}\Big([R(X_1), R(X_2), \ldots], [R(Y_1), R(Y_2), \ldots]\Big). \quad (10)$$

E Redaction Protocol

For redaction, we identify six specific categories requiring redaction: place, time, day, phone number, code, (such as reservation numbers and train numbers), and address (including postcodes). All pertinent entities within the source text are redacted using their category name and angle brackets "<>". Furthermore, each redacted entity in the same category is indexed numerically and recurring entities are consistently labeled with the same index number. An example of an source text and its redacted version is available in Figure 12.

This method of indexing is crucial for distinguishing between different entities within the same categories. Without such systematic indexing, the semantic integrity of the sentences can be severely damaged, resulting in a convoluted text that obscures the intended relationships inherent in the source text. This can make human annotation virtually impossible, as the relational context critical for understanding the dialogue can be lost.



Original MultiWOZ Document

USER : I am looking for a hotel, **Carolina Bed and Breakfast.**
SYSTEM : Sure, I have it right here. Its a moderate priced guesthouse with a 4 star rating, free internet and parking. Phone number is **01223247015** can I book it for you?
USER : No, that is all for now.
SYSTEM : Ok. I can also help with restaurant, train or taxi reservations. Do you need any of these?
USER : I am needing a train to arrive in **Cambridge** by **14:45.**
SYSTEM : What day did you need a train for? Also, where will you be departing from?
USER : I want to travel on **Tuesday**. I'm coming from **Broxbourne** to **Cambridge.**
SYSTEM : **TR8199** matches your criteria. It leaves at **13:32** on **Tuesday.** Would you like to book this train?
USER : As long as it arrives by **14:45**, then it's good. Can you book it for 7 people?
SYSTEM : It arrives at **14:32**. Your reference number is **R6MSYW4P** and the total fee for 7 people is 125.29 GBP. Is there anything else I can help you with?
USER : No, that's everything. Thank you.
SYSTEM : You are very welcome!

Redacted MultiWOZ Document

USER : I am looking for a hotel, **<PLACE_1>.**
SYSTEM : Sure, I have it right here. It's a moderate priced guesthouse with a 4 star rating, free internet and parking. Phone number is **<PHONENUMBER_1>** can I book it for you?
USER : No, that is all for now.
SYSTEM : Ok. I can also help with restaurant, train or taxi reservations. Do you need any of these?
USER : I am needing a train to arrive in **<PLACE_2>** by **<TIME_1>.**
SYSTEM : What day did you need a train for? Also, where will you be departing from?
USER : I want to travel on **<DAY_1>.** I'm coming from **<PLACE_3>** to **<PLACE_2>.**
SYSTEM : **<CODE_1>** matches your criteria. It leaves at **<TIME_2>** on **<DAY_1>.** Would you like to book this train?
USER : As long as it arrives by **<TIME_1>,** then it's good. Can you book it for 7 people?
SYSTEM : It arrives at **<TIME_3>**. Your reference number is **<CODE_2>** and the total fee for 7 people is 125.29 GBP. Is there anything else I can help you with?
USER : No, that's everything. Thank you.
SYSTEM : You are very welcome!

Figure 12: An example of a redacted MultiWOZ document.

F Additional Analysis

This section provides further analysis of the collected human annotations and summary evaluator performance.

F.1 Challenges in Automated Evaluation

Table 12 shows the ratios of human corrections to LLM's sentence-level binary labels in the fact verification task. It reveals that human annotators identify the highest frequency of errors for summaries generated by the proprietary LLMs, followed by those generated by the open-source LLMs, and the least for those generated by the non-LLMs. This trend highlights that automated faithfulness evaluation is more challenging for summaries generated by recent LLMs compared to those generated by non-LLMs. This can be attributed to the fact that factual errors in LLM-generated summaries are more complex and nuanced, often involving subtle misrepresentations that are harder to detect.

F.2 Detailed Human Annotation Result

Tables 14 –17 presents a comprehensive breakdown of the human annotation results for each domain and model, separately for faithfulness, completeness, conciseness, and the composite score. We present additional domain-level findings.

Faithfulness Score. Table 14 indicates that the faithfulness scores are fairly high across all domains except for non-LLMs. The input type (dialogue vs. non-dialogue) causes more significant variations in faithfulness scores than the domain itself. Specifically, faithfulness scores for dialogue generally range from 61.4% to 99.2%, which are lower compared to non-dialogue scores, which range from 74.7% to 100.0%.

Model type	Human Correction to AI-Claimed Error*	Human Correction to AI-Claimed Non-Error**
Non LLM	17.83% (168/942)	1.79% (62/3,459)
Open Source LLM	20.00% (174/870)	2.66% (287/10,803)
Prop. LLM	22.49% (56/249)	2.84% (229/8,076)
Total	19.31% (398/2,061)	2.59% (578/22,338)

Table 12: Human correction to AI factuality evaluation labels. *Cases where the AI inaccurately flags sentences as factually incorrect, and human annotators correct these errors. **Cases where the AI deems sentences error-free, but human annotators identify factual inaccuracies.

Model type	ρ
Non-LLM	-0.24*
Open-source LLM	-0.14*
Proprietary LLM	0.05

Table 13: Pearson correlation coefficients between human score in faithfulness evaluation and abstractiveness scores. *p-value < 0.05

Completeness Score. Table 15 reveals that, across all language model categories, completeness scores generally drop significantly in three domains: Report, Medical Literature, and Sci-fi. This suggests that recent summarizers struggle to identify key information in documents characterized by specialized terminologies, as in the Report and Medical Literature domains, or by intricate plots and unique vocabulary, as in the Sci-fi domain.

Conciseness Score. Table 16 demonstrates that the conciseness scores are generally high, with the exception of the Sci-fi domain. This finding suggests that verbose summaries generated by recent language models may be attributed to the imaginative content and unconventional plot structures typical of the Sci-fi domain.

F.3 The Relationship between Abstractiveness and Faithfulness

Table 13 shows the Pearson correlation coefficients between the human scores in faithfulness evaluation and abstractiveness scores across the three summarizer categories. The trade-off between abstractiveness and faithfulness exists only for the non-LLMs and open-source LLMs.

Model Type	Summ. Model	Non-Dialogue					Dialogue				Avg. Score	DoS
		News	Lifestyle	Report	Med lit	Sci-fi	Daliy life	Booking	Interview	Meeting		
Non LLM	$BART_{large}$	91.5	83.7	91.0	91.9	85.7	80.3	82.8	81.1	74.7	84.7	82.8
	$Flan\text{-}T5_{large}$	84.7	96.0	85.3	90.0	74.7	66.0	78.6	89.5	61.4	80.7	65.4
Open Source LLM	Phi-2	81.8	62.8	**100.0**	75.8	77.6	83.6	90.0	81.1	81.3	81.5	62.8
	$Mistral_{7B\text{-}Inst}$	94.7	97.2	97.4	98.8	94.7	91.4	97.9	92.8	92.9	95.3	92.7
	$Llama2_{13B\text{-}Chat}$	92.7	**100.0**	96.7	89.7	91.0	91.0	92.5	89.6	83.7	91.9	83.7
	$Mixtral_{8x7B\text{-}Inst}$	97.4	98.0	99.2	98.2	93.7	98.3	93.9	**98.9**	92.4	96.7	93.1
Prop. LLM	$GPT\text{-}3.5_{turbo}$	95.0	96.0	99.6	97.6	94.2	**99.0**	98.0	93.2	94.4	96.3	**93.6**
	$GPT\text{-}4_{turbo}$	97.3	97.2	98.8	96.3	**99.3**	98.0	100.0	93.2	92.1	96.9	92.1
	Claude2.1	**99.3**	93.3	**100.0**	**100.0**	97.9	97.6	**99.2**	92.3	**98.6**	**97.6**	92.3
Avg. Score in Domain		92.7	91.6	96.5	93.1	89.9	89.5	92.5	90.6	90.2		

Table 14: Faithfulness scores of each summarizer across the nine domains. "DoS" indicates domain stability scores for faithfulness.

Model Type	Summ. Model	Non-Dialogue					Dialogue				Avg. Score	DoS
		News	Lifestyle	Report	Med lit	Sci-fi	Daliy life	Booking	Interview	Meeting		
Non LLM	$BART_{large}$	49.8	86.0	17.0	23.8	8.0	40.1	55.4	30.8	31.0	38.0	22.0
	$Flan\text{-}T5_{large}$	43.2	43.7	21.2	19.2	8.4	46.5	42.5	31.1	32.1	38.1	**61.9**
Open Source LLM	Phi-2	63.9	49.2	19.5	18.2	13.0	46.4	27.4	27.8	23.4	32.1	49.1
	$Mistral_{7B\text{-}Inst}$	**71.9**	83.5	46.8	44.9	**37.6**	63.3	69.1	55.0	63.8	59.5	54.1
	$Llama2_{13B\text{-}chat}$	50.7	72.6	27.9	48.8	10.8	37.6	43.0	37.7	33.9	40.3	38.2
	$Mixtral_{8x7B\text{-}Inst}$	69.4	**92.4**	**47.8**	50.7	34.2	64.1	72.7	58.3	62.7	61.4	41.8
Prop. LLM	$GPT\text{-}3.5_{turbo}$	53.1	90.4	45.1	**53.0**	29.0	66.0	59.6	47.3	51.3	55.0	38.7
	$GPT\text{-}4_{turbo}$	66.2	**92.4**	47.6	50.9	36.6	**76.4**	**76.2**	**61.5**	**67.4**	**63.9**	44.2
	Claude2.1	53.4	75.0	26.4	36.1	33.4	67.5	53.4	45.0	44.7	48.3	51.4
Avg. Score in Domain		57.9	76.1	33.3	38.4	23.4	56.4	55.5	43.8	45.7		

Table 15: Completeness scores of each summarizer across the nine domains. "DoS" indicates domain stability scores for completeness.

Model Type	Summ. Model	Non-Dialogue					Dialogue				Avg. Score	DoS
		News	Lifestyle	Report	Med lit	Sci-fi	Daliy life	Booking	Interview	Meeting		
Non LLM	$BART_{large}$	**90.9**	82.3	73.3	81.3	25.8	75.7	83.4	68.5	75.9	73.0	34.9
	$Flan\text{-}T5_{large}$	89.7	78.0	**88.3**	83.3	35.3	80.3	72.3	74.2	60.6	73.6	46.7
Open Source LLM	Phi-2	73.8	68.3	40.1	63.4	27.3	75.4	46.3	46.0	53.0	54.8	51.9
	$Mistral_{7B\text{-}Inst}$	83.6	95.0	78.7	72.8	59.8	78.5	80.7	75.1	75.0	77.7	64.8
	$Llama2_{13B\text{-}chat}$	90.5	94.7	65.7	85.6	39.4	88.2	71.6	76.8	69.4	75.8	44.7
	$Mixtral_{8x7B\text{-}Inst}$	82.8	92.5	70.7	70.2	49.8	84.2	70.0	79.5	68.4	74.2	57.3
Prop. LLM	$GPT\text{-}3.5_{turbo}$	84.9	98.0	87.8	**91.4**	65.1	**94.1**	72.4	79.7	78.7	83.6	67.1
	$GPT\text{-}4_{turbo}$	85.8	**95.2**	87.8	90.0	**70.1**	87.7	**86.1**	84.5	79.3	**85.2**	74.9
	Claude2.1	85.2	84.0	76.9	85.4	61.7	83.1	68.9	**85.2**	**86.5**	79.7	**75.3**
Avg. Score in Domain		85.3	87.5	74.4	80.4	48.3	83.0	72.4	74.4	71.9		

Table 16: Conciseness scores of each summarizer across the nine domains. "DoS" indicates domain stability scores for conciseness.

Model Type	Summ. Model	Non-Dialogue					Dialogue				Avg. Score	DoS
		News	Lifestyle	Report	Med lit	Sci-fi	Daliy life	Booking	Interview	Meeting		
Non LLM	$BART_{large}$	77.4	84.0	60.5	65.6	39.8	65.4	73.9	60.1	60.5	65.2	55.8
	$Flan\text{-}T5_{large}$	72.5	72.6	65.0	64.2	39.5	64.3	64.5	64.9	51.8	62.1	66.9
Open Source LLM	Phi-2	73.1	60.1	53.2	52.5	39.3	68.5	54.6	51.6	52.6	56.2	66.1
	$Mistral_{7B\text{-}Inst}$	**83.4**	91.9	74.3	72.2	64.0	77.8	82.6	74.3	77.3	77.5	72.1
	$Llama2_{13B\text{-}chat}$	78.0	89.1	63.4	74.7	47.0	72.3	69.0	68.1	62.4	69.3	58.0
	$Mixtral_{8x7B\text{-}Inst}$	83.2	94.3	72.6	73.0	59.2	82.2	78.9	78.9	74.5	77.4	64.9
Prop. LLM	$GPT\text{-}3.5_{turbo}$	77.7	94.8	77.5	**80.7**	53.1	64.4	56.8	57.7	57.7	68.9	58.3
	$GPT\text{-}4_{turbo}$	83.1	**94.9**	**78.1**	79.0	**68.7**	**87.4**	**87.5**	**79.7**	**79.6**	**82.0**	73.8
	Claude2.1	79.3	84.1	67.8	73.8	64.3	82.7	73.8	74.2	76.6	75.2	**80.2**
Avg. Score in Domain		78.6	85.1	68.0	70.6	53.6	75.9	73.3	69.3	67.5		

Table 17: Composite scores of each summarizer across the nine domains. "DoS" indicates domain stability scores for the composite scores.

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-Based	$UniEval_{faith}$	-0.52	0.74*	-0.60	-0.83*	-0.68*	-0.07	-	-0.22	-0.33
	QAFactEval	-0.73*	0.07	-0.79*	0.03	-0.72*	0.08	-0.10	-0.28	-0.50
NLI-Based	$SummaC_{Conv}$	-0.73*	0.25	-0.44	-0.43	-0.67*	-0.38	-0.72*	-0.72*	-0.63
	AlignScore	-0.40	0.24	-0.28	-0.10	0.45	0.80*	0.43	-0.35	0.65*
	MiniCheck	-0.37	0.71*	-0.31	-0.47	-0.90*	-0.10	0.27	-0.22	-0.05
LLM-Based	$G\text{-}Eval_{faith}$	**0.93***	**0.78***	**0.55**	0.32	**0.93***	0.83*	0.92*	**0.68***	**0.80***
	$G\text{-}Eval+_{faith}$	0.78*	0.75*	0.37	**0.68***	0.90*	**0.97***	**0.93***	0.55	0.72*
	FactScore	0.13	-0.01	-0.08	-0.10	-0.43	0.40	0.12	0.23	0.38

Table 18: System level rank correlation with human scores in faithfulness evaluation across nine domains (*: p-value < 0.05). For LLM-based methods, G-Eval, G-Eval+, and FactScore are the summary, sentence, and atomic level evaluators.

F.4 System-level Evaluator Benchmark Result

We report system-level results (See Appendix D.2 for the calculation details) of evaluator performance on our benchmark. Table 18–20 present the correlations between the scores predicted by the automated evaluators and the human scores at the system-level across the three dimensions.

F.5 Comparison with Similarity-based Metric

Table 21 shows the summary-level agreement with human scores for conventional similarity-based metrics (i.e., ROUGE-1/2/L (Lin, 2004) and BERTScore (Zhang et al., 2019)) across three dimensions (faithfulness, conciseness, and completeness) and composite score (the average of the three dimensions). In all dimensions, similarity-based evaluators show performance comparable to G-Eval+ in a few domains, such as Report, Medical Literature and Sci-fi. However, in general, they exhibit significantly weaker agreement with human scores in all dimensions compared to the LLM-based evaluator.

G Details of Manual Annotation

G.1 Annotator Qualification Requirements

For qualification requirements of annotators on MTurk, we select only those with an approval rating above 95% and at least 1,000 accepted HITs. Also, we administer a qualification test comprising 10 English comprehension questions that simulate the actual annotation tasks. We limit our pool of crowd-sourced workers to those who score 100 on this test and are based in AU, CA, NZ, GB, or US.

G.2 Annotator Compensation

Annotators are paid 50% above the average American minimum wage. We provided a $25 bonus to annotators who deliver 500 consecutive high-quality annotations.The total cost of obtaining fine-grained human annotations for the three evaluation dimensions exceeded $30K for 2,025 summaries.

G.3 Attention Check

Our human annotation process involves stringent protocols and detailed strategies that filter out

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-based	$\text{UniEval}_{\text{coh}}$	0.55	0.68*	-0.08	0.1	0.7*	0.23	-	0.5	0.70*
NLI-based	$\text{Lite}^3\text{Pyramid}$	0.8*	-0.6	0.53	**0.77***	0.92*	0.8*	-	0.47	**0.87***
	A^3CU	0.8*	0.32	0.65	0.57	**0.92***	0.57	-	0.67*	0.53
LLM-based	$\text{G-Eval}_{\text{coh}}$	0.58	0.83*	**0.83***	0.73*	0.72*	0.8*	0.72*	0.80*	0.82*
	$\text{G-Eval+}_{\text{com}}$	**0.92***	**0.88***	0.57	0.68*	0.75*	**0.88***	**0.75***	**0.92***	0.82*

Table 19: System level rank correlation with human scores in completeness evaluation across nine domains (*: p-value < 0.05). For UniEval and G-Eval, we use their coherence scores since the completeness dimension is not directly supported.

Model Type	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
QA-based	$\text{UniEval}_{\text{rel}}$	-0.27	-0.03	-0.08	**0.53**	0.77*	0.13	-	0.78*	0.47
NLI-based	$\text{Lite}^3\text{Pyramid}$	-0.77*	-0.02	0.03	0.43	0.65	0.35	-	0.75*	0.03
	A^3CU	-0.5	0.00	-0.45	0.00	0.52	0.57	-	0.58	-0.18
LLM-based	$\text{G-Eval}_{\text{rel}}$	-0.43	0.60	0.05	**0.53**	**0.88***	**0.58**	0.20	0.58	0.67*
	$\text{G-Eval+}_{\text{conc}}$	-0.1	**0.85***	0.05	**0.53**	**0.88***	**0.58**	0.02	0.73*	**0.7***

Table 20: System level rank correlation with human scores in conciseness evaluation across nine domains (*: p-value < 0.05). For UniEval and G-Eval, we use their relevance scores since the conciseness dimension is not directly supported.

Dimension	Evaluator	Non-Dialogue					Dialogue			
		News	Lifestyle	Report	Med lit	Sci-fi	Daily Life	Booking	Interview	Meeting
Faithfulness	ROUGE-1	0.12	0.08	0.27*	0.3*	0.18*	0.04	-	0.07	0.14*
	ROUGE-2	0.06	0.04	0.20	0.20	0.20	-0.03	-	0.09	0.04
	ROUGE-L	0.05	0.08	0.22*	0.26*	0.16*	0.03	-	0.08	0.05
	BERTScore	0.10	0.10	0.29*	0.25*	0.03	0.05	-	0.16*	0.17*
	$\text{G-Eval+}_{\text{faith}}$ †	**0.63***	**0.57***	**0.46***	**0.55***	**0.38***	**0.46***	**0.59***	**0.52***	**0.53***
Completeness	ROUGE-1	0.08	-0.09	**0.61***	0.47*	0.60*	-0.02	-	-0.03	0.17*
	ROUGE-2	0.10	0.01	0.46*	0.39*	0.54*	0.01	-	0.01	0.02
	ROUGE-L	-0.01	-0.09	0.46*	0.38*	0.57*	-0.02	-	-0.04	0.02
	BERTScore	0.18*	-0.10	0.49*	0.26*	0.48*	-0.07	-	0.07	0.11
	$\text{G-Eval+}_{\text{com}}$ †	**0.57***	**0.61***	0.59*	**0.65***	**0.68***	**0.56***	**0.32***	**0.63***	**0.66***
Conciseness	ROUGE-1	0.13	0.04	-0.01	0.09	0.24	0.27	-	0.15	0.19
	ROUGE-2	0.14	0.04	0.00	0.08	0.22	0.21	-	0.10	0.09
	ROUGE-L	**0.18**	0.05	0.03	0.15	0.23	0.28	-	0.15	0.13
	BERTScore	0.11	0.05	0.16*	0.14	0.21*	0.26*	-	0.24*	0.17*
	$\text{G-Eval+}_{\text{con}}$ †	0.11	**0.39***	**0.17***	**0.24***	**0.44***	**0.36***	0.02	**0.49***	**0.45***
Composite	ROUGE-1	0.16*	0.00	0.39*	**0.38***	0.42*	0.13	-	0.09	0.25**
	ROUGE-2	0.16*	0.03	0.30*	0.30*	0.40*	0.09	-	0.09	0.06
	ROUGE-L	0.11	0.01	0.32*	0.35*	0.40*	0.14*	-	0.09	0.10
	BERTScore	0.20*	0.01	0.43*	0.28*	0.30*	0.11	-	0.21*	0.20*
	G-Eval+ †	**0.47***	**0.68***	**0.50***	-0.04	**0.68***	**0.63***	**0.36***	**0.66***	**0.73***

Table 21: Agreement with human scores for similarity-based evaluators in evaluations of three dimensions and a composite score (Pearson correlation on summary-level percentage scores). † : For comparative purposes, we include the results for G-Eval+, which is identified as the best-performing LLM-based evaluator in our main analysis.

low-quality responses. We extensively incorporates novel attention check methods across the three annotation stages – fact verification, key-fact validation, and key fact alignment – to eliminate low-quality and insincere submissions. For any submissions where annotators fail the attention checks, we reject their submissions. We ban repeated offenders from participating in future tasks.

These methods enable us to selectively collect high-quality annotations while leveraging the cost-effectiveness and time efficiency of crowd-sourcing platforms, which are crucial for scaling our annotation protocol to larger datasets.

Fact Verification Annotation. For each source text-summary sentences pair, we include two types of attention checks using a random summary sentence that should always be labeled as factually incorrect. We manually assign a machine label to this sentence as either "factually correct" or "fac-

tually incorrect." Annotators should always "disagree" with the "factually correct" machine label and "agree" with the "factually incorrect" machine label.

Key-Fact Validation Annotation. For each key-fact validation annotation task, we include questions with a random, irrelevant key fact that should always be answered as an invalid key fact.

Key-Fact Alignment Annotation. We introduce two types of attention checks. One type presents a random sentence as if it were a key fact, ensuring that the alignment with all summary sentences is "not aligned." The other type involves inserting fake summary sentences: one is a slight paraphrase of a key fact, which should always be answered as "aligned," and another is a random sentence consistently appearing across all assessments, which should always be classified as "not aligned".