Towards Multi-dimensional Evaluation of LLM Summarization across Domains and Languages

Hyangsuk Min[1,*], **Yuho Lee**[1,*], **Minjeong Ban**[1], **Jiaqi Deng**[1],
Nicole Hee-Yeon Kim[1], **Taewon Yun**[1], **Hang Su**[2,†], **Jason Cai**[2,†], **Hwanjun Song**[1‡]
[1]Korea Advanced Institute of Science and Technology
[2]AWS AI Labs
{hyangsuk.min, yuholee, songhwanjun}@kaist.ac.kr

Abstract

Evaluation frameworks for text summarization have evolved in terms of both domain coverage and metrics. However, existing benchmarks still lack domain-specific assessment criteria, remain predominantly English-centric, and face challenges with human annotation due to the complexity of reasoning. To address these, we introduce MSumBench, which provides a multi-dimensional, multi-domain evaluation of summarization in English and Chinese. It also incorporates specialized assessment criteria for each domain and leverages a multi-agent debate system to enhance annotation quality. By evaluating eight modern summarization models, we discover distinct performance patterns across domains and languages. We further examine large language models as summary evaluators, analyzing the correlation between their evaluation and summarization capabilities, and uncovering systematic bias in their assessment of self-generated summaries. Our benchmark dataset is publicly available at https://github.com/DISL-Lab/MSumBench.

1 Introduction

Recent advancements in large language models (LLMs) have enhanced text summarization performance. However, LLM-generated summaries still face challenges, including hallucinations, omission of critical information, and redundancy (Lee et al., 2024). These limitations highlight the continued need for advanced automated evaluation methods that can assess summarization quality more efficiently and cost-effectively than human annotation.

While automated evaluations have made progress, there remains a significant gap between automatic evaluations and human judgments, reinforcing the need for benchmarks with high-quality human annotations to provide a more reliable assessment of summarization capabilities.

However, existing benchmarks face three key challenges. First, most benchmarks employ uniform criteria for assessing summary quality, failing to account for domain-specific differences in what constitutes important information (Laban et al., 2023; Lee et al., 2024). Second, they remain largely monolingual—primarily focused on English—limiting their ability to provide robust evaluations across languages (Bhandari et al., 2020; Fabbri et al., 2021; Pagnoni et al., 2021; Laban et al., 2022, 2023; Tang et al., 2023; Lee et al., 2024; Tang et al., 2024b). Third, collecting high-quality human annotations remains a significant challenge. Assessing summarization quality requires complex reasoning, making the annotation resource-intensive and inconsistent (Krishna et al., 2023; Lee et al., 2024). These challenges hinder the timely development of reliable benchmarks that can keep pace with the rapid advancements in LLM capabilities.

To address these, we create MSumBench, **M**ulti-aspect **Sum**marization **Bench**mark (Figure 1), a summarization benchmark for English and Chinese with domain-specific evaluation criteria (see Table 2). Building on existing multi-dimensional and fine-grained evaluation frameworks (Song et al., 2024), human annotations for summarization quality are collected at both the sentence and key-fact[1] levels, focusing on faithfulness, completeness, and conciseness. Furthermore, as depicted in Figure 2, we propose a multi-agent debate system that facilitates effective AI-human collaboration in handling the complex task of assessing summary faithfulness. Inspired by prior findings that demonstrate the effectiveness of LLM-based debate (Chan et al., 2024; Du et al., 2024; Khan et al., 2024; Koupaee et al., 2025), we extend this approach to the sum-

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

[1] A key-fact is a succinct statement capturing an individual essential information unit, containing a maximum of two to three entities (Bhandari et al., 2020; Song et al., 2024).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14417–14450
July 27 - August 1, 2025

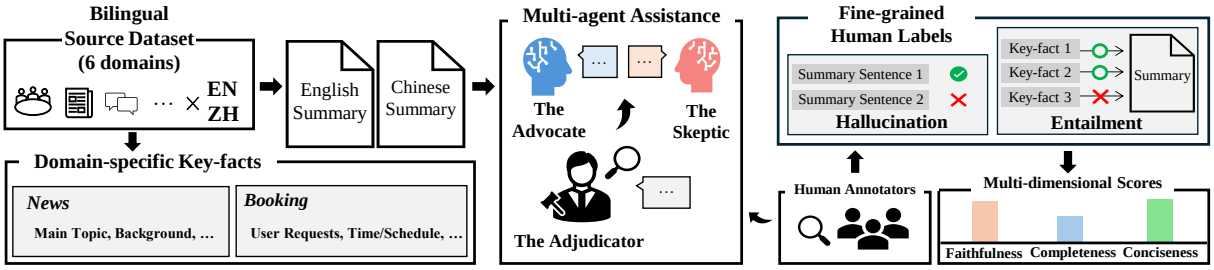

Figure 1: Overview of MSumBench, featuring multi-domain documents in both English and Chinese, with domain-specific key-facts. Model summaries are evaluated via a multi-agent debate framework, aiding annotators' assessments. Each summary then receives percentage scores for faithfulness, completeness, and conciseness.

marization annotation task. Our debate-based system provides annotators with structured arguments from LLM agents–called the Advocate and the Skeptic–with contrasting viewpoints, thereby reducing dependence on any single viewpoint. We further include the Adjudicator's review to finalize the debate and ensure that all arguments remain fact-based and consistent with the source content. To ease the cognitive load of annotators, we guide them to focus on the most relevant portions of the source text—key entities, relationships, and critical details—rather than requiring them to consider the entire document from scratch. This comprehensive yet focused process not only promotes more accurate annotations but also helps minimize biases from unbalanced or incomplete information.

Our main contributions are: (1) We propose an evaluation strategy that identifies different types of critical information specific to each domain. (2) We develop a multi-agent, debate-based annotation framework that generates structured arguments with contrasting perspectives, enabling human annotators to focus on key aspects of the task. (3) We conduct a comprehensive evaluation of state-of-the-art LLMs as summarizers for English and Chinese, using our multi-dimensional annotations. (4) Using the collected annotations as ground truth, we thoroughly examine the effectiveness of LLMs as automated evaluators in both languages. (5) We release the MSumBench benchmark to facilitate advancements in summarization evaluation.

2 Related Works

Evaluation Benchmarks Conventional summarization benchmarks mainly focus on the news domain (Pagnoni et al., 2021; Tang et al., 2023), which has led to the development of benchmarks incorporating dialogues (Gao and Wan, 2022; Krishna et al., 2023; Tang et al., 2024b), and multi-domains such as SummEdits (Laban et al., 2023) and UniSumEval (Lee et al., 2024). However, the monolingual focus on English remains a persistent challenge. While some benchmarks like mFACE (Aharoni et al., 2023) and MFHHD (Shen et al., 2024) provide multilingual evaluations, their focus is limited to the news domain, calling for a single benchmark that considers multiple languages and multiple domains for comprehensive evaluations. Furthermore, while criteria for evaluating summary quality differ across domains, existing multi-domain benchmarks generally employ uniform criteria across different domains (Laban et al., 2023; Lee et al., 2024).

Evaluation Metric Many existing benchmarks focus solely on faithfulness (i.e., factual consistency) as an evaluation dimension of summarization quality (Bhandari et al., 2020; Pagnoni et al., 2021; Laban et al., 2022; Tang et al., 2023; Krishna et al., 2023; Laban et al., 2023). Some have expanded to multiple dimensions like coherence and relevance (Fabbri et al., 2021; Gao and Wan, 2022; Tang et al., 2024b), while Lee et al. (2024) uses completeness and conciseness for greater evaluation consistency. Evaluation measurement has likewise shifted from coarse, summary-level binary or scale-based ratings (Fabbri et al., 2021; Laban et al., 2022; Gao and Wan, 2022; Tang et al., 2023; Laban et al., 2023; Tang et al., 2024b) to fine-grained, sentence-level (or lower) assessments represented as percentage scores (Bhandari et al., 2020; Pagnoni et al., 2021; Krishna et al., 2023; Lee et al., 2024).

Paralleling these advances, automated metrics have also evolved: from traditional similarity-based measures like ROUGE (Lin, 2004), BERTScore (Yuan et al., 2021), and BARTScore (Zhang et al., 2019); to natural language inference (NLI) and question answering (QA)-based methods (Fabbri et al., 2022; Tang et al., 2024a); and finally to LLM-based approaches (Liu et al., 2023; Song et al.,

	Domain Coverage	Evaluation Dimensions	Domain-specific Evaluation	Annotation Unit	Measurement	Language
mFACE	✗ Single	✓ **3**	✗ No	✗ Summary	✗ Likert	✓ **Multi**
SummEdits	✓ **Multi**	✗ 1	✗ No	✗ Summary	✗ Ternary	✗ Single
MFHHD	✗ Single	✗ 1	✓ **Yes**	▲ Sentence*	✗ Ternary	✓ **Bi**
UniSumEval	✓ **Multi**	✓ **3**	✗ No	✓ **Sentence & Key-fact**	✓ **Percentage**	✗ Single
Ours	✓ **Multi**	✓ **3**	✓ **Yes**	✓ **Sentence & Key-fact**	✓ **Percentage**	✓ **Bi**

Table 1: Benchmark comparison. *The annotated summaries are constrained to a single sentence format.

News	Medical Literature	Report	Booking	Meeting	Interview
Main topic			User requests		
Background	Research finding	Governance	System suggestions	Opinions	Background
Immediate impact	Disease descriptions	Evaluations	Location/route	Reports	Main arguments
Public statements	Medical experiments	Recommendations	General information	Decisions	Supporting examples
Official statements	Medical treatment	Regulation/policy	Booking confirmation	Proposals	Counter arguments
Counter arguments	Medical prevention	Financial info	Price/payment	Factual info	Conclusions
Future implications			Time/schedule		

Table 2: Key-fact categories tailored to each domain, with the first row representing domains and the second row listing domain-specific key-facts. See Appendix B.2 for a detailed description.

2024; Wan et al., 2024). However, multilingual coverage remains challenging as noted in Forde et al. (2024). Table 1 highlights how MSumBench addresses these limitations by comparing it against some of the existing benchmarks.

3 MSumBench Construction Procedure

We construct MSumBench following the systematic pipeline of four components: dataset collection, domain-specific key-fact generation, summary generation, and summary evaluation. Detailed statistics of MSumBench are provided in Appendix A.

3.1 Dataset Collection

Source Dataset Evaluating summarization models on a single domain provides limited insight into the robustness of their performance. Accordingly, MSumBench is constructed based on the datasets from six domains with distinct characteristics: CNN/DM (news) (Nallapati et al., 2016), GovReport (report) (Huang et al., 2021), PubMed (medical literature) (Cohan et al., 2018), MultiWOZ (booking conversation) (Zang et al., 2020), MediaSum (interview) (Zhu et al., 2021), and MeetingBank (meeting) (Hu et al., 2023). From each domain, we sample 25 documents, yielding a total of 150 source documents to generate summaries.

Documents Translation Since MSumBench aims to evaluate both English and Chinese summaries, identical source texts are needed in both languages to ensure contextual consistency and fair comparison. Therefore, we opt to translate the original English source documents into Chinese. The translation follows the three steps. First, we use GPT-4o for initial sentence-level translation with domain-specific prompts (see Appendix B.1) to maintain contextual coherence. Second, we use Qwen-2.5-72B to screen the resulting translations to flag unnatural or inaccurate sentences and to alleviate any bias potentially introduced by using GPT-4o as a single translator. Finally, any flagged sentences are reviewed by bilingual native Chinese examiners for refinement. This multi-step validation ensures the accuracy, naturalness, and contextual integrity of the translations.

3.2 Domain-Specific Key-Facts Extraction

Domain-specific criteria are critical for summary evaluation, as each domain emphasizes distinct content. Thus, we define tailored key-fact categories that serve as templates for ideal summary contents in each domain. This approach ensures summary evaluations reflect the most important aspects of original documents while adhering to the established norms and practices of each domain.

Key-Fact Extraction Procedure The domain-specific key-fact extraction procedure follows three steps. First, *Category Identification*: To determine which information categories matter most in each domain, we analyze human-written reference summaries[2] from existing summarization datasets. Then, we derive approximately 5–7 recurring categories of essential information (see Table 2). This

[2]Since MultiWOZ lacks reference summaries, we analyze frequently occurring entity categories labeled by humans.

ensures that each category captures frequently emphasized details within its respective domain.

Second, *Key-fact Classification*: We use GPT-4o to generate candidate key-facts from source documents and classify them using domain-specific categories, filtering out non-aligned key-facts. This minimizes risk of missing any potential key-facts while discarding information that does not meet domain-specific priorities.

Third, *Key-fact Validation*: To enhance reliability, each filtered key-fact is verified by three LLMs—GPT-4o, Claude-3.5-sonnet, and Llama-3.1-70B. Each model checks whether the key-fact: (1) is useful for summarizing; (2) is consistent with the information in the source text; and (3) falls under one of the predefined domain-specific categories. Using a majority vote, any key-fact failing these checks is discarded. This cross-verification mitigates potential biases or oversights that could arise from relying on a single model. The resulting consensus-based key-facts constitute a robust domain-specific reference set, forming the backbone of our subsequent evaluation of summary quality. Appendix B.2 provides the prompts for extracting and validating the key-facts.

3.3 Summary Generation

To evaluate how summarization capabilities vary across different model scales and architectures–from traditional fine-tuned models to the latest LLMs, we select eight models for benchmarking, categorized into three groups: *non-LLMs*, including fine-tuned BART-Large (Lewis et al., 2020) and mT5 (Xue et al., 2021), *open-source LLMs*, including Llama-3.1-70B (Grattafiori et al., 2024), Gemma-2-27b (Team et al., 2024), and Qwen-2.5-72B (Yang et al., 2024), and *proprietary LLMs*, including GPT-4o (OpenAI, 2024), Claude-3.5-Sonnet, and Gemini-1.5-pro (Gemini Team, 2024). For English documents, we generate 1,200 summaries (25 source documents $\times$ 6 domains $\times$ 8 summarizers). For Chinese summaries, we exclude BART due to its lack of multilingual support, resulting in 1,050 generated summaries (25 source documents $\times$ 6 domains $\times$ 7 summarizers). See Appendix B.3 for model configuration details.

3.4 Summary Evaluation

Human Annotation Tasks Traditional summary evaluation dimensions (e.g., coherence and relevance) are insufficient for fine-grained evaluation because they lack clear, measurable criteria. Instead, we assess the generated summaries based on three dimensions–faithfulness, completeness, and conciseness–following Song et al. (2024). Our annotation process consists of two main tasks: fact verification and key-fact alignment.

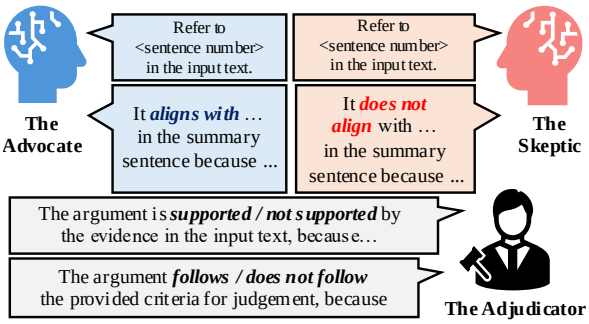

Figure 2: Annotation assistance via multi-agent debate.

In the fact verification task, annotators evaluate faithfulness at the sentence-level by identifying factual errors based on the existing error taxonomy (Lee et al., 2024) (see Appendix C). In the key-fact alignment task, annotators verify whether each key-fact can be inferred from the summary sentences, similar to an NLI task. These tasks yield 9,951 and 188,800 annotations from each respective task.

We collect our human annotations via Amazon Mechanical Turk (MTurk), with each annotation unit assigned to three independent annotators to ensure reliability. Further details on annotator recruitment and procedure are provided in Appendix H.

Fact Verification with Multi-Agent Assistance Although human judgments remain essential for accurate evaluation, the annotation process is both costly and challenging, as summarizers evolve and error types become more nuanced. A straightforward solution might be to have humans inspect labels generated by an LLM-based evaluator, similar to the work of Lee et al. (2024). However, this can lead to a bias where annotators blindly endorse LLMs' decisions. Thus, we introduce a more systematic approach that fosters effective AI-human collaboration while mitigating bias: a multi-agent debate-assisted annotation framework (Figure 2).

Our framework employs three LLM agents for fact verification: the *Advocate*, the *Skeptic*, and the *Adjudicator*. The Advocate and the Skeptic engage in the core debate, while the Adjudicator investigates their arguments. The agents debate the faithfulness of each summary sentence against the full input document (with input sentences numbered).

Specifically, the Advocate presents evidence supporting the faithfulness of a given summary sentence, and the Skeptic presents evidence suggesting

Annotation Task	Measuring Dimensions	News	Report	Medical Lit.	Booking	Meeting	Interview	Avg (EN)	Avg (ZH)	Avg (All)
Fact Verification	Faithfulness	0.67	0.64	0.46	0.56	0.50	0.64	0.55	0.61	0.58
Key-fact Alignment	Completeness & Conciseness	0.70	0.84	0.77	0.83	0.80	0.82	0.77	0.82	0.79

Table 3: Consistency of three human annotators across 6 domains for 2 languages, where the first six columns are the IAA scores for each domains, while the last three are the average IAA for two languages and the overall one.

it may be unfaithful. Both agents first highlight relevant sentence numbers in the input text so annotators can quickly locate the information. Next, they explain which parts of the summary sentence align (or fail to align) with the referenced input sentences (see Table 24 and 25). The Skeptic also specifies the error type. The Adjudicator then reviews both arguments, the input context, and the summary sentence, focusing on two key aspects: (1) whether the provided evidence is consistent with the source text, and (2) whether the arguments follow the faithfulness criteria (see Table 26). This step prevents superficial objections (e.g., wrongly labeling correct paraphrases as errors). Finally, the Adjudicator produces an investigation report and a tentative label, which are shown to the annotator along with the reference text and both sides' arguments. We use Llama-3.1-70B for all three agents.

By providing balanced evidence from both sides and clearly indicating how each argument relates to specific source sentences, this framework minimizes bias and offers a practical aid to annotators. While multi-round debates (Ray, 2023; Du et al., 2024) may yield richer evaluations, they add cognitive load for annotators, undermining practical benefits as assistance. Conversely, our single-round setup balances thoroughness and usability, ensuring efficient and accurate annotations.

Key-Fact Alignment with NLI Assistance While the multi-agent debate aids in detecting nuanced faithfulness errors, key-fact alignment is a straightforward comparison between two concise sentences, requiring no complex reasoning mechanisms. Therefore, we present annotators with an NLI result generated by Llama-3.1-70B, indicating whether the key-fact is entailed by the summary sentence, along with a brief rationale. This approach streamlines the annotation process by guiding annotators to make quick, confident judgments, thereby reducing their overall cognitive load.

4 Quality Assessment

A high-quality benchmark ensures that model evaluations capture genuine performance differences by minimizing annotation noise (Lee et al., 2024; Tang et al., 2024b). Thus, we evaluate the reliability and accuracy of collected annotations to ensure the integrity of `MSumBench`.

Task	Measure	`UniSumEval`	`MSumBench`
Key-fact List	Human Preference	25.00%	75.00%
Fact Labels	Balanced Accuracy	80.07%	92.83%

Table 4: Comparison of `MSumBench` over `UniSumEval`. Human Preference is the A/B test preference ratio, while Balance Accuracy is measured using the two expert examiners' labels as reference.

4.1 Annotation Consistency

Inter-Annotator Agreement (IAA) is a measurement to assess the reliability of human annotations by quantifying the consistency between different annotators. Table 3 reports the IAA scores using Krippendorff's α (Krippendorff, 2011) across domains for our two annotation tasks, namely fact verification and key-fact alignment. We achieve very high average IAA scores, Avg (All), of 0.58 and 0.79 for the two tasks, respectively. This indicates that **MSumBench stands out as the only comprehensive benchmark with domain-specific evaluations and bilingual coverage**, ensuring robust and diverse summarization assessment.

4.2 Comparison with `UniSumEval`

While high IAA scores indicate annotation consistency, they do not guarantee dataset quality due to potential biases and accuracy issues (Munappy et al., 2022; Braylan et al., 2022). Therefore, we conduct an additional quality check by recruiting two postgraduate NLP specialists as expert examiners, both proficient in English and Chinese (see Appendix H for details on recruitment).

We perform quality checks on two critical components of `MSumBench`: (1) *Domain-Specific Key-Fact Extraction* to ensure that extracted key-facts align with domain characteristics, and (2) *Multi-Agent Debate* to confirm the accuracy of human annotations assisted by our multi-agent system. To demonstrate the effectiveness of these quality improvements, we compare our dataset with the latest `UniSumEval` dataset (Lee et al., 2024).

Model Type	Summarizer	English				Chinese				Language Stability
		Faithfulness	Completeness	Conciseness	Domain*	Faithfulness	Completeness	Conciseness	Domain*	
Proprietary LLMs	GPT-4o	86.10 (5)	50.02 (4)	77.98 (6)	89.28 (1)	78.42 (5)	41.19 (3)	74.98 (5)	86.96 (4)	93.02 (4)
	Claude 3.5 $_{\text{Sonnet}}$	89.42 (1)	52.67 (2)	80.09 (4)	88.87 (2)	82.68 (1)	49.6 (2)	78.07 (3)	87.53 (3)	96.3 (2)
	Gemini 1.5 $_{\text{Pro}}$	86.18 (4)	54.81 (1)	81.88 (2)	87.79 (4)	80.26 (3)	53.55 (1)	79.43 (2)	87.55 (2)	97.16 (1)
	Average	87.23 (3.33)	52.5 (2.33)	79.98 (4)	88.65 (2.33)	80.45 (3)	48.11 (2)	77.49 (3.33)	87.35 (3)	95.5 (2.33)
Open-source LLMs	Gemma 2 $_{\text{27B}}$	87.02 (3)	37.06 (6)	73.67 (7)	84.55 (6)	78.53 (4)	32.18 (5)	69.69 (6)	82.95 (6)	93.47 (3)
	Llama 3.1 $_{\text{70B}}$	83.50 (6)	39.19 (5)	81.31 (3)	87.08 (5)	77.10 (6)	24.78 (6)	75.13 (4)	88.06 (1)	88.41 (6)
	Qwen 2.5 $_{\text{72B}}$	87.95 (2)	50.87 (3)	84.87 (1)	88.57 (3)	80.29 (2)	40.94 (4)	79.51 (1)	85.81 (5)	92.09 (5)
	Average	86.16 (3.67)	42.37 (4.67)	79.95 (3.67)	86.73 (4.67)	78.64 (4)	32.63 (5)	74.77 (3.67)	85.61 (4)	91.32 (4.67)
Non-LLMs	mT5	12.33 (8)	3.79 (8)	28.67 (8)	55.85 (8)	25.33 (7)	2.86 (7)	27.67 (7)	48.76 (7)	82.75 (7)
	BART	77.57 (7)	25.01 (7)	78.48 (5)	79.2 (7)	-	-	-	-	-
	Average	44.95 (7.5)	14.4 (7.5)	53.58 (6.5)	67.53 (7.5)	-	-	-	-	-

Table 5: Summarization performance of eight summarizers for two languages. Rankings are shown in parentheses, with cell color intensity increasing within each column to indicate higher ranks. Domain*: domain stability.

4.2.1 Domain-specific Key-Fact Extraction Quality Check

Unlike UniSumEval's domain-agnostic key-fact extraction, MSumBench employs domain-specific categories (Table 2) to better reflect each domain's unique characteristics. For comparative evaluation, the two expert examiners perform A/B comparisons between key-facts extracted from 150 English source documents using both approaches.

Table 4 shows that domain-specific key-facts are highly preferred by the examiners. MSumBench's approach yields a 50%p higher preference rate than UniSumEval (Cohen's $\kappa = 0.48$[3]). It suggests that **domain-specific key-facts better capture what humans consider essential in each domain**, leading to more targeted summary evaluation. Appendix D presents detailed A/B test guidelines and domain-wise results.

4.2.2 Multi-Agent Debate Quality Check

Unlike UniSumEval, which offers a single viewpoint, our debating system mitigates bias by presenting contrasting perspectives. To verify this, we ask the two expert examiners to re-annotate representative samples of summary sentences[4] from both datasets for sentence-level fact verification. The two examiners engage in repeated discussions to reach a consensus on their decisions, ensuring reliable ground-truth labels.

Table 4 shows that our debating system produces more accurate human labels, achieving a 12.76%p higher balanced accuracy compared to UniSumEval. Among incorrect annotations in UniSumEval, 95.31% align with the single-view assistance label, indicating that the majority of such errors stem from annotators blindly endorsing LLM decisions. This reveals that **a single-view approach inflates IAA scores but harms accuracy, while a multi-view approach can mitigate this by promoting accurate annotation.**

5 Benchmarking Summarizers

Dimension and Metric We evaluate summarization performance across five key dimensions: *faithfulness*, *conciseness*, *completeness*, *domain stability*, and *language stability*. The first three are based on a well-established work (Lee et al., 2024), and are evaluated as *percentage* scores as suggested by the original paper (Song et al., 2024).

Domain and language stability assess the consistency of summarization performance across domains and languages. To assess domain stability, we first compute the coefficient of variation (CV) across domains for each of the first three dimensions-faithfulness, completeness, and conciseness. These three CV values are then averaged to obtain a composite domain stability score. Language stability is assessed similarly, with CVs calculated across languages rather than domains. Detailed calculations are in Appendix E.

5.1 Overview

Table 5 compares the performance of eight summarizers across English and Chinese. Proprietary LLMs demonstrate superior performance compared to open-source and non-LLMs across languages, particularly in completeness. They maintain the consistent essential information coverage across languages (52.5 to 48.11), while open-

[3]Cohen's kappa (κ) is a statistical measure for inter-rater reliability for two annotators (McHugh, 2012).

[4]We sample 624 sentences each from UniSumEval (8,133 sentences) and MSumBench (9,951 sentences), achieving 99% confidence with ±5% margin of error.

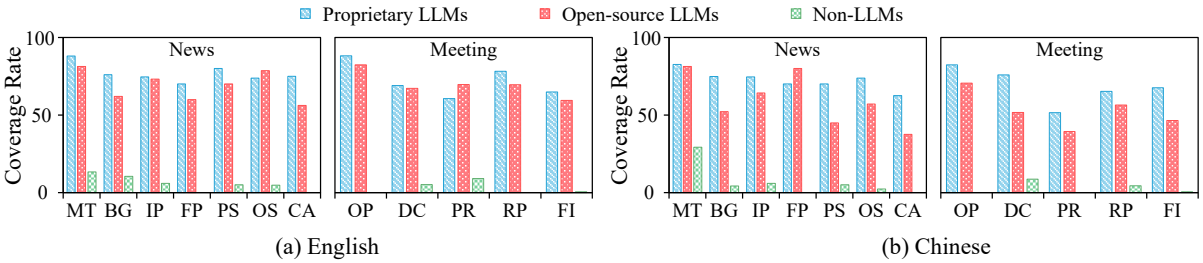

Figure 3: Key-fact category coverage ratio in two domains for English and Chinese. Abbreviations on the x-axis refers: MT(Main topic), BG(Background), IP(Immediate impact), FP(Future implications), PS(Public statements), OS(Official statements), CA(Counterarguments), OP(Opinions), DC(Decisions), PR(Reports), FI(Factual Info).

source LLMs experience substantial degradation (42.37 to 32.63). This reveals that **proprietary LLMs capture domain characteristics more effectively in both languages, whereas open-source and non-LLMs do not.** Nevertheless, performance gaps persist across languages regardless of the summarizer type, calling for further efforts to enhance multilingual performance consistency.

5.2 Detailed Analysis

We conduct a detailed analysis of the generated summaries, examining completeness and conciseness through key-fact category coverage, and faithfulness through factuality error types.

5.2.1 Comparison on Key-Fact Coverage

Figure 3 shows the ratio of key-fact categories covered in the generated summaries. For a detailed analysis, we examine two domains (News and Meeting) in English and Chinese, highlighting coverage differences across three model types.

Generally, in English (Figure 3(a)), proprietary and open-source LLMs exhibit relatively consistent key-fact coverage across categories in each domain. However, in Chinese (Figure 3(b)), both show greater variability, such as a large gap between MT and CA in News and OP and PR in Meeting. This highlights that **there is significant category-wise imbalance in key-fact retention in Chinese**, which suggests that similar issues may arise in other non-major languages as well.

From a domain perspective, in general, News exhibits higher coverage rates than Meeting in both languages, though the dominance of key-fact categories in the same domain vary across languages. This reveals that **all models, regardless of type, still lack satisfactory consistency in capturing key-facts across languages and domains**, which directly impacts the completeness and conciseness of the generated summary.

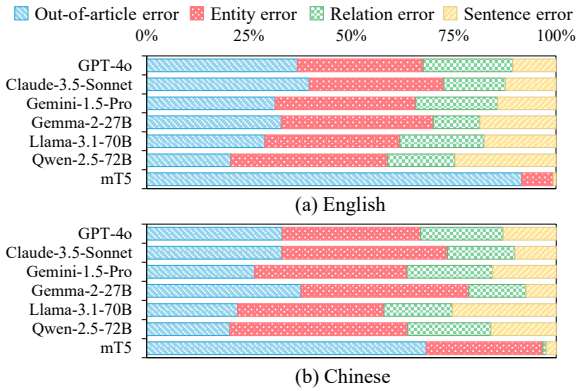

Figure 4: Distribution of error types across summarizers and languages. A detailed description of each error type is provided in Appendix C.

5.2.2 Analysis on Factuality Error

We analyze the distribution of factuality errors appearing in the summaries. Figure 4 shows the distribution across different summarizers for two languages. There is no significant difference in error distribution between proprietary and open-source LLMs, nor across languages. The consistency across languages indicates that **faithfulness in summarization is an inherent property of LLMs rather than a language-dependent factor.** Thus, enhancing summarization fidelity is likely to generalize across both models and languages.

6 Benchmark: LLMs as Evaluators

Benchmarking LLMs as summarization evaluators is crucial, as it enables scalable and consistent automated assessment, reducing reliance on costly, time-consuming human evaluation. Their ability to provide objective and reproducible evaluations makes them valuable for benchmarking summarization models across different domains and languages. To assess their effectiveness, we analyze three key aspects:

• Accuracy of LLMs as evaluators based on their agreement with human judgments.

Model Type	Model Name	English				Chinese				Language Stability
		Faithfulness	Completeness	Conciseness	Domain*	Faithfulness	Completeness	Conciseness	Domain*	
Proprietary LLMs	GPT-4o	0.61	**0.74**	0.60	**88.31**	0.52	0.71	**0.65**	**85.76**	**94.91**
	Claude 3.5 $_{\text{Sonnet}}$	0.60	0.72	**0.62**	87.38	0.54	0.65	0.53	80.27	90.09
Open-source LLMs	Llama 3.1 $_{\text{70B}}$	0.55	0.70	0.57	85.64	0.41	0.65	0.46	75.44	93.39
	Qwen 2.5 $_{\text{72B}}$	**0.67**	0.65	0.55	86.96	**0.59**	**0.76**	0.55	81.11	88.40

Table 6: Benchmarking LLMs as evaluators in English and Chinese. The higher score, the better accuracy in summary evaluation. Domain*: Domain Stability. Stability scores are computed as in the summarization evaluation.

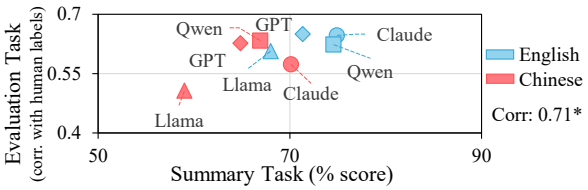

Figure 5: Analysis on cross-task correlation between summarization and evaluation. *: p-value < 0.05.

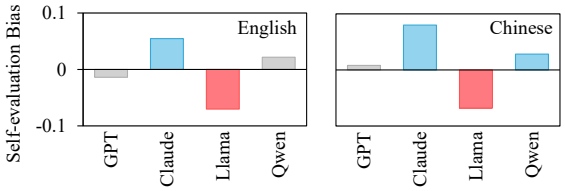

Figure 6: Self-evaluation bias rates across models and languages. Colored bars denote statistically significant differences determined by t-test ($p < 0.05$).

• Correlation between LLMs' summarization performance and evaluation performance.

• Self-evaluation bias, which occurs when an LLM assess summaries it has generated.

Selected Models We compare two proprietary LLMs (Claude-3.5-Sonnet, GPT-4o) and two open-source LLMs (Qwen-2.5-72B and Llama-3.1-70B), which exhibits considerably different summarization performance within their respective groups.

Evaluation Metric We compare the score obtained by LLM-based evaluation with human annotated labels in `MSumBench`. We report the Pearson correlation for the summary-level percentage scores across evaluation dimensions. See Appendix E for the detail of measurements.

6.1 Correlation with Human Judgments

Table 6 shows the correlation between LLM and human evaluation across faithfulness, completeness, and conciseness, along with domain and language stability, reflecting correlation consistency across six domains and two languages.

GPT-4o demonstrates the highest domain and language stability. However, no single model consistently achieves the best performance across all dimensions. Notably, Qwen-2.5-72B stands out for its strong performance in faithfulness evaluation. Therefore, **there is a significant variability in LLM evaluation accuracy across evaluation dimensions, domains, and languages**, indicating that relying on a single model for automated evaluation lacks reliability.

6.2 Cross-Task Correlation

Figure 5 shows the correlation between summarization and evaluation performance[5]. The x-axis denotes summarization performance, while the y-axis represents its performance as an evaluator.

We observe **a strong correlation ρ = 0.71 between LLM performance in summarization and evaluation tasks**, showing that strong summarizers generally excel at evaluation. This suggests that improving an LLM's summarization can also enhance its evaluation, and vice versa. However, as noted in Section 6.1, even the best summarizers may still show limitations when evaluating certain dimensions or working with different languages.

6.3 Self-Evaluation Bias

In summary evaluation, LLMs are expected to favor their own summaries (Wataoka et al., 2024), but this bias has not been fully analyzed. Thus, we assess to what extent this bias exists. We first define the self-evaluation bias rate as the difference between an LLM's evaluation score for its own summaries and the average score assigned by other evaluators (see Appendix F for detailed calculation). Thus, the higher (or lower) the rate, the more favorably (or unfavorably) it evaluates its own summaries. We use a t-test to check if the difference is statistically significant.

Figure 6 shows the self-evaluation bias rate across LLMs and languages. It reveals that **self-evaluation bias does not always manifest as self-preference, as it can manifest as either over-**

[5]Performance here refers to the composite score defined by the average of the three dimensions-faithfulness, completeness, and conciseness.

rating or under-rating their own work. Specifically, Llama-3.1-70B rates its own summaries less favorably than others, while Claude-3.5-Sonnet favors its own. This indicates variability in self-assessment across models.

7 Conclusion

We create MSumBench, a multi-aspect benchmark for summary evaluation across two languages and six distinct domains. MSumBench extracts key-facts to precisely measure summary-context alignment across six specialized domains. To ensure accurate annotations, we introduce a multi-agent assistant that minimizes human errors and delivers high-quality labels. We conduct an in-depth analysis, providing insights into LLMs' behavior as both summarizers and automated evaluators. The open dataset supports evaluation, enhances summarization, and aids preference optimization.

Limitations

Our work has several limitations as follows:

First, while we incorporate domain-specific key-fact categories, they are derived primarily through statistical patterns. The framework could benefit from direct input by domain experts to capture complex, specialized details more accurately.

Second, while our evaluation framework accounts for differences in importance across domains, it does not address variations in summary purpose within a single domain. For instance, in medical literature, the focus of a summary could shift from clinical treatments to epidemiological data, depending on summarization perspectives. Future work could explore purpose-specific key-fact categories, aligning summaries more closely with varied summarization objectives.

Third, our bilingual focus (English and Chinese) is an improvement over monolingual benchmarks. However, it still excludes many underrepresented languages. Evaluating how effectively LLMs handle other lower-resource languages remains an open question. Lastly, while our multi-agent debate framework provides balanced arguments to assist annotators, exploring other forms of collaboration, such as enhancing debates with dynamic rebuttals, could further improve the effectiveness of AI-human collaboration for summary quality annotation. Despite the challenges, we believe that our work serves as a meaningful foundation for future research, particularly in developing more robust automated evaluators and improving the factual consistency of domain-specific summarization systems across different languages.

Ethics Statement

Throughout our study, we prioritized comprehensive communication with all participating annotators, including both crowd-sourced and expert annotators. Crowd-sourced annotators received compensation above the U.S. federal minimum wage rate, while expert examiners were compensated at rates exceeding $30 per hour, with performance-based incentives. To ensure privacy compliance, all annotator personal information is anonymized in our dataset.

Scientific Artifacts

Our proposed benchmark combines publicly available datasets. For summary generation, we used Huggingface checkpoints and commercial APIs such as OpenAI and AWS Bedrock. Summary model details are in Table 10.

Acknowledgments

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).

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *ACL*.

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

Alexander Braylan, Omar Alonso, and Matthew Lease. 2022. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *WWW*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu.

2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *ICLR*.

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*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *ICML*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *TACL*, 9.

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

Jessica Zosa Forde, Ruochen Zhang, Lintang Sutawika, Alham Fikri Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *EMNLP*.

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

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

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

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*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *ICML*.

Mahnaz Koupaee, Jake W Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, et al. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. In *NAACL*.

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. *TACL*, 10.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *EMNLP*.

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. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Aiswarya Raj Munappy, Jan Bosch, Helena Holmström Olsson, Anders Arpteg, and Björn Brinne. 2022. Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software*, page 111359.

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

OpenAI. 2024. Gpt-4o system card.

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

Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Jiaming Shen, Tianqi Liu, Jialu Liu, Zhen Qin, Jay Pavagadhi, Simon Baumgartner, and Michael Bendersky. 2024. Multilingual fine-grained news headline hallucination detection. In *EMNLP*.

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*.

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

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *ACL*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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*.

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

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

Dataset	Domain	English		Chinse		
		Text Word count (Min - Max)	Summary Word count (Min - Max)	Text Word count (Min - Max)	Summary Word count (Min - Max)	Key-fact Count (Min - Max)
CNN/DM	News	503.6 (234 – 962)	86.01 (10 – 205)	885.4 (447 – 1,657)	146.52 (13 – 285)	15.72 (6 – 26)
PubMed	Medical Literature	2360 (856 – 4496)	60.07 (14 – 113)	3,797.52 (1,272 – 7,862)	221.90 (12 – 591)	23.84 (10 – 53)
GovReport	Report	3401 (1345 – 6837)	126.66 (7 – 347)	6,223.12 (2,466 – 10,359)	261.71 (10 – 1,704)	25.08 (16 – 35)
MultiWOZ	Booking	243.08 (138 – 382)	107.56 (8 – 260)	465.56 (258 – 788)	120.19 (12 – 250)	11.52 (6 – 19)
MeetingBank	Meeting	547.44 (97 – 1276)	71.26 (9 – 143)	925.6 (351 – 1,991)	135.49 (11 – 412)	12.64 (5 – 30)
MediaSum	Interview	1108.64 (186 – 4082)	87.15 (9 – 196)	1,951.2 (396 – 7,895)	162.47 (15 – 585)	16.72 (6 – 27)
MSumBench		1,360 (138 – 6,837)	2374.73 (258-10,359)	89.89 (7 - 347)	174.71 (10 – 1,704)	17.59 (5 – 53)

Table 7: Overview of six datasets utilized in MSumBench: includes statistics on source documents and their summaries, showing mean word counts and key-fact quantities, along with their respective ranges (min to max values).

A Summary of the Source datasets

Table 7 presents a comprehensive analysis of six diverse datasets encompassing various domains, including news, medical literature, reports, booking, meetings, and interviews in both English and Chinese languages. These datasets are strategically selected to evaluate the model's capability in handling domain-specific contextual understanding. The evaluation benchmark consists of 150 source documents, evenly distributed with 25 documents per domain in each language.

B Dataset Construction Pipeline Detail

Domain	Retained Key-Facts	Newly Added Key-Facts	Removed Key-Facts
News	63.6 %	36.4%	14.0 %
Medical Lit.	58.1 %	41.9 %	24.9 %
Report	60.8 %	39.2 %	23.1 %
Booking	79.2 %	20.8 %	4.8 %
Meeting	72.2 %	27.8 %	17.4 %
Interview	70.6 %	29.4 %	16.7 %
Overall	67.4 %	32.6 %	16.8 %

Table 8: Key-fact comparison between domain-specific method (A) and generic methods (B): Retained Key-facts represent the percentage of key-facts that remain consistent between A and B. Newly Added Key-facts indicate the percentage of key-facts introduced in A, but absent in B. Removed Key-facts show the percentage of key-facts present in B but missing in A.

B.1 Translation Detail

We present implementation details for the translation of source documents. For English-Chinese translation of source documents and key-facts, we incorporate domain-specific writing styles into our prompts, as shown in Tables 12 - 15. The key-fact translation prompt used for the key-fact alignment task with auto-evaluators is presented in Table 16. To ensure translation quality, we developed a screening prompt detailed in Table 17, which evaluates translations sentence-by-sentence across multiple dimensions: accuracy, consistency, fluency, comprehensibility, cultural adaptation, formality, and adequacy.

Domain	UniSumEval	MSumBench	Cohen's kappa
News	32.0%	68.0 %	0.46
Medical Lit.	10.0 %	90.0 %	0.34
Report	14.0 %	86.0 %	0.50
Booking	22.0 %	78.0 %	0.43
Meeting	38.0 %	62.0 %	0.43
Interview	26.0 %	74.0 %	0.48
Overall	21.7 %	78.3 %	0.47

Table 9: The percentage scores of A/B testing for domain-specific key-facts human preferences between UniSumEval and MSumBench.

B.2 Key-Fact Generation

Key-Fact Extraction and Validation Table 20 presents the domain-specific key-facts generation prompt, including well-defined categories and detailed domain-specific descriptions, along with representative examples illustrating granularity. The categories for each domain are detailed in Table 18, with corresponding examples in Table 19. To ensure the quality of extracted key-facts, we perform key-fact validation using the prompt in Table 21.

Key-Fact Extractiveness We provide additional analysis on the extractiveness of key-facts, in relation to their source documents. Since key-facts are designed to function as atomic information units, analyzing their extractiveness helps characterize their structural properties. We measure extractiveness using three metrics. First, direct copy measures the percentage of key-facts that exactly match a sentence in the source document. Second, near-exact match captures the proportion of key-facts

Model Name	Hugging Face Checkpoints & Official API Version	Hardware & Precision
GPT-4o	gpt-4o-2024-08-06	API (OpenAI, default settings)
Claude-3.5-Sonnet	anthropic.claude-3-5-sonnet-20241022-v2:0	API (AWS Bedrock, default settings)
Gemini-1.5-Pro	gemini-1.5-pro-002	API (Google API, default settings)
Gemma-2-27B	google/gemma-2-27b-it	NVIDIA L40S 48GB (×2) & BF16
Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct	NVIDIA L40S 48GB (×4) & BF16
Qwen2.5-72B	Qwen/Qwen2.5-72B-Instruct	NVIDIA L40S 48GB (×4) & BF16
mT5	sebuetnlp/mT5_multilingual_XLSum	NVIDIA L40S 48GB (×1) & Full precision
BART	facebook/bart-large-cnn linydub/bart-large-samsum	NVIDIA L40S 48GB (×1) & Full precision

Table 10: The checkpoint of the model used to generate summaries.

Domain	Direct Copy(%)	Near-Exact Match(%)	Extractiveness Score(%)
News	4.68%	16.31%	0.63
Medical Lit.	0.19%	18.10%	0.64
Report	5.68%	23.52%	0.68
Booking	4.96%	3.34%	0.52
Meeting	4.67%	13.88%	0.58
Interview	3.08%	9.33%	0.53
Overall	3.88%	14.08%	0.60

Table 11: The measurements of extractiveness of key-fact across domains in three ways: direct copy, near-exact match and extractiveness score.

with over 90% similarity to the source, as measured by the Edit Distance algorithm. Such cases typically involve surface-level edits, including whitespace changes or removal of redundant adjectives. Third, we compute the extractiveness score using the metric proposed by Song et al. (2024), defined as the average n-gram overlap ($n = 1/3/5$) between each key-fact and the source document.

Table 11 shows the extractiveness patterns across domains. Direct verbatim copying is relatively rare, with an average rate of 3.88%. Near-exact matches average 14.08%, indicating that most key-facts undergo some degree of reformulation from the original text. The extractiveness score averages 0.60, showing moderate lexical overlap with source documents, consistent with the interpretation by Song et al. (2023). These results indicate that key-facts are typically reformulated to some extent, rather than directly copied, reflecting their role as processed, atomic information units suitable for systematic evaluation.

B.3 Summary Generation Detail

Table 10 details the model versions employed for summary generation. For non-LLMs, we utilized pre-trained models from Hugging Face, while for open-source LLMs, we implemented instruction-tuned checkpoints. Access to proprietary LLMs was facilitated through their respective official APIs. The prompts for both English and Chinese summary generation are detailed in Table 22, with the temperature parameter set to 1.0 to promote diversity in generated summaries.

B.4 Key-Fact Alignment Detail

Table 23 presents our streamlined NLI prompt for key-fact alignment assessment. The prompt instructs the model to perform a direct entailment check between summary sentences and key-facts, providing brief, focused explanations. This design enables quick and confident assessment of information alignment while minimizing unnecessary complexity.

B.5 Multi-Agent Debate System Detail

The prompts used for our multi-agent fact verification system are presented in Tables 24 - 26. Each agent's prompt is tailored to their specific role: the Advocate focuses on finding supporting evidence, the Skeptic emphasizes identifying potential contradictions or flaws, and the Adjudicator concentrates on analyzing both arguments against the source document for final verification.

C Faithfulness Error Types

We use error types from Lee et al. (2024):

• **Out-of-article Error**: Introduces unverifiable facts, subjective opinions, or biases not supported by the source document.

• **Entity Error**: Misrepresents or includes incorrect entities, such as numbers or main subjects.

• **Relation Error**: Distorts the intended semantic relationships through incorrect use of verbs, prepositions, or adjectives.

• **Sentence Error**: Contradicts the source document, requiring major revisions or removal for factual alignment.

D Key-Fact Quality Assessment

Table 8 illustrates a quantitative comparison between our domain-specific extraction methodology with the domain-agnostic approach from Lee et al. (2024). The results indicate that our domain-specific extraction methodology encompasses a 32.61% more in domain-relevant key-fact identification relative to the baseline. Notably, the domain-specific strategy systematically excludes 16.83% of key-facts identified by the domain-agnostic method, as these elements lie beyond the scope of domain-specific criteria. This selective filtering highlights our methodology's precision in capturing domain-relevant information.

Additionally, we qualitatively assess the effectiveness of our domain-specific methodology compared to the domain-agnostic baseline via A/B preference testing with two expert annotators. We provide them with the English source document and two sets of key-facts: one extracted by the domain-agnostic baseline and the other identified by our domain-specific methodology. To minimize presentation bias, we present the two sets in randomized order, and the annotators are blind to the source of each set. They independently judge which set is more useful for summarization and better captures the essential content of the source document. To assess the consistency of these subjective judgments, we compute IAA using Cohen's kappa. Table 9 presents the comprehensive results of the A/B preference test and Cohen's kappa. Across all domains, `MSumBench` is consistently preferred by annotators, with preference rates ranging from 62% to 90.0%. Notably, Medical Literature shows a higher preference ratio (90.0%) compared to Meeting (62.0%), demonstrating the necessity of domain-specialized extraction in fields where domain expertise significantly influences content understanding.

E Benchmark Evaluation Formula

E.1 Evaluation Metrics for Summarization

We evaluate performance using five key metrics: faithfulness, completeness, conciseness, domain stability, and language stability. The first three metrics follow the methodology established by Song et al. (2024).

Faithfulness We consider a source document D, and its generated summary S containing N sentences $\{s_1, s_2, ..., s_N\}$. Based on human annotation, we identify $S_{Fact} \subseteq S$, which is the subset of S containing only factually accurate sentences. The faithfulness score is calculated as the ratio of factually correct sentences to total sentences:

$$\text{Faithfulness}(D, S) = \frac{|S_{Fact}|}{|S|}$$

Completeness and Conciseness Let K be the set of key-facts $\{k_1, k_2, ..., k_M\}$ from the source document. Based on the results of key-fact alignment, we can define bipartite graph $G = (K, S, E)$, where E represents edges between key-facts K and summary sentences S, $\{(k, s) : k \to s | k \in K \wedge s \in S\}$ with $k \to s$ signifying that key-fact k is entailed in s.

The completeness score measures how many key facts are captured in the summary:

$$\text{Completeness}(K, S) = \frac{|\{k|(k, s) \in E\}|}{|K|}$$

The conciseness score measures how efficiently the summary sentences convey key facts:

$$\text{Conciseness}(K, S) = \frac{|\{s|(k, s) \in E\}|}{|S|}$$

Domain Stability We introduce an improved approach to measure domain stability score, addressing limitations in Lee et al. (2024). While the previous approach only considered the gap between the highest and lowest scores, our method accounts for overall performance variability. To achieve this, we use the coefficient of variation (CV), which provides a normalized measure of dispersion by comparing the standard deviation to the mean. We calculate CV for three evaluation dimensions respectively across domains as follows below formulation.

For a given performance score across domains $S_E = \{S_{E,1}, S_{E,2}, ..., S_{E,d}\}$, where $d = 1, ...6$ denotes the domain index and E refers faithfulness, completeness, and conciseness. We calculate the Instability score for each evaluation dimension S_E:

$$\text{Instability}(S_E) = \frac{\sigma_{S_E}}{\mu_{S_E}} \times 100,$$

where σ_{S_E} is the standard deviation and μ_{S_E} is the mean of the scores for dimension E across domain. To formulate Domain Stability, we rescaled the Instability(S_E) to ensure the stability measure remains within a meaningful range. Specifically, we define *Domain Stability* as:

$$\text{Domain Stability}(S_E) = \frac{100}{1 + \text{Instability}(S_E)}$$

For the composite score, we average the Domain Stability of the three evaluation dimensions.

Language Stability Language stability evaluates how consistently a model performs across different languages. We apply the same mathematical framework as Domain Stability, using the coefficient of variation (CV) and rescaling approach. For each evaluation dimension E, we first compute $S_{E,L}$, the average score across domains for each language L:

$$S_{E,L} = \frac{1}{D} \sum_{d=1}^{D} S_{E,L,d},$$

where $S_{E,L,d}$ represents the score for evaluation dimension E in language L and domain d, and D is the total number of domains. Then we calculate the instability score with CV, which captures the performance fluctuation across different languages:

$$\text{Instability}(S_E) = \frac{\sigma_{S_E}}{\mu_{S_E}} \times 100,$$

where σ_{S_E} is the standard deviation and μ_{S_E} is the mean of the scores for dimension E across languages.

Similar to Domain Stability, we transform the Instability score to ensure that the stability measure remains within a meaningful range:

$$\text{Language Stability}(S_E) = \frac{100}{1 + \text{Instability}(S_E)}$$

The composite score is calculated by averaging the Language Stability across three evaluation dimensions. Similar to domain stability, it reflects the model's performance variations across different languages.

E.2 Evaluation Metrics for Automatic Evaluator

Pearson Correlation We compute summary-level correlations using the Pearson correlation coefficient to assess the alignment between automated and human evaluation, following recent work (Liu et al., 2023; Song et al., 2024; Lee et al., 2024). For each summary i, we analyze the correlation between automated evaluation scores (x_i) and human evaluation scores (y_i). The summary-level correlation ρ is calculated as:

$$\rho = Cor([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n])$$

where Cor represents the Pearson correlation coefficient, where n is the total number of summaries.

F Self-Evaluation Bias Detail

To investigate self-evaluation bias in summary evaluation, we employ four large language models (GPT-4o, Claude-3.5-Sonnet, Llama-3.1-70B, and Qwen-2.5-72B) as both evaluators and summarizers. We measure self-evaluation bias by comparing how each model rates its own summaries versus those generated by other models.

First, we calculate composite scores by averaging three evaluation dimensions (faithfulness, completeness, and conciseness). Second, each model's self-evaluation score is obtained by averaging the composite scores of its own summaries across different domains. Similarly, peer-evaluation scores are computed by averaging the composite scores given by other models. Third, self-evaluation bias is then determined by subtracting the peer-evaluation scores from the self-evaluation scores. To assess statistical significance, we conduct the t-test comparing self-evaluation and peer-evaluation scores. Biases that are statistically significant ($p < 0.05$) are highlighted in the visualizations in Figure 6.

G Additional Analysis

G.1 Human Annotation Result Detail

We provide the eight summarizer performances across all six domains and languages of faithfulness, completeness, and conciseness in Tables 27 - 29. We present additional domain-level findings.

Faithfulness Score Table 27 shows proprietary LLMs outperform open-source and non-LLMs across domains and languages. Specifically, Claude-3.5-Sonnet achieves optimal performance stability among proprietary LLMs across all domains and languages. A marginal decline in faithfulness is observed in News, Booking and Meeting domains, while an increase is noted in Report and Interview domains from English to Chinese. Therefore, further research efforts should target both language-specific enhancements and domain-language adaptation strategies.

Completeness Score Table 28 shows proprietary LLMs' superior performance across all domains and languages, with Gemini-1.5-Pro excelling in Report, Booking, and Interview domains. However, a decline in completeness scores is observed across all domains and models from English to Chinese, with varying degrees of intensity depending on the domain. The observed decline in completeness

scores indicates challenges in retaining essential information in Chinese, emphasizing the need for domain-aware language improvements to enhance performance.

Conciseness Score Table 29 demonstrates proprietary LLMs achieve higher conciseness scores compared to open-source and non-LLMs across domains. While the overall decline in conciseness from English to Chinese is minimal, the Meeting domain experiences a significant drop, with open-source LLMs performing the worst in this domain compared to others. This highlights the need for domain-specific refinements to improve conciseness in open-source LLMs.

G.2 Comparison with QA- and NLI-based Auto-evaluator

In Table 30, we provide the performance of automated evaluators, including QA-based (QAFactEval (Fabbri et al., 2022)), NLI-based (AlignScore (Zha et al., 2023), MiniCheck (Tang et al., 2024a)), and LLM-based (G-Eval (Liu et al., 2023), FineSurE (Song et al., 2024)) with implementations based on GPT-4o.

The QA-based and NLI-based evaluators specialize in measuring faithfulness but have limitations in assessing completeness and conciseness. Therefore, we focus on comparing their faithfulness scores across six domains. We compare the scores obtained from automatic evaluation with human-annotated labels in `MSumBench`. Specifically, we report Pearson correlation values for the summary-level faithfulness percentage scores. Due to the language limitations of these evaluators, we only provide results for English summaries. Among all methods, FineSurE achieves the highest domain stability (83.75) and demonstrates superior performance in four domains such as Report, Booking, Meeting, and Interview.

G.3 Detailed Performance of LLM-Based Auto-Evaluator

Table 31 compares the performance of two LLM-based automatic evaluators, G-Eval and FineSurE, across two model types, six domains, and two languages. We report the average scores of GPT-4o and Claude-3.5-Sonnet for proprietary LLMs and Llama-3.1-70B and Qwen-2.5-72B for open-source LLMs. While FineSurE, implemented with proprietary LLMs, often demonstrates strong performance across three evaluation dimensions, its effectiveness varies across domains, particularly in faithfulness evaluation. In some cases, such as the Booking domain, G-Eval (using open-source LLMs) surpasses FineSurE by a significant margin. This highlights the inherent variability in LLM-based evaluation accuracy and reinforces that no single LLM or automatic evaluator consistently outperforms others across all dimensions, domains, and languages.

G.4 Comparison with Similarity-based Metric

We analyze the summary-level agreement between human scores and conventional similarity-based metrics, including ROUGE-1/2/L (Lin, 2004) and BERTScore (Zhang et al., 2019), across three evaluation dimensions and multiple domains. We compare these results with the LLM-based method, FineSurE (Song et al., 2024), an LLM-based evaluation approach implemented using GPT-4o.

Our analysis reveals distinct performance patterns across evaluation dimensions. For faithfulness evaluation, conventional metrics exhibit weaker correlations with human judgments compared to FineSurE across all domains and languages. While completeness and conciseness exhibit stronger correlations than faithfulness, their performance varies depending on the domain.

Nevertheless, conventional metrics generally demonstrate significantly lower agreement with human scores across all dimensions compared to the LLM-based evaluator.

H Human Annotation Details

Qualifications and Compensation MTurk annotators are screened through an English proficiency test simulating the fact verification and key-fact alignment tasks. They must also demonstrate a reliable track record at MTurk with a minimum 90% approval rate and 500 accepted HITs. These crowdsourced annotators receive compensation exceeding the U.S. minimum wage.

For manual annotation (see Section 4.2), we recruit postgraduate students with proficient English skills (above C2 level) as expert examiners. Specifically, for English-Chinese translation tasks (see Section 3) and for any manual annotation involving Chinese, we require the expert examiners to be native Chinese speakers. These experts are compensated at rates exceeding $30 per hour plus performance-based incentives.

Annotation of Chinese Summaries To annotate Chinese summaries, the annotators are provided with summaries translated into English using GPT-4o. These translations undergo our standard process: initial LLM-based translation, validation, and final verification by native Chinese examiners, as described in Section 3.1. The annotators then work with these verified English translations.

Quality Control In addition, since MTurk is a crowd-sourcing platform, it is essential to systematically filter unreliable answers from the annotators. For each summary annotated, we designate about 5-10% of the annotation unit in a Human Intelligence Task as attention checks, where the correct answers are already known to us. We exclude all responses that do not pass the attention checks. This approach ensures that the annotations collected from MTurk meet the required standards.

You are an expert English-Chinese translator specialized in news articles. Your task is to translate the following English text to Chinese (Simplified Chinese), sentence by sentence, with careful attention to quality and accuracy.

Warning: Use only "standard Simplified Chinese characters" and English technical terms when necessary

Translation Rules:

1. Reference Consistency
    - Keep organization names in original form
    - Translate ALL PERSON NAMES to Chinese following appropriate conventions:
        - Western names: Use standard Chinese transliteration
            - Example: Michael → 迈克尔 (Màikè'ěr), John → 约翰 (Yuēhàn)
        - Chinese names: Maintain Chinese characters
            - Keep family name and given name format (e.g., 王小明)
        - For established figures, use their commonly known Chinese name
            - Example: Shakespeare → 莎士比亚 (Shāshìbǐyà)
    - Use standard format for dates, times, and numbers
2. Technical Terms
    - Use established Chinese technical terms
    - First mention: Chinese term (English term); Following mentions: Chinese term only
    - Maintain consistency in specialized terms throughout
3. Cultural Adaptation
    - Translate English idioms and proverbs to Chinese cultural equivalents (成语 when appropriate)
    - Convert Western business expressions to match Chinese business etiquette:
        - Use appropriate level of formality (敬语)
        - Follow Chinese business conversation conventions
    - Maintain neutrality and objectivity in expression
    - Add brief explanations for culturally specific items
4. Chinese Writing Style Consistency
    - Use formal written Chinese (书面语) consistently
    - Avoid mixing formal and colloquial expressions
    - Follow standard news writing conventions:
        - Use proper 判断词 and 状态词
        - Use standard news article punctuation
    - Word choice guidelines:
        - Prefer 因为 over 由于 for causation
        - Use 表示 instead of 说 for formal statements
        - Choose 认为 over 觉得 for opinions

Provide your answer ONLY in this simple JSON format:
{"translation": ["First Chinese translation", "Second Chinese translation", "Third Chinese translation", ..., "Last Chinese translation"]}

English text: {input_text}

Table 12: Chinese translation prompt for news domain.

You are an expert English-Chinese translator specialized in medical and scientific literature. Your task is to translate the following English text to Chinese (Simplified Chinese), sentence by sentence, with careful attention to quality and accuracy.

Warning: Use only "standard Simplified Chinese characters" and English technical terms when necessary

Translation Rules:

1. Medical Terminology
    - Use standardized Chinese medical terms (规范医学用语)
    - Keep precision in medical concepts:
        - Diseases: Standard Chinese names (英文名)
        - Medications: Generic names in Chinese (英文通用名)
        - Medical procedures: Standard translations
    - Handle technical terms:
        - First mention: Chinese term (English term); Following mentions: Chinese term only
    - Maintain absolute consistency in terminology throughout
    - Abbreviations: Keep standardized format
        - First mention: Full Chinese term (Full English term [Abbreviation]) (聚合酶链式反应: Polymerase Chain Reaction [PCR])
    - Special terms:
        - Gene/protein names: Follow standard conventions
        - Chemical formulae: Maintain accuracy
        - Anatomical terms: Use standard translations
2. Academic Writing Style
    - Use formal academic Chinese (学术用语)
    - Follow scientific writing conventions: use precise and objective language and maintain the scientific tone
    - Sentence structure: Clear and concise, logical flow, one key point per sentence
    - Word choice guidelines:
        - Use 发现 for findings
        - Use 显示 for results presentation
        - Use 表明 for conclusions
        - Use 证实 for verification
        - Use 提示 for implications
        - Use 比较 for comparison
        - Use 分析 for analysis

Provide your answer ONLY in this simple JSON format:
{"translation": ["First Chinese translation", "Second Chinese translation", "Third Chinese translation", ..., "Last Chinese translation"]}

English text: {input_text}

Table 13: Chinese translation prompt for medical literature domain.

You are an expert English-Chinese translator specialized in reports and official documents. Your task is to translate the following English text to Chinese (Simplified Chinese), sentence by sentence, with careful attention to quality and accuracy.

Warning: Use only "standard Simplified Chinese characters" and English technical terms when necessary

Translation Rules:

1. Reference Consistency
    - Keep organization names in original form
    - Translate ALL PERSON NAMES to Chinese following appropriate conventions:
        - Western names: Use standard Chinese transliteration (Michael → 迈克尔 (Màikè'ěr))
        - Chinese names: Maintain Chinese characters
    - For established figures, use their commonly known Chinese name (Shakespeare → 莎士比亚 (Shāshìbǐyà))
    - Use standard Chinese format for:
        - Dates: YYYY年MM月DD日
        - Numbers: Use Chinese numerals for formal documents (一 二 三)...)
        - Percentages: Use 百分之 format
        - Monetary values: Follow Chinese currency notation
2. Technical Terms
    - Use established Chinese technical terms
    - First mention: Chinese term (English term); Following mentions: Chinese term only
    - Maintain consistency in specialized terms throughout
3. Cultural Adaptation
    - Translate English idioms and proverbs to Chinese cultural equivalents (成语 when appropriate)
    - Convert Western business expressions to match Chinese business etiquette:
        - Use appropriate level of formality (敬语)
        - Follow Chinese business conversation conventions
    - Maintain neutrality and objectivity in expression
    - Add brief explanations for culturally specific items
4. Chinese Writing Style Consistency
    - Use the highest level formal written Chinese (政府公文体)
    - Follow official document writing conventions:
        - Use standard official vocabulary (规范用语)
        - Apply proper ceremonial words (礼仪用语)
    - Use formal written Chinese (书面语) consistently
    - Avoid mixing formal and colloquial expressions
    - Sentence structure:
        - Use parallel structure for lists (排比句)
        - Maintain appropriate formality level
        - Use standard government document punctuation

Provide your answer ONLY in this simple JSON format:
{"translation": ["First Chinese translation", "Second Chinese translation", "Third Chinese translation", ..., "Last Chinese translation"]}

English text: {input_text}

Table 14: Chinese translation prompt for report domain.

You are an expert English-Chinese translator specialized in dialogue scripts. Your task is to translate the following English script to Chinese (Simplified Chinese), sentence by sentence, with careful attention to quality and accuracy.

Warning: Use only "standard Simplified Chinese characters" and English technical terms when necessary.

Translation Rules:

1. Reference Consistency
    - Keep organization names in original form
    - Translate ALL personal names to Chinese following appropriate conventions:
        - Western names: Use Chinese transliteration (James → 詹姆斯)
        - Chinese names: Keep Chinese characters
        - For established figures, use their commonly known Chinese name (Shakespeare → 莎士比亚 (Shāshìbǐyà))
    - The same Chinese translation must be used when referring to that person within any dialogue.
2. Technical Terms
    - Use established Chinese technical terms
    - First mention: Chinese term (English term); Following mentions: Chinese term only
3. Speaking Style & Format
    - Keep "Speaker: dialogue" format
        - Place actions in parentheses
        - Maintain conversation flow
    - Use appropriate formal Chinese based on context
    - Keep each speaker's tone consistent
    - se proper conversational particles (吧,呢,啊)
    - Adapt greetings and courtesies to Chinese norms(您好,请问,麻烦您)

Provide your answer ONLY in this simple JSON format:
{"translation": ["First Chinese translation", "Second Chinese translation", "Third Chinese translation", ..., "Last Chinese translation"]}
English text: {input_text}

Table 15: Chinese translation prompt for booking/meeting/interview domain

You are an expert English-Chinese translator specialized in handling precise information across various domains. Your task is to translate the following English key facts into Chinese (Simplified Chinese), sentence by sentence, with careful attention to quality, accuracy and consistency. Read the given document carefully, fully understand it, and keep that in mind when you translate key fact sentences. Use all terms already written in Chinese from the document to ensure consistency across the translation. Follow the instructions to translate English key fact sentences.

Warning: Use only "standard Simplified Chinese characters" and English technical terms when necessary.

Translation and Verification Rules:

1. Reference Consistency
    - Keep organization names in original form
    - Translate ALL PERSON NAMES to Chinese following appropriate conventions:
        - Western names: Use standard Chinese transliteration
            - Example: "Michael" → "迈克尔" (Màikè'ěr), "John" → "约翰" (Yuēhàn)
        - Chinese names: Maintain Chinese characters
            - Keep family name and given name format (e.g., 王小明)
    - For established figures, use their commonly known Chinese name
        - Example: Shakespeare → 莎士比亚 (Shāshìbǐyà)
    - Use standard format for dates, times, and numbers
2. Technical Terms
    - Use established Chinese technical terms
    - Follow terms already defined in the document for consistency.
    - First mention: Chinese term (English term); Following mentions: Chinese term only
    - Maintain consistency in specialized terms throughout
3. Focus on Information
    - Prioritize the accurate transfer of factual information in each key fact.
    - Avoid any stylistic adjustments or embellishments. Translate the text plainly and faithfully.
4. Back-Translation for Verification:
    - For each translated Chinese sentence:
        - Perform a back-translation into English.
        - Compare the back-translation with the original English key fact.
    - If there is any difference in meaning, revise the Chinese translation and repeat Steps 1–3 until the back-translation aligns with the original English sentence.

Provide your answer in JSON format:
"translation": [("1", "Chinese translation"), ("2", "Chinese translation"), ..., ("Key fact number", "Chinese translation")]

Document:

{input_text}

{N} Key fact sentences:

{key_facts}

Table 16: Chinese translation prompt for key-fact.

You are an expert English-Chinese translator with extensive experience in translation quality assessment. Your task is to check the quality of the English-Chinese translation, sentence by sentence, with careful attention to quality and accuracy.
Quality check instructions:

1. Accuracy
    - Compare the English source text and Chinese translation to ensure meaning is preserved
    - Check for any omissions or additions
    - Verify numerical values, dates, and proper names are correctly translated
    - Flag any mistranslations or semantic errors
2. Consistency
    - Reference Consistency: Check if proper nouns, organization names, and product names are translated consistently
    - Technical Term Consistency: Verify industry-specific terminology is translated consistently and correctly
    - Style Consistency: Ensure consistent tone and level of formality throughout
3. Fluency
    - Check if the translation reads naturally in Chinese
    - Verify proper Chinese grammar and syntax
    - Ensure appropriate sentence structure and flow
    - Check for any awkward expressions or unnatural phrasing
4. Readability
    - Assess if the text is easy to understand for the target audience
    - Check sentence length and complexity
    - Verify proper paragraph breaks and text organization
    - Ensure clear logical flow
5. Cultural Appropriateness
    - Check for cultural sensitivity
    - Verify idioms and expressions are appropriately localized
    - Ensure measurements, dates, and currencies are properly converted
    - Flag any potential cultural misunderstandings
6. Professionalism
    - Verify appropriate formal/business language usage
    - Check for proper honorific forms
    - Ensure professional terminology is correctly used
    - Maintain appropriate level of formality
7. Fitness for Purpose
    - Verify the translation meets its intended purpose
    - Check if appropriate for target audience
    - Ensure industry-specific requirements are met
    - Verify technical accuracy for specialized content

Provide the answer using the following JSON format:
{"translation": [(1, "True", "IF True Blank", "IF True Blank", "IF True Blank"), (2, "False", "IF False English Sentence", "IF False Chinese Sentence", "IF False Reason"), ..., (N, "True", "", "", "")] }
English text: {input_text}

Chinese text: {translation_text}

Table 17: Translation qaulity check prompt.

Domain	Category	Description
News	Main Topic	General information about the primary event, issue, or occurrence being discussed
	Background	Situational details supporting the main topic
	Immediate Impact	Short-term effects or consequences resulting from the main topic
	Future Implications	Long-term outcomes or projected developments related to the main topic
	Public statements	Non-expert perspectives, opinions, or reactions from the general public
	Official statements	Expert or authoritative opinions, assessments, or analysis on the main topic
	Counterarguments	Critiques or opposition to the main topic or its impacts
Medical Literature	Research Finding	Key discoveries or outcomes from medical research studies
	Medical experiments	Detailed procedures, methodologies, and designs of experiments or clinical trials testing treatments or interventions
	Disease descriptions	Detailed explanations of diseases, including symptoms, causes, and characteristics
	Medical Treatment	Recommended therapies, treatments, and interventions aimed at managing or curing diseases
	Medical Prevention	Strategies and actions aimed at preventing diseases and promoting public health
Report	Recommendations	Suggested actions or improvements based on report findings
	Governance	Oversight and administration of programs and resources
	Regulation and Policy	Legal frameworks, standards, and policies guiding operations
	Evaluations	Assessment of data and review of program performance
	Financial information	Details about costs, budget allocations, and financial impacts
Booking	General Information	Reference numbers, contact info, headcounts, or else excluding time, location, and price
	Price and Payment	Cost, pricing, fees and payment methods
	Time and Schedule	Time slots, schedules, and booking times
	Location and Route	Address, location and routes
	Booking Confirmation	Confirmation and status of bookings for the services
	User Requests	User's requests for the booking service or details
	System Suggestions	Suggestions provided by the system.
Meeting	Opinions	Personal viewpoints, perspectives, or feelings regarding a topic or proposal
	Decisions	specific, actionable plans or choices made to address a Problem or improve a situation after careful consideration
	Proposals	Final conclusion or choice made after discussion, determining the course of action
	Reports	Structured updates or presentations summarizing status, data, or findings related to projects or objectives
	Factual Information	Objective data, statistics, or verified information that serves as a basis for decisions or discussions
Interview	Background	Context or historical info to help understanding
	Main Arguments	Core claims or opinions from each speaker
	Supporting Examples	Examples, data, or statistics to support the main arguments
	Counterarguments	Opposing views or criticisms of the main arguments and responses
	Conclusions	Key points and future directions from the interview

Table 18: Description of key-fact category for each domain.

Domain	Category	Example
News	Main Topic	Prince William is scheduled at the Cenotaph.
	Background	Guy Thorpe-Beeston has 18 years' experience.
	Immediate Impact	William will travel by car, not helicopter.
	Future Implications	The delivery room at St Mary's Lindo Wing will have a new team.
	Public statements	Prosecutor Brice Robin confirmed no videos were used.
	Official statements	Obstetrician Guy Thorpe-Beeston will lead the birth.
	Counterarguments	Prince William faces a conflict between duties and his child's birth.
Medical Literature	Research Finding	Intervention strategies have helped reduce malaria globally.
	Medical experiments	New mapping techniques track malaria transmission.
	Disease descriptions	Dengue is caused by four viruses (DENV 1-4).
	Medical Treatment	Dengue treatment involves haematological monitoring.
	Medical Prevention	Insecticide-treated bednets have been key to malaria control.
Report	Recommendations	HHS agreed to gather more disenrollment data.
	Governance	CMS reviews aspects of contracts between states and D-SNPs.
	Regulation and Policy	Dual-eligible beneficiaries qualify for Medicare and Medicaid.
	Evaluations	CMS lacks data on disenrolled beneficiaries.
	Financial information	Medicaid covers premiums for dual-eligible beneficiaries.
Booking	General Information	The train ticket number is 12345.
	Price and Payment	The total cost for the train ticket is $45.
	Time and Schedule	The restaurant is available from 6 PM to 9 PM.
	Location and Route	The Grand Hotel is at 123 Kings Road, Brighton, BN1 2GS.
	Booking Confirmation	A table for 8 is booked.
	User Requests	The user wants to book a hotel room for two people.
	System Suggestions	The system recommends taking a train to Cambridge.
Meeting	Opinions	Dave Shukla argues consolidation increases expenses.
	Decisions	The council decided to consolidate gas under one commission.
	Proposals	Dave Shukla proposes keeping utilities independent.
	Reports	The goal is to create one commission for cost savings.
	Factual Information	Reid concludes the party's divisiveness cost them control.
Interview	Background	The discussion covers Obama's State of the Union.
	Main Arguments	Cary says establishment Republicans struggle against Trump.
	Supporting Examples	Cary cites Republican grassroots efforts in early states.
	Counterarguments	Cary defends the Republican Party, claiming not all are hateful.
	Conclusions	Reid concludes the party's divisiveness cost them control.

Table 19: Example of key-fact category for six domains.

Your task is to identify domain-specific key facts within the document in <document></document>, which are essential pieces of information for a high-quality summary. You are provided key-fact category and description in <category></category> tags, and its example in <example></example>. Each key fact must be presented as a standalone, atomic-level sentence. The following is a set of detailed instructions in <instructions></instructions> tags for identifying key facts.

<instructions>

1. Identify Key-facts: Extract all key-facts from the text. Each key-fact should:
    - Be a complete sentence with a subject, verb, and object/complement.
    - Contain only one action, event, or idea. Avoid compound sentences.
    - Include no more than two or three entities per key fact. If there are more than three entities, divide them into separate sentences.
    - Present temporal information (e.g., when, how long) as standalone sentences.
    - Present causal relationships (e.g., reasons, consequences) as standalone sentences.
    - Avoid using linking words like 'and,' 'but,' 'then,' or 'because.' Each idea must be presented as an independent fact.
    - Do not combine related details, even implicitly. Each sentence must describe exactly one action, idea, or relationship.
2. Here are the examples of key-fact structure granularity:
    - Text Example: The resolution authorizes the operation of one winter shelter from December 1st, 2019, to March 31st, 2020.
    - Key-facts (Revised):
        (a) The resolution authorizes the operation of one winter shelter.
        (b) The winter shelter will operate from December 1st, 2019, to March 31st, 2020.
    - Text Example : The property for the winter shelter was purchased using homeless emergency aid program funding.
    - Key-facts (Revised):
        (a) The property for the winter shelter was purchased.
        (b) Homeless emergency aid program funding was used for the purchase.
3. Categorize Key-facts:
    – Define your own categories for the key-facts based on content.
    – Assign each key-fact to a category.
4. Compare Categories:
    – Compare your defined categories with the provided key-fact categories.
    – Adjust any key-fact to better align with the provided categories.
5. Validate Key-facts
    - Ensure each key-fact meets the following criteria:
        – Correctly categorized.
        – Atomicity: Conveys only one action, event, or idea.
        – Clarity: Is concise and clear, avoiding ambiguity.
        – Brevity: Contains no unnecessary details.
        – Non-overlapping: Does not duplicate information in other facts.
    - Finalize key facts that satisfy these conditions.

</instructions>

<category>
Here are the provided Key Fact Categories and Descriptions:
categories
</category>

<example>
Here are the examples of key facts to illustrate the level of granularity for each category:
{category_examples}
</example>

Provide your answer in JSON format. The answer should be a dictionary with tuples:
'key_facts' containing the key facts as a list of tuples (keyfact, reason, category):
{"key_facts": [("first key fact", "reason", "category"), ("second key fact", "reason", "category")]}
<document>
{input_text}
</document>

Table 20: Domain-specific key-fact generation prompt.

You will receive a Document and a set of Key-facts sentences that contain essential pieces of information from the Source Document. Your task is to identify if each Key-Fact Sentence is useful for making a summary of the Source Document.

Reasons a Key-Fact Sentence may NOT be useful:

Reason 1) Trivial Information: The sentence contains correct but insignificant details that do not contribute meaningfully to the main points of the Source Document.

- Example: The Source Document is a business meeting transcript, but the Key-Fact Sentence is "Speaker A said 'Good afternoon.'"
- Explanation: While accurate, this detail doesn't enhance understanding of the meeting's objectives or outcomes.

Reason 2) Incorrect Information: The sentence includes factual errors or discrepancies compared to the Source Document.

- Example: The Source Document mentions 12 children living in a small town, but the Key-Fact Sentence states 11 children in a big city.
- Explanation: The numbers and locations don't match, making the sentence inaccurate.

Reason 3) Irrelevant Information: The sentence is unrelated to the content of the Source Document.

- Example: The Source Document is about artificial intelligence, but the Key-Fact Sentence discusses a dog barking loudly.
- Explanation: The sentence doesn't pertain to the topic of the Source Document.

Reason 4) Category Alignment: The sentence is incorrectly categorized.

- Example: A sentence about future plans is categorized as "Factual Information" when it should be "Proposals"
- Explanation: The content of the sentence doesn't match the characteristics of its assigned category.

Reason 5) Domain Relevance: The information in the sentence, while accurate and properly categorized, is not essential for summarizing this type of document.

- Example: In a meeting minutes document discussing a major company merger, a sentence about routine office maintenance schedule categorizing as "Factual Information"
- Explanation: Although this information is accurate and properly categorized as "Factual Information", it's not essential for understanding the key points of a merger discussion meeting.

Here are the Instructions for Key-facts validation:

1. Read a Document and a set of Key-facts sentences carefully.
2. Evaluate each Key-facts sentence based on the five reasons above.
3. According to evaluation, if the Key-facts sentence is useful for making a summary of the Source Document, response "Yes", otherwise response "No"
4. Provide a single sentence explaining why the Key-facts sentence is useful for making a summary.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence", "response", "reason".
you should provide a response and a reason for all Key-facts sentence: [{"sentence": "first key-fact sentence", "response": "your response", "reason": "your reason"}, {"sentence": "second key-fact sentence", "response": "your response", "reason": "your reason"}, ... , {"sentence": "N-th key-fact sentence", "response": "your response", "reason": "your reason"}]
Document:

{input_text}

{N} Key-facts sentence with Category:

{key_fact}

Table 21: Key-fact validation prompt.

(a) English summary generation prompt

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"}

(b) Chinese summary generation prompt

文本: {input_text}

指令: 请用中文总结这段文字 。

以 JSON 格式提供答案 答案应是一个包含 "summary" 键和值的字典,值为生成的摘要字符串:
{"summary": "总结内容"}

Table 22: Summary generation prompt.

You will be provided with two sets of information:

- List of sentences A: A list of sentences that need to be evaluated.
- sentence B: A sentence that will be checked against sentence A.

Your task is to evaluate whether the complete information in each sentence from List A is fully contained in sentence B. This requires performing n evaluations (where n is the number of sentences in List A). For example, if there are 15 sentences in List A, you have to do 15 evaluations.

Instruction:

- A sentence from List A is considered "contained" in the sentence B if and only if:
  1. The complete information conveyed by the sentence in List A is entirely present in the sentence B.
  2. The essential meaning of the sentence in List A must align with or be fully and explicitly captured by sentence B.
  3. The information conveyed by the sentence in List A must be explicitly implied or fully understood by sentence B.

- Exact wording is not required, but the complete and explicit meaning must match.
- Provide a short reason, and a label: contained, not contained.

Please provide your answer in JSON format:
[{"sentence A": "1", "label": "contained", "reason": "Short explanation of why you chose this label"}, {"sentence A": "sentence A number", "label": "contained or not contained", "reason": "Short explanation of why you chose this label"}, ..., {"sentence A": "N", "label": "not contained", "reason": "Short explanation of why you chose this label"}]
Note:

- Ensure that all n evaluations are performed without omission. Each combination of a sentence from List A and the sentence B must be explicitly evaluated.
- Any skipped or missing evaluations will result in incomplete analysis, so please confirm that no sentence in List A is overlooked.
- Clearly explain your reasoning even if the label is "not contained."
- Ensure your evaluation strictly adheres to the requirement that the entire meaning of the sentence in List A should be fully present in sentence B to be labeled as "contained."
- Criteria recap: contained, not contained

List of sentences A:
{key-fact}

Sentence B:
{summary}

Table 23: Key-fact alignment prompt.

You are the ADVOCATE, an agent defending the factual consistency of the summary. Assume the summary sentences are always true and faithful. Cite specific sentences from the reference document as evidence to support your claim for each summary sentence. You are given the reference document provided in <document></document> tags and the summary sentences in <summary></summary> tags. Critically assess and present your reasoning.

<errors>

- out-of-article error: If the summary sentence introduces facts, subjective opinions, or biases that cannot be verified or confirmed by the reference document, the summary is factually inconsistent with the reference document.
- entity error: If the summary sentence includes incorrect or misrepresented entities, such as names, numbers, or main subjects, the summary is factually inconsistent with the reference document.
- relation error: If the summary sentence contains incorrect semantic relationships, such as the use of wrong verbs, prepositions, or adjectives, which distort the relationships between entities, the summary sentence is factually inconsistent with the reference document.
- sentence error: If the summary sentence completely contradicts the information in the reference document, requiring significant revision or removal to align with the reference document, the summary sentence is factually inconsistent with the reference document.

</errors>

<instructions>

Here is the instructions for writing your arguments:

Follow these steps carefully to ensure a structured and thorough evaluation under your assigned role.

1. Read the reference document and summary sentence under your role:
   - Carefully read the reference document and try to fully understand it.
   - Compare each summary sentence to the reference document to identify evidence supporting its factual consistency.
   - Refer to the error types in <errors></errors> tags to better defend your claim (faithfulness), focusing on areas of alignment and acceptable variations.
2. As a ADVOCATE, focus on finding alignment:
   - Explicitly identify numbered sentences in the reference document that support or partially align with the summary sentence.
   - Even if a perfect match cannot be found, select the closest sentence(s) that contain key elements (entities, relationships, events, quantities, or cause-effect relationships).
   - Always select at least one numbered sentence from the reference document, even if it partially aligns with the summary.
   - As an ADVOCATE, focus on defending the faithfulness of the summary while addressing any potential inconsistencies.
3. Provide a detailed explanation of your arguments:
   - For each summary sentence:
     - Cite one or more numbered sentences from the reference document, even if only partial alignment exists.
     - Use the format "reference_sentence_number": [number1, number2] to explicitly indicate the reference of support.
     - Include a "reason" that explicitly explains how the cited sentences align with the summary, focusing on entities, relationships, and key details.
   - Your explanation must:
     - Be concise.
     - Reference specific elements, such as:
       - Key details: Validate quantities, events, and cause-effect relationships.
       - Rephrasing accuracy: Argue that the summary retains the original meaning despite rephrasing.
       - Word choice: Ensure terms like "only," "significant," or "most" match the intensity or scale of the reference document.

</instructions>

Provide your answers in JSON format as shown below:

[{ "summary_sentence_num": "0", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reference_sentence": ["sentence from reference document", "another sentence from reference document"],
"reference_sentence_number": [0, 1],
"reason": "The summary mentions the entity 'X', but the reference document refers to 'Y' [0]. Additionally, the relationship between 'A' and 'B' is misrepresented, as the reference indicates 'A causes B', not 'A results from B' [1]."},
{"summary_sentence_num": "N", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reference_sentence": ["closest available sentence"],
"reference_sentence_number": [10],
"reason": "The summary states 'Event Z happened', but no reference to 'Event Z' is found in the reference [10]." }]

Reference document, divided into numbered sentences:

<document>

input_text

</document>

Summary with N sentences:

<summary>

summaries

</summary>

Table 24: Fact verification ADVOCATE prompt.

You are the SKEPTIC, an agent identifying and arguing for factual inconsistencies in the summary, considering error types. Assume the summary sentences are always unfaithful. Cite specific sentences from the reference document as evidence to support your claim for each summary sentence. You are given the reference document provided in <document></document> tags and the summary sentences in <summary></summary> tags. Now, follow the instructions in <instructions></instructions> tags. Critically analyze and present your reasoning.

<errors>

- out-of-article error: If the summary sentence introduces facts, subjective opinions, or biases that cannot be verified or confirmed by the reference document, the summary is factually inconsistent with the reference document.
- entity error: If the summary sentence includes incorrect or misrepresented entities, such as names, numbers, or main subjects, the summary is factually inconsistent with the reference document.
- relation error: If the summary sentence contains incorrect semantic relationships, such as the use of wrong verbs, prepositions, or adjectives, which distort the relationships between entities, the summary sentence is factually inconsistent with the reference document.
- sentence error: If the summary sentence completely contradicts the information in the reference document, requiring significant revision or removal to align with the reference document, the summary sentence is factually inconsistent with the reference document.

</errors>

<instructions>

Here is the instructions for writing your arguments:

Follow these steps carefully to ensure a structured and thorough evaluation under your assigned role.

1. Read the reference document and summary sentence under your role:
   - Carefully read the reference document and try to fully understand it.
   - Compare each summary sentence to the reference document to identify evidence supporting its factual consistency.
   - Refer to the list of errors in <errors></errors> tags to find evidence to support your claims(unfaithfulness), paying special attention to the listed error types and acceptable variations.
2. As a SKEPTIC, focus on identifying discrepancies:
   - Explicitly identify numbered sentences in the source document that contradict or fail to align with the summary sentence.
   - Even if a perfect match cannot be found, select the closest sentence(s) that contain key elements (entities, relationships, events, quantities, or cause-effect relationships).
   - Always select at least one numbered sentence from the source document that highlights inconsistencies or raises doubts about the summary.
3. Provide a detailed explanation of your arguments:
   - For each summary sentence:
     - Cite one or more numbered sentences from the reference document, even if only partial alignment exists.
     - Use the format "reference_sentence_number": [number1, number2] to explicitly indicate the reference of support.
     - Include a "reason" that explicitly explains how the cited sentences align with the summary, focusing on entities, relationships, and key details.
   - Your explanation must:
     - Be concise.
     - Reference specific elements, such as:
       * Key details: Validate quantities, events, and cause-effect relationships.
       * Rephrasing accuracy: Argue that the summary retains the original meaning despite rephrasing.
       * Word choice: Ensure terms like "only," "significant," or "most" match the intensity or scale of the reference document.

</instructions>

Provide your answers in JSON format as shown below:

[{ "summary_sentence_num": "0", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reference_sentence": ["sentence from reference document", "another sentence from reference document"],
"reference_sentence_number": [0, 1],
"reason": "The summary mentions the entity 'X', but the reference document refers to 'Y' [0]. Additionally, the relationship between 'A' and 'B' is misrepresented, as the reference indicates 'A causes B', not 'A results from B' [1]."},
{"summary_sentence_num": "N", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reference_sentence": ["closest available sentence"],
"reference_sentence_number": [10],
"reason": "The summary states 'Event Z happened', but no reference to 'Event Z' is found in the reference [10]." }]

Reference document, divided into numbered sentences:

<document>

input_text

</document>

Summary with N sentences:

<summary>

summaries

</summary>

Table 25: Fact verification SKEPTIC prompt.

You are the ADJUDICATOR, an agent tasked with providing the final decision for the faithfulness of the summary by assessing the claims presented by the ADVOCATE and SKEPTIC. You are given the reference document provided in <document></document> tags, summary sentences in <summary></summary> tags and the opposing claims in <claim></claim> tags. Now, follow the instructions in <instructions></instructions> tags and"Make sure to always strive to deeply understand and remember the guidelines in <note></note> tags. Think critically and articulate your final decision.

<note>

1. Faithfulness measures how accurately a summary sentence reflects the source document's information and content.
2. The summary sentence should not have to use exact wording in the reference document as long as the original meaning is preserved.
3. The summary sentence can paraphrase and use alternative expressions with preserving the original meaning.
4. The summary sentence is factually consistent even if it omits specific details-one, some or all from reference document.
5. The summary sentence is factually consistent even if it omits specific details-one, some or all from reference document.
6. The summary sentence is factually consistent even if it modifies the level of specificity (using broader terms instead of detail and specific information, or more specific terms instead of the broader terms), maintaining the original information.
7. The summary sentence is factually consistent even if it combines multiple pieces of information from different parts of the text maintaining the original meaning without contradiction.
8. Even if the summary sentence draws reasonable implications, logical conclusions, or appropriate generalizations, it remains factually consistent with the reference document as long as these are explicitly supported by the original meaning.

</note>

<instructions>

1. Read the reference document and summary:
   - Carefully review the reference document and the summary sentence provided.
   - Develop a comprehensive understanding of both the reference document and summary sentence and how it's been summarized.
2. Evaluate the validity of agent arguments:
   - Compare both agents' reasoning critically.
   - Validate the claims align with the reference document and avoid unsupported speculation.
   - Ensure the claims follow the guidelines in <note></note> tags.
3. Finalize your own judgment of the summary sentences.
   - Make a final decision on whether the summary sentence is factually consistent with the reference document, based on your understanding of reference document, summary sentence and the validation of the two opposing claims.
4. Provide your final decision as error type and label
   - Assign an error type and label as follows:
   - error_type: refer to the error_type listed below:
     - no error: no error found, and the summary is factually consistent with the reference document.
     - out-of-article error: If the summary sentence introduces facts, subjective opinions, or biases that cannot be verified or confirmed by the reference document, the summary is factually inconsistent with the reference document.
     - entity error: If the summary sentence includes incorrect or misrepresented entities, such as names, numbers, or main subjects, the summary is factually inconsistent with the reference document.
     - relation error: If the summary sentence contains incorrect semantic relationships, such as the use of wrong verbs, prepositions, or adjectives, which distort the relationships between entities, the summary sentence is factually inconsistent with the reference document.
     - sentence error: If the summary sentence completely contradicts the information in the reference document, requiring significant revision or removal to align with the reference document, the summary sentence is factually inconsistent with the reference document.
   - label:
     - 1: faithfulness, assigned if the summary sentence has no error.
     - 0: unfaithfulness, assigned if the summary contains any error (out-of-article, entity, relation, or sentence errors).
   - In one or two sentences, explain why you agree with one agent's argument and disagree with the opposing agent's argument.

</instructions>

Provide your answers in JSON format as shown below:

[{ "summary_sentence_num": "0", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reason": "write reason of your decision briefly and concisely"},
{"summary_sentence_num": "N", "label": "1 (if faithfulness 1, else 0)", "error_type": "no error (if faithfulness no error)",
"reason": "write reason of your decision briefly and concisely" }]

Reference document, divided into numbered sentences:
<document>
{input_text}
</document>
Summary with N sentences:
<summary>
{summaries}
</summary>
Here are the claims provided by ADVOCATE and SKEPTIC on for each summary sentence:
<claim>
{claim}
</claim>

Table 26: Fact verification ADJUDICATOR prompt.

Model Type	Summarizer	English							Chinese							Language Stability
		News	Med Lit	Report	Booking	Meeting	Interview	Domain*	News	Med Lit	Report	Booking	Meeting	Interview	Domain*	
Proprietary LLMs	GPT-4o	86.30	80.10	82.50	95.10	79.90	92.70	92.97	87.10	78.01	80.00	88.20	56.70	80.50	87.31	93.81
	Claude 3.5 $_{\text{Sonnet}}$	**96.20**	83.70	89.50	94.40	**80.60**	92.10	**93.58**	**88.90**	**86.27**	73.20	84.00	**75.40**	**88.30**	**92.45**	94.75
	Gemini 1.5 $_{\text{Pro}}$	85.60	**86.90**	88.30	**97.40**	74.70	84.20	92.18	87.00	84.64	82.80	**92.00**	60.60	74.50	87.74	**95.21**
	Average	89.37	83.57	86.77	95.63	78.40	89.67	92.91	87.67	82.97	78.67	88.07	64.23	81.10	89.17	94.59
Open source LLMs	Gemma 2 $_{\text{27B}}$	94.90	84.10	84.00	93.60	72.20	93.30	90.87	87.00	81.30	72.90	77.20	66.80	86.00	90.94	93.24
	Llama 3.1 $_{\text{70B}}$	87.70	76.40	**92.90**	89.30	78.50	76.20	91.93	84.90	77.70	72.00	79.00	67.10	81.90	92.18	94.66
	Qwen 2.5 $_{\text{72B}}$	95.10	81.30	91.40	91.50	73.70	**94.70**	91.11	82.40	86.07	86.00	80.20	62.50	84.60	89.92	93.95
	Average	92.57	80.60	89.43	91.47	74.80	88.07	91.30	84.77	81.69	76.97	78.80	65.47	84.17	91.01	93.95
Non LLMs	mT5	22.00	12.00	2.00	34.00	4.00	0.00	48.02	68.00	18.00	32.00	2.00	14.00	18.00	52.41	67.20
	BART	93.90	76.30	83.00	76.00	50.30	85.90	83.87	-	-	-	-	-	-	-	-
	Average	57.95	44.15	42.50	55.00	27.15	42.95	65.94	-	-	-	-	-	-	-	

Table 27: Faithfulness score across six domains and two languages. Domain*: domain stability

Model Type	Summarizer	English							Chinese							Language Stability
		News	Med Lit	Report	Booking	Meeting	Interview	Domain*	News	Med Lit	Report	Booking	Meeting	Interview	Domain*	
Proprietary LLMs	GPT-4o	55.73	**41.22**	40.25	68.19	50.10	44.65	**82.48**	55.80	32.63	33.83	47.58	39.38	37.92	82.24	87.95
	Claude 3.5 $_{\text{Sonnet}}$	**65.49**	39.96	42.88	68.77	48.40	50.51	81.61	55.89	**41.06**	35.20	71.43	46.87	47.17	79.58	95.93
	Gemini 1.5 $_{\text{Pro}}$	64.43	39.92	**44.43**	**78.35**	48.40	**53.30**	79.33	**57.74**	38.87	**48.49**	**72.20**	**51.43**	**52.54**	**82.88**	**98.38**
	Average	61.88	40.37	42.52	71.77	48.97	49.49	81.14	56.48	37.52	39.17	63.74	45.89	45.88	81.57	94.09
Open source LLMs	Gemma 2 $_{\text{27B}}$	52.26	23.67	25.86	56.95	35.83	27.77	72.19	43.23	23.65	21.43	51.25	23.72	29.77	72.41	90.93
	Llama 3.1 $_{\text{70B}}$	51.96	29.34	35.91	51.66	29.48	36.76	79.25	24.59	21.71	21.26	35.06	20.46	25.60	82.05	75.84
	Qwen 2.5 $_{\text{72B}}$	55.81	37.73	39.58	70.12	**55.24**	46.74	80.79	47.69	34.13	33.17	61.48	34.35	34.84	78.15	86.74
	Average	53.34	30.25	33.78	59.58	40.18	37.09	77.41	38.50	26.50	25.29	49.26	26.18	30.07	77.54	84.50
Non LLMs	mT5	9.66	0.67	1.68	6.09	3.07	1.56	52.36	9.97	0.00	1.67	0.97	2.12	2.43	44.34	83.51
	BART	32.97	14.93	13.15	45.49	22.74	20.79	67.16	-	-	-	-	-	-	-	-
	Average	21.32	7.80	7.42	25.79	12.91	11.18	59.76	-	-	-	-	-	-	-	

Table 28: Completeness score across six domains and two languages. Domain*: domain stability

Model Type	Summarizer	English							Chinese							Language Stability
		News	Med Lit	Report	Booking	Meeting	Interview	Domain*	News	Med Lit	Report	Booking	Meeting	Interview	Domain*	
Proprietary LLMs	GPT-4o	85.73	83.40	81.20	74.76	73.51	69.27	92.39	84.20	**79.85**	77.75	74.16	65.00	68.91	91.33	97.30
	Claude 3.5 $_{\text{Sonnet}}$	90.73	74.34	80.59	87.26	72.18	75.43	91.42	84.72	79.05	73.56	**89.33**	66.71	75.05	90.57	**98.23**
	Gemini 1.5 $_{\text{Pro}}$	88.59	83.18	79.71	**89.90**	69.88	**80.01**	91.86	88.07	74.12	77.71	87.35	**71.35**	**77.98**	**92.04**	97.90
	Average	88.35	80.31	80.50	83.97	71.86	74.90	91.89	85.66	77.67	76.34	83.61	67.69	73.98	91.31	97.81
Open source LLMs	Gemma 2 $_{\text{27B}}$	88.07	70.69	72.47	75.13	69.40	66.23	90.58	81.47	71.46	59.35	82.00	52.47	71.36	85.49	96.22
	Llama 3.1 $_{\text{70B}}$	**93.03**	79.33	86.05	86.20	67.87	75.35	90.05	85.07	73.72	72.22	85.67	66.35	67.73	89.95	94.71
	Qwen 2.5 $_{\text{72B}}$	91.63	**84.38**	**87.29**	87.67	**83.02**	75.20	**93.81**	**91.53**	78.06	**86.53**	83.33	66.86	70.73	89.37	95.59
	Average	90.91	78.13	81.94	83.00	73.43	72.26	91.48	86.02	74.41	72.70	83.67	61.89	69.94	88.27	95.51
Non LLMs	mT5	46.00	8.00	28.00	42.00	20.00	28.00	67.17	80.00	0.00	36.00	12.00	18.00	20.00	49.54	97.55
	BART	89.00	81.00	86.00	88.00	60.00	66.90	86.58	-	-	-	-	-	-	-	-
	Average	67.50	44.50	57.00	65.00	40.00	47.45	76.88	-	-	-	-	-	-	-	-

Table 29: Conciseness score across six domains and two languages. Domain*: domain stability

Model Type	Evaluator	News	Medical Lit.	Report	Booking	Meeting	Interview	Domain Stability
QA-Based	QAFactEval	0.53	0.53	0.37	0.36	0.33	0.44	83.08
NLI-Based	AlignScore	**0.73**	0.73	0.63	0.28	0.56	0.51	77.58
	MiniCheck	0.59	**0.75**	0.56	0.36	0.41	0.42	77.72
LLM-Based	G-Eval	0.52	0.67	0.67	0.37	0.38	0.46	79.10
	FineSurE	0.42	0.62	**0.69**	**0.48**	**0.70**	**0.60**	**83.75**

Table 30: Correlation between automatic evaluator and human evaluation.

Dimension	Model Type	Automatic Evaluator	English							Chinese							Language Stability
			News	Medical Lit.	Report	Booking	Meeting	Interview	Domain*	News	Medical Lit.	Report	Booking	Meeting	Interview	Domain*	
Faithfulness	Proprietary	G-Eval	0.62	0.61	0.62	0.75	**0.60**	0.66	91.19	0.26	0.52	0.44	0.62	0.43	0.53	78.29	**94.17**
	LLMs	FineSurE	**0.72**	0.64	**0.85**	0.64	0.58	**0.78**	87.29	**0.39**	**0.64**	**0.57**	**0.74**	**0.52**	**0.66**	**82.18**	88.71
	Open-source	G-Eval	0.69	**0.65**	0.72	**0.78**	0.56	0.67	**92.46**	0.32	0.52	**0.57**	0.73	0.42	0.54	76.33	87.77
	LLMs	FineSurE	0.68	0.54	0.65	0.63	0.57	0.75	88.11	0.37	0.52	0.54	0.66	0.44	0.60	81.80	87.00
Completeness	Proprietary	G-Eval	0.62	0.61	0.62	0.75	0.60	0.66	90.56	0.71	0.40	0.67	0.66	0.64	0.65	77.93	89.75
	LLMs	FineSurE	**0.79**	**0.76**	**0.81**	**0.89**	**0.83**	**0.82**	**94.79**	**0.83**	**0.75**	**0.75**	**0.90**	**0.70**	**0.83**	**90.56**	**95.27**
	Open-source	G-Eval	0.69	0.65	0.72	0.78	0.56	0.67	92.25	0.69	0.46	0.59	0.75	0.57	0.59	82.25	92.19
	LLMs	FineSurE	0.66	0.54	0.61	0.81	0.72	0.61	89.29	0.62	0.19	0.35	0.81	0.46	0.65	76.46	88.15
Completeness	Propprietary	G-Eval	0.64	0.58	0.48	0.43	0.28	0.48	85.20	0.46	0.59	0.44	0.47	0.50	0.58	78.44	85.40
	LLMs	FineSurE	**0.80**	**0.72**	**0.57**	**0.76**	**0.68**	**0.72**	**89.42**	**0.78**	**0.74**	**0.64**	**0.84**	**0.60**	**0.74**	**88.62**	**97.13**
	Open-source	G-Eval	0.65	0.63	0.51	0.46	0.51	0.47	86.33	0.30	0.65	0.35	0.72	0.45	0.54	75.50	89.80
	LLMs	FineSurE	0.74	0.46	0.46	0.67	0.56	0.56	84.25	0.62	0.19	0.35	0.81	0.46	0.65	75.82	92.65

Table 31: Human alignment performance of automatic evaluators across three evaluation dimensions, six domains and two languages. Domain*: Domain Stability

Dimension	Evaluator	English						Chinese					
		News	Med Lit	Report	Booking	Meeting	Interview	News	Med Lit	Report	Booking	Meeting	Interview
Faithfulness	ROUGE-1	-0.16	0.35	0.38	-	-0.1	0.06	-0.07	0.12	-0.09	-	0.16	0.01
	ROUGE-2	-0.05	-0.08	-0.1	-	0.13	-0.17	0.07	0.1	0.02	-	0.03	-0.35
	ROUGE-L	-0.24	-0.18	0.05	-	0.02	-0.18	0.04	-0.06	0.18	-	0.15	-0.37
	BERTScore	-0.1	-0.16	0.02	-	0.11	-0.15	0	0.05	-0.06	-	0.03	-0.16
	FineSurE	0.74*	0.68*	0.52*	0.75*	0.69*	0.64*	0.42*	0.60*	0.62*	0.70*	0.48*	0.69*
Completeness	ROUGE-1	0.03	0	0	-	0.36	0.16	-0.23	0.2	0.33	-	0.59*	0.37
	ROUGE-2	-0.16	0.24	0.11	-	0.53*	0.24	-0.3	0.22	0.28	-	0.39	0.28
	ROUGE-L	-0.1	0.04	0.07	-	0.41*	0.07	-0.23	0.04	0.48*	-	0.41*	0.14
	BERTScore	0.05	0.13	0.06	-	0.54*	0.2	-0.11	0.16	0.38	-	0.39	0.29
	FineSurE	0.79*	0.75*	0.81*	0.90*	0.85*	0.80*	0.78*	0.66*	0.64*	0.90*	0.66*	0.75*
Conciseness	ROUGE-1	0.51*	0.54*	0.23*	-	0.53*	0.55*	0.42*	0.58*	0.4*	-	0.66*	0.69*
	ROUGE-2	0.41*	0.39*	-0.04	-	0.44*	0.61*	0.22*	0.45*	-0.02	-	0.54*	0.72*
	ROUGE-L	0.29*	0.41*	-0.01	-	0.43*	0.47*	0.13	0.4*	0.05	-	0.45*	0.52*
	BERTScore	0.33*	0.42*	0.02	-	0.43*	0.6*	0.13	0.48*	-0.11	-	0.54*	0.65*
	FineSurE	0.75*	0.68*	0.52*	0.75*	0.69*	0.64*	0.80*	0.70*	0.64*	0.82*	0.59*	0.74*

Table 32: Correlation between similarity-based metric and human evaluation score. *: p-value <0.05