A Tale of Evaluating Factual Consistency: Case Study on Long Document Summarization Evaluation

Yang Zhong
Department of Computer Science
University of Pittsburgh
yaz118@pitt.edu

Diane Litman
Department of Computer Science and LRDC
University of Pittsburgh
dlitman@pitt.edu

Abstract

Ensuring factual consistency in summarization remains a challenge, especially for long-document evaluation. While automated, reference-free evaluation models are essential given the impracticality of large-scale human assessment for lengthy texts, challenges persist in evaluating different systems on how to handle different summary granularities and evolving model generations. In this work, we conduct a systematic study on diverse factual-consistency evaluation systems across four long-document datasets, encompassing summaries generated by models from non-LLMs to proprietary LLMs. Our analysis reveals that fine-grained continuous scores can provide more reliable assessments of different evaluation systems' capabilities than binary classification. We also examine the relationship between sentence-level and summary-level model performance, highlighting its dependency on dataset characteristics. Moreover, our study reveals that advanced systems can achieve higher recall in error detection for older summaries, yet struggle with false positives and fine-grained error detection. Our analysis and case studies provide further insights into designing robust factuality evaluation systems, which are becoming increasingly in demand as generative models advance rapidly.

1 Introduction

Despite the recent progress for summarization models in producing fluent summaries, they still encounter challenges in producing summaries that are **factually consistent** with the source context (Maynez et al., 2020; Kryscinski et al., 2020; Goyal and Durrett, 2021; Cao and Wang, 2021; Zhang et al., 2024). While human evaluation remains the best practice to judge the factual consistency of automatically generated summaries, it becomes increasingly challenging when *long* sequences of generated texts (close to or over 100 words) and inputs (over thousands of words) need to be evaluated (Goyal et al., 2022; Krishna et al., 2023). Thus, building reference-free evaluation metrics becomes much more appealing. Yet, testing and developing reliable evaluation systems[1] faces challenges, especially for long document summarization.

Different from *short* news/dialogue summarization benchmarks–such as SUMMAC (Laban et al., 2022), AGGREFACT (Tang et al., 2023) and DIALSUMM (Gao and Wan, 2022) – which provide summary-level binary labels (1/0) indicating the factual consistency of brief (1-3 sentence) summaries, *long-document* evaluation benchmarks (Koh et al., 2022; Zhang et al., 2024; Lee et al., 2024) typically feature summaries of three or more sentences and are over 100 words in length. This discrepancy raises the question about how to produce the summary-level predictions for long document summaries. While some works (Zhang et al., 2024; Scirè et al., 2024) opt for a binary classification setup, others aggregate human-annotated sentence- or clause-level scores to produce a continuous measure of the overall factual consistency. While different evaluation systems are proposed to optimize for individual datasets, clarifying the impact of label setups is crucial to provide fair comparisons of different approaches.

Moreover, inspired by the study by Pagnoni et al. (2021), several long-document evaluation datasets collect sentence-level fine-grained error types. Although recent work (Xu et al., 2024; Song et al., 2024) started incorporating error prediction at the sentence level, their approaches relied on LLM's zero-shot capability, and there lacks a systematic evaluation of how different systems perform in de-

[1] In this paper, we distinguish between **evaluation systems** and **metrics** when reporting benchmarking results. The former refers to models that take the source document and summary as input to produce a summary-level prediction, while the latter refers to the measures (e.g., correlation, accuracy) used to compare these predictions against human-annotated scores in benchmark datasets.

Findings of the Association for Computational Linguistics: ACL 2025, pages 12511–12532

tecting the fine-grained error types.

Meanwhile, although summarization models have quickly evolved from pre-trained models such as BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020) to Large Language Models (LLMs) (Achiam et al., 2023; Dubey et al., 2024; Anthropic, 2024), the development of benchmarking datasets has lagged behind (i.e., the human-annotations of BART-generated summaries (Koh et al., 2022) still serve as one primary reference to assess the performance of varied strong LLM-based evaluators (Xu et al., 2024)). As found by Lee et al. (2024), proprietary LLM-generated summaries have a lower rate of intrinsic errors compared to those of non-LLM-generated summaries. Before attempting to excel on any given benchmark, it is vital to understand the strengths and limitations of different evaluation systems across model-generated summary types.

To overcome the limitations mentioned above, we conduct a systematic study over four different long-document summarization evaluation datasets that cover different source text domains, with summaries generated by models ranging from non-LLMs to proprietary LLMs. We empirically evaluate different factual consistency evaluation systems over the combined datasets, motivated by multiple research questions. Exploring the optimal utilization of summary-level annotation in evaluating long summaries (RQ1), we find that fine-grained, continuous scoring methods are more effective than binary classification (which is heavily relied on in short document summarization evaluations) (§5.1.1). We also examine the transferability of models' sentence-level prediction capabilities to summary-level evaluations (RQ2), noting a high correlation between model efficiency in detecting sentence-level inconsistencies and the performance on summary-level tasks, supposing utilization of suitable datasets and metrics (§5.1.2). Additionally, we assess the strengths and limitations of different evaluation systems in capturing fine-grained error types (RQ3), observing that more advanced systems can significantly improve the recall in detecting errors, yet still face challenges in minimizing false-positives (§5.2). Lastly, our analyses on how factual-inconsistency errors evolve as summarization systems are updated (§3.2 and §6) and the systematic comparison of system performances across time (RQ4, §5.3) underscores the need for continuous efforts in co-designing benchmarks and iteratively refining evaluation systems to enhance factual error detection. Lastly, we conduct both quantitative and qualitative analyses of model predictions, which reveal key takeaways about the strengths and limitations of different approaches, as well as actionable insights for guiding the development of future methods (§6).[2]

2 Related Work

Long Document Evaluation Benchmarks Research on automatic factual consistency evaluation metrics and resources for long document summarization is limited. Koh et al. (2022) released annotated model-generated summaries assessing factual consistency at the **sentence** and **summary** levels for GovReport (Huang et al., 2021) and arXiv (Cohan et al., 2018) and Krishna et al. (2023) introduced human-evaluation guidelines for long document summaries and released their crowdsourced datasets. Furthermore, Bishop et al. (2024), Zhang et al. (2024), Ramprasad et al. (2024), and Lee et al. (2024) introduced multiple benchmarks on long documents and covered diverse domains. However, most prior studies proposed and validated factual consistency evaluation methods using their own released corpora, which often lacked standardized reference labels (some used binary summary labels while others prefer continuous scores). *We aggregate multiple long document evaluation datasets that span multiple domains, unify the sentence-level and summary-level prediction tasks, and conduct a comprehensive study on the evaluation tasks.*

Evaluation Systems To tackle the challenges of long documents, numerous approaches leveraging either natural language inference (NLI) models (Zha et al., 2023; Zhang et al., 2024; Zhong and Litman, 2025) or Question-Answering-based models (Deutsch et al., 2021; Fabbri et al., 2022) were explored. With the rise of more capable LLMs, evaluating summaries on either sentence or summary level through zero-shot prompting became increasingly popular in the community (Liu et al., 2023; Song et al., 2024; Li et al., 2025). *Our study benchmarks different evaluation systems and systematically evaluates the strengths and limitations of those systems, providing discussions on the future directions in developing more robust factual consistency evaluation systems.*

Systematic Evaluation on Summarization Tasks Recent studies have aggregated benchmarking datasets and conducted meta-evaluations of differ-

[2]Scripts, predictions and analyses are available at https://github.com/cs329yangzhong/TaleOfFactualEval.

	Text Diversity	Summary Generation	Human Annotation		Dataset Statistics			
	Doc.Src	Gen.Model	Granularity	Sum.Label	Size	Doc.Word	Sum.Sent	Sum.Word
DIVERSUMM	GovReport ArXiv	*non-LLMs:* PEGASUS, BART	Sent Level	Binary	147 146	2008 4407	15 6	391 150
LONGEVAL	SQuALITY PubMed	*non-LLMs:* BART, BART-DPR BIGBIRD-Pegasus, LongT5	Clause Level	Percent.	40 40	7457 3888	19 8	388 190
RAMPRASAD'24	BillSum PubMed	*non-LLM:* Flan-T5-XL *proprietary LLM:* GPT-3.5	Sent Level	Percent.	100 100	1681 1797	3 4	86 100
UNISUMEVAL	GovReport PubMed SQuALITY MediaSum MeetingBank	*non-LLMs:* BART, T5 *open-source LLMs:* Phi-2, Mistral7B Llama2-13B-chat, Mixtral-8x7B *proprietary LLMs:* GPT3.5-turbo GPT-4-turbo, Claude2.1	Sent Level	Percent.	182 193 217 194 194	6296 3187 6083 1618 978	6 7 5 5 4	156 165 110 113 89

Table 1: Dataset statistics on our selected portions of DIVERSUMM (Zhang et al., 2024), LONGEVAL (Krishna et al., 2023), the dataset in Ramprasad et al. (2024), and the long split of UNISUMEVAL (Lee et al., 2024). We report the source of input (Doc.Src) and the summarization models (Gen.Model). Regarding human annotation details, LongEval annotates at the clause level by breaking down sentences into atomic units, while the other datasets annotate at the sentence level. Summary labels are either aggregated in a binary form (label 1 if there are no inconsistent sentences and 0 otherwise) or reported as a percentile of annotated "consistent" sentences/clauses. Finally, we report the number of test cases (Size), document length in the average number of words (Doc.Word), summary length in the average number of sentences (Sum.Sent) and words (Sum.Word).

ent factual consistency evaluation systems across these datasets (Gabriel et al., 2021; Pagnoni et al., 2021; Fabbri et al., 2021; Laban et al., 2022; Tang et al., 2023; Laban et al., 2023; Tam et al., 2023; Aharoni et al., 2023; Chrysostomou et al., 2024). However, many focus solely on summary-level performance metrics, such as correlation and accuracy, without offering insights into improving these evaluation systems. Notably, Pagnoni et al. (2021) performed error analyses of evaluation system predictions over fine-grained error types, and Tang et al. (2023) studied the different error types included in the benchmark's system summaries and further compared evaluation systems performances spanning both pre-LLM and LLM models. Yet, both studies focused on the *short* summarization tasks, with conclusions based on summary-level labels. *Our work addresses the more challenging long-document summarization tasks by considering the nuanced sentence-level evaluations. We introduce a comprehensive evaluation that spans different systems, including both the older specialized fact-checkers and more recent LLM-based systems, and analyze their performances over summaries produced by varied generations of generative models.*

3 Datasets and Analysis

3.1 Long Datasets Studied

We choose four data sources in our experiments: (1) DIVERSUMM (Zhang et al., 2024), specifically its *ArXiv* and *GovReport* split; (2) LONGEVAL (Krishna et al., 2023) (the machine-generated summary portion) [3]; (3) RAMPRASAD'24 (Ramprasad et al., 2024), which spans both the legal domain (Billsum) (Kornilova and Eidelman, 2019) and the medical domain (PubMed (Cohan et al., 2018)); as well as (4) UNISUMEVAL (Lee et al., 2024), which includes long document splits spanning five text domains, and we use its faithfulness annotations.

All datasets contain (annotated/derived) fine-grained sentence-level labels and aggregated summary-level labels. Noticeably, these datasets reflect the NLP community's ongoing efforts to construct annotated instances derived from summaries produced by state-of-the-art generative models, ranging from *non-LLMs* such as BART to *LLMs* such as GPT-4-turbo. We include detailed dataset comparisons in Table 1 and a more extended table about the dataset annotation quality in Table 8 of Appendix A.1. Compared to shorter summarization benchmarks such as AggreFact (Tang et al., 2023), which typically include inputs of fewer than 500 words and summaries consisting of just one to three sentences in the news domain, the corpora examined in our work cover a broader range of domains and feature much longer inputs and summaries. This complexity poses greater challenges for factual consistency detection (Koh et al., 2022).

3.2 Analyses on Dataset Labels

Analysis 1: How do factual inconsistency scores distribute across the binary summary-level la-

[3] This aligns with the setup in prior work (Wu et al., 2024) by excluding human-written reference summaries.

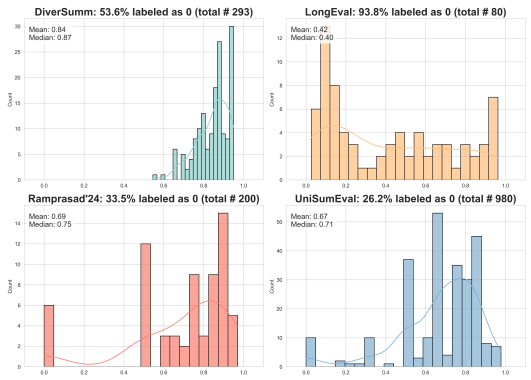

Figure 1: Distribution of summary-level continuous scores for summaries assigned the binary label of 0. For each dataset, we report the proportion of summaries containing errors (% labeled as 0) along with the total number of annotated article-summary pairs.

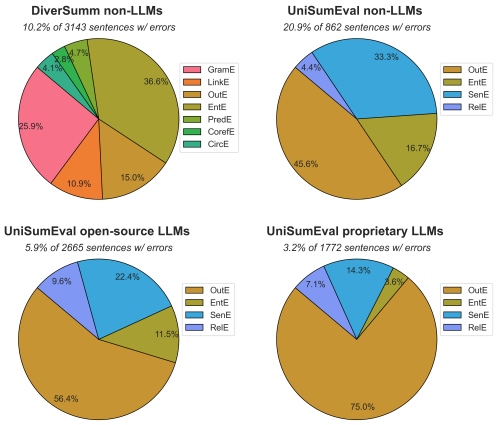

Figure 2: Error distribution across datasets and summarization models. Non-LLM summaries exhibit higher proportions of errors, of which entity (EntE) and out-of-article errors (OutE) are the most prevalent.

bels? In DiverSumm, summary-level labels are determined based on sentence-level annotations. A summary is labeled as *inconsistent* (0) if at least one sentence is annotated as inconsistent. Otherwise, it is labeled as consistent (1). In Figure 1, we plot the distributions of summary-level scores for all four datasets. For those annotated with default percentage scores, we transform them into the binary version as in DIVERSUMM. We observe that the distribution of datasets can differ, despite all containing factual consistency errors. While DIVERSUMM has a relatively higher proportion of consistent summary sentences, LONGEVAL has a left-skewed distribution, highlighting the high error rates. We hypothesize that treating all inconsistent summaries as 0 overlooks fine-grained differences, which may hinder the evaluation result's credibility when applying binary classification-based metrics. We further validate this hypothesis in RQ1 of §5.1.

Analysis 2: The distribution of fine-grained sentence-level factual inconsistency errors DIVERSUMM and UNISUMEVAL provide fine-grained fact verification error types (Table 9 and Table 10). In Figure 2, we plot the distributions of fine-grained error types among all errors for each dataset, with the breakdown over the summarization models (non-LLMs / open-source LLMs / proprietary LLMs). We observe that extrinsic errors, i.e., out-of-article errors (OutE), are prevalent in UNISUMEVAL, denoting the persisting challenges of models in introducing non-verifiable facts while generating summaries. The proportion of entity errors (EntE) is also reduced. In a later section (§5.2), we benchmark the capabilities of different evaluation systems in detecting various error categories.

4 Experimental Setup

With the varied research questions introduced, our main focus is to perform a systematic study on existing factual consistency evaluation systems under the long document evaluation setting. Below, we describe the systems we experiment with to produce sentence-level labels (§4.1), as well as the evaluation metrics suitable for sentence-level and summary-level evaluations (§4.2).

4.1 Factual-Consistency Evaluation System Selection

We split our evaluated systems into three main categories: (1) ***Specialized fact-checkers***, a term defined in prior work (Tang et al., 2024a), which are powered mainly by trained natural language inference models over fact-checking tasks and developed before LLMs. One of our baselines is **AlignScore** (Zha et al., 2023), an NLI-based metric used as one strong baseline in prior work. We also include **INFUSE** (Zhang et al., 2024), which sets the SOTA on DIVERSUMM, as well as **MINICHECK FT5-SENT** (MC-FT5) that is a best-performing non-LLM fact-checker over multiple benchmarks, which computes the individual summary sentences' scores using MINICHECK FT5 (Tang et al., 2024b). Additionally, we include (2) ***LLM-based systems*** that mainly rely on the LLM's zero-shot capability through prompting. We include **GPT4o** (OpenAI et al., 2024) and **Gemini** as the LLM fact-checkers, using a prompt adopted from Tang et al. (2024b). Additionally, we add **Llama-3.1-**

BeSpoke-MiniCheck-7B (BeSpoke), the SOTA fact-checking model on the LLM-AggreFact benchmark (Tang et al., 2024b), and the faithfulness component of **FineSurE** (Song et al., 2024), which evaluates summarization quality at a fine-grained level by assessing individual summary sentences. It introduces reasoning and error predictions on top of the binary classification.

Lastly, we include two variants of **StructScore** (Zhong and Litman, 2025), (3) ***the linguistic-informed approach*** to enhance long-document factual consistency evaluation. The initial sentence-level scores can be generated using different backbone systems. In this study, we consider **MC+rew.** and **BeSpoke+rew.**, where the sentence-level scores are re-weighted based on computed discourse-structure features. Additionally, we include **StructS-MC** and **StructS-BS** (BeSpoke), which incorporate both document segmentation and post-hoc reweighting for improved evaluation, as proposed in Zhong and Litman (2025). Implementation details are provided in Appendix B.

4.2 Evaluation Metrics

For *sentence-level* tasks, we evaluate in two settings. Since sentence-level annotations can be converted into a binary format (indicating whether the sentence contains errors or not), we report both ROC-AUC (AUC) and balanced-accuracy (bAcc) across all sentences within the summary, as done in prior work (Zhang et al., 2024; Song et al., 2024). On the fine-grained level evaluation, we measure the recall of individual fine-grained types when the system prediction does not have fine-grained types. On *summary-level* tasks with binary labels, we report balanced-accuracy bAcc and ROC-AUC (AUC). For tasks with continuous scores (i.e. Ramprasad et al. (2024) and UniSumEval), we report Pearson correlation r (Benesty et al., 2009) and Spearman correlation ρ (Zar, 2005).

5 Experiments

5.1 Which granularity-level metrics to use when evaluating different systems?

Despite the growing number of datasets, determining the most effective granularity-level evaluation remains challenging. As found in §3.2, the wide distribution of continuous scores suggests that a simple binary classification setting may be less convincing. In the following section, we empirically examine the impacts of summary-level evaluation settings, exploring the sentence-level prediction capabilities of different evaluation systems, as well as uncovering the connections between sentence-level and summary-level evaluations.

5.1.1 RQ1: What is the better way to utilize summary-level annotations in evaluating long summaries?

We standardize summary-level evaluations across the four datasets in both binary and continuous settings. For evaluating systems that produce continuous sentence-level scores, we aggregate them by computing the average to produce continuous summary-level scores. We choose the threshold 0.5 to generate binary labels, aligning with the zero-shot setup in Tang et al. (2024a). For LLM-based systems which only produce binary sentence-level labels, we report the percentage of predicted consistent sentences for LLM-based systems, as they only output binary sentence-level labels.

Table 2 shows the performances of different factual consistency evaluation systems on the full dataset.[4] We observe different trends between the binary and continuous evaluation settings. Among evaluation systems, the linguistic-informed approaches generally rank in top-3 on DIVERSUMM (row 8-11) and on RAMPRASAD'24 (i.e. row 8 obtains the highest AUC and ρ for the two settings). Specialized fact-checkers, such as ALIGNSCORE and MC-FT5 (SENT), obtain better AUC over RAMPRASAD'24 and LONGEVAL, yet perform the worst on r. We attribute this to those systems' tendency to assign low scores to summaries with factual consistency near zero, effectively handling skewness but failing to differentiate varying levels of inconsistency. On UNISUMEVAL, LLMs achieve significantly higher correlations on continuous labels (rows 4-6) and AUC.

To further validate our hypothesis on the effects of dataset fine-grained score distributions (§3.2 Analysis 1), we compare the ranking correlations among the 11 systems under the two evaluation setups (Binary vs. Continuous) in the bottom block of the table. AUC and Pearson's r exhibit significant correlations, as measured by both Spearman's ρ and Kendall's τ, particularly on DIVERSUMM and UNISUMEVAL, which follow normal distributions. In contrast, on LONGEVAL, rank correlations between AUC, Pearson's r, and bAcc are generally close to zero or negative, indicating that binary

[4] Alternative colorings by quantile ranking as well as domain-specific breakdowns are in Appendix C.

ID	Eval System	DIVERSUMM				RAMPRASAD'24				LONGEVAL				UNISUMEVAL			
	Ref. Label	Binary		Contin.		Binary		Contin.		Binary		Contin.		Binary		Contin.	
	Eval Metric	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ
Specialized fact-checkers																	
1	INFUSE	87.15	59.31	0.55*	0.60*	74.16	60.86	0.36*	0.41*	**96.53**	**84.67**	0.72*	0.69*	52.91	50.86	0.05	0.04
2	AlignScore	84.59	**74.00**	0.53*	0.57*	75.55	54.46	0.42*	0.45*	91.20	76.00	0.74*	0.74*	57.41	53.74	0.21*	0.12*
3	MC-FT5	88.05	55.27	0.57*	0.62*	75.34	54.27	0.48*	0.45*	86.93	79.33	0.80*	0.77*	57.31	49.96	0.10*	0.12*
LLM-based																	
4	GPT4o	82.45	53.19	0.59*	0.59*	60.24	53.97	0.44*	0.33*	84.40	71.33	0.88*	0.87*	69.67	52.65	0.51*	0.41*
5	Gemini	83.11	55.10	0.52*	0.55*	56.58	54.35	**0.54***	0.33*	89.20	80.67	0.90*	0.86*	**71.18**	52.16	**0.54***	**0.47***
6	BeSpoke	61.92	53.21	0.35*	0.30*	66.91	53.58	0.49*	0.33*	90.13	68.67	**0.94***	**0.92***	69.61	53.29	0.41*	0.32*
7	FineSurE	81.36	68.37	0.49*	0.53*	62.63	51.86	0.32*	0.28*	84.93	**84.67**	0.86*	0.85*	69.24	**57.84**	0.41*	0.34*
Lingustic-informed																	
8	MC+rew.	88.62	61.27	0.59*	0.64*	**76.94**	54.23	0.49*	**0.47***	88.53	81.33	0.81*	0.76*	56.05	50.44	0.08	0.10
9	StructS-MC	**91.19**	62.54	**0.63***	**0.68***	71.68	58.08	0.43*	0.38*	89.60	80.67	0.90*	0.86*	54.91	51.30	0.08*	0.08*
10	BeSpoke+rew.	60.78	55.45	0.34*	0.29*	71.40	53.85	0.51*	0.40*	90.13	70.67	0.93*	0.90*	64.81	55.94	0.32*	0.25*
11	StructS-BS	89.65	69.72	0.62*	0.66*	72.03	**61.63**	0.47*	0.40*	86.40	74.67	0.86*	0.85*	58.99	55.67	0.17*	0.15*
System-level Eval Metrics Ranking Correlation																	
AUC vs. bACC		ρ = 0.45, τ = 0.35				ρ = 0.41, τ = 0.24				ρ = -0.02, τ = 0.00				ρ = 0.50, τ = 0.27			
AUC vs. Pearson		ρ = 0.89*, τ = 0.84*				ρ = -0.18, τ = -0.11				ρ = -0.09, τ = -0.04				ρ = 0.99*, τ = 0.94*			
BACC vs. Pearson		ρ = 0.25 , τ =0.18				ρ = -0.19, τ = -0.15				ρ = -0.57, τ = -0.45				ρ = 0.52, τ = 0.28			

Table 2: Summary-level results for all summarization evaluation systems on DIVERSUMM, RAMPRASAD'24, LONGEVAL and UNISUMEVAL. We study two types of summary-level reference labels (Ref Label): (1) Binary, evaluated using ROC-AUC (AUC) and balanced accuracy (bAcc), and (2) Continuous, for which we report Pearson correlation r and Spearman correlation ρ, with * indicating statistical significance ($p < 0.05$). The best model performance per column is **bold**. Green, orange and red indicate system performance ranking intervals per column based on three-digit value ranks, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively. We further examine system-level rank correlations across different metric pairs, finding relatively low correlations, which highlights the challenges in comprehensively evaluating a model's capabilities.

evaluation metrics struggle to provide fair assessments when label distributions are relatively flat. **In summary, given the nature of long summaries, fine-grained labels provide a more comprehensive evaluation than binary conversion, as long as the dataset is not overly skewed (i.e., most sentences are not factually inconsistent).**

5.1.2 RQ2. Is sentence-level prediction performance transferable to summary-level results?

Sent-level Evaluation. We present sentence-level performance results and include an additional column for default summary-level evaluation, following prior work (Zhang et al., 2024; Ramprasad et al., 2024; Wu et al., 2024) in Table 3. Looking at sentence-level prediction columns, we find that evaluation systems perform differently across datasets. The linguistic-informed approach StructS-BS obtains better AUC than zero-shot LLMs on DiverSumm. The BeSpoke that is fine-tuned on a large amount of inference tasks also demonstrates its capability in obtaining the best or second best AUC on LONGEVAL and UNIEVALSUMM. To analyze the correlation between sentence-level and summary-level performance, we conduct a ranking comparison across the 11 evaluation systems in the bottom section of Table 3. Over DIVERSUMM and RAMPRASAD'24, where summaries contain fewer inconsistent sentences, we observe a discrepancy between sentence-level and summary-level performances. **On the other datasets that are sufficiently large and contain summaries with varying degrees of factual inconsistency, there exist significant correlations (measured in Spearman's ρ and Kendall's τ) between the two sets of evaluations, suggesting that optimizing sentence-level detections could benefit the summary-level development.**

Instance-level Analysis. We introduce a new approach to analyze the relation between sentence-level accuracy and summary-level evaluation. We adopt the reliability diagram (Murphy and Winkler, 1977) (which is traditionally used to compare predicted probabilities to actual outcomes in classification tasks (Guo et al., 2017))—by treating each summary's average sentence-level accuracy as the "predictor" and comparing it against the model's summary-level factuality predictions and ground-truth labels (here we use the continuous ground truth labels across all four datasets).[5] We first bin the summaries by their sentence-level accuracy range and compute each bin's corresponding average predicted vs. actual summary-level scores.[6] Plotting the error between predicted and ground-

[5]We provide an illustrative example in Appendix D.

[6]By default, we form 10 bins within the range of 0-1.

ID	Eval System	DIVERSUMM			RAMPRASAD'24			LONGEVAL			UNISUMEVAL		
	Eval. Task	Sent-level		Summ-level	Sent-level		Summ-level	Sent-level		Summ-level	Sent-level		Summ-level
	Eval. Metric	*AUC*	*bAcc*	*AUC*	*AUC*	*bAcc*	ρ	*AUC*	*bAcc*	r	*AUC*	*bAcc*	r
Specialized fact-checkers													
1	INFUSE	75.62	67.96	87.15	67.30	63.02	0.41*	77.84	72.89	0.72*	62.31	61.52	0.05
2	AlignScore	72.70	68.73	84.59	**74.78**	**66.06**	0.45*	76.76	72.24	0.74*	66.88	61.90	0.21*
3	MC-FT5	74.45	68.40	88.05	68.33	58.82	0.45*	86.47	77.88	0.80*	68.94	61.19	0.10*
LLM-based													
4	GPT4o	74.88	74.88	82.45	61.54	61.54	0.33*	88.72	**88.72**	0.88*	75.89	75.89	0.51*
5	Gemini	70.47	70.47	83.11	60.33	60.33	0.33*	88.49	88.49	0.90*	70.50	70.50	**0.54***
6	BeSpoke	76.53	70.32	61.92	74.35	60.13	0.33*	**95.11**	87.94	**0.94***	**82.39**	71.64	0.41*
7	FineSurE	76.69	**76.69**	81.36	64.93	64.93	0.28*	84.64	84.64	0.86*	77.23	**77.23**	0.41*
Lingustic-informed													
8	MC+rew.	75.89	69.16	88.62	73.41	60.39	**0.47***	85.97	79.53	0.81*	67.73	61.14	0.08*
9	StructS-MC	75.74	69.75	**91.19**	68.66	57.95	0.38*	90.37	83.80	0.90*	65.57	60.31	0.08*
10	BeSpoke+rew.	78.58	71.34	60.78	74.46	63.46	0.40*	94.41	87.80	0.93*	79.81	72.69	0.32*
11	StructS-BS	**81.58**	75.89	89.65	71.83	**66.06**	0.40*	91.88	86.19	0.86*	73.36	68.36	0.17*
System-level Eval Metrics Ranking Correlation													
sent-AUC vs. sent-bAcc		ρ = 0.54, τ = 0.38			ρ = 0.25, τ = 0.18			ρ = 0.70*, τ =0.56*			ρ = 0.81*, τ = 0.60*		
sent-AUC vs. summ-metric		ρ = -0.11, τ = -0.05			ρ = 0.51, τ = 0.38			ρ = 0.86*, τ = 0.69*			ρ = 0.72*, τ = 0.54*		
sent-bAcc vs. summ-metric		ρ = -0.36, τ = -0.24			ρ = 0.06, τ = 0.02			ρ = 0.80*, τ = 0.65*			ρ = 0.81*, τ = 0.61*		

Table 3: Sentence-level performance of different evaluation systems on DIVERSUMM, RAMPRASAD'24, LONGEVAL and UNISUMEVAL. We report both AUC and bAcc and include the official summary-level metric released for each dataset. Green, orange and red indicate system performance ranking intervals per column based on three-digit value ranks, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively. The best model performance per column is **bold**. We report the rank correlations across sentence-level and summary-level ratings. * indicates statistical significance (p<0.05).

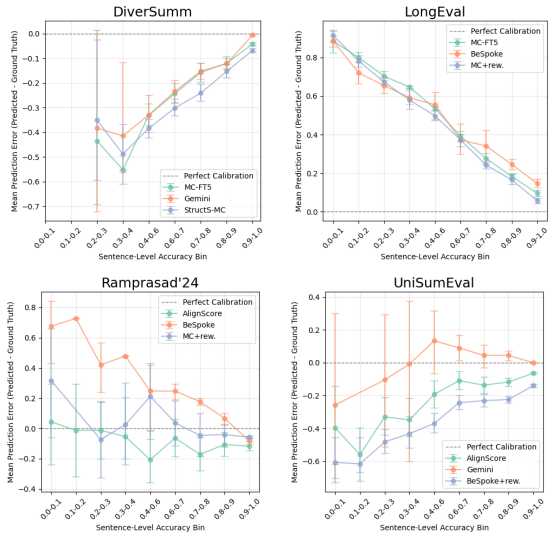

Figure 3: The reliability diagrams of the top-performing summary-level evaluation systems for each dataset. On the y-axis, values above and below zero indicate overestimation (predicts too high) and underestimation of the true score, respectively. Points close to 0 suggest that the model is well-aligned with the ground truth in the accuracy range. The x-axis refers to the bins of summaries with sentence-level accuracy from less (left) to more (right) accurate.

truth scores along these bins allows us to visualize whether the model consistently overestimates or underestimates summary-level factual consistency when sentence-level accuracy is high or low.

We select the best-performed summary-level models within the three model categories. Figure 3 reports the diagrams of all four datasets. Looking at the y-axis, we observe that all models underestimate the scores of DIVERSUMM (predicting lower scores than ground truth labels) and overestimate over LONGEVAL, which aligns with the factual consistency score distribution in §3.2. The upward trends from left to right in DIVERSUMM and UNISUMEVAL suggest that the model becomes more calibrated when sentence-level accuracy is increasing, becoming more accurate with higher-accuracy summaries. Models turn more precise (closer to 0) by reducing the positive errors in the remaining two datasets. Comparing different models, we observe Gemini obtains the best-calibrated performances on UNISUMEVAL, affirmed by its high summary-level correlation score in Table 3.

5.2 Fine-grained Error Detection

RQ3. What are the strengths and limitations of different evaluation systems in detecting fine-grained error types? We investigate the capability of different models in capturing the fine-grained types on DIVERSUMM and UNISUMEVAL. We select the models with the highest sentence-level bAcc per dataset (Table 3 of each model category). Compared to the specialized fact-checker AlignScore, LLM-based model FineSurE improves the recall of EntE and GramE in Table 4 and the recalls of all errors in Table 5. However, current

Error Cat. (error/total: 320/3143)	OutE (48)	EntE (117)	PredE (15)	CircE (13)	GramE (83)	LinkE (35)	CorefE (9)
AlignScore (1388)	97.9%	74.4%	73.3%	92.3%	66.3%	91.4%	55.6%
FineSurE (1312)	93.8%	89.7%*	93.3%	84.6%	92.8%*	74.3%	100%
StructS-BS (1082)	91.7%	79.5%	93.3%	100.0%	71.1%	80.0%	88.9%

Table 4: Recall analysis in assessing factual consistency on DIVERSUMM. We report the sentence count for each erroneous type. We also include the numbers of predicted error sentences for each model, which indicate high false positives. (*) indicates a significant difference in recall compared to AlignScore.

Error Cat. (error/total : 392/5299)	OutE (212)	SenE (103)	EntE (50)	RelE (27)
AlignScore (904)	42.5%	29.1%	46.0%	40.7%
FineSurE (1129)	64.6%*	85.4%*	84.0%*	66.7%*
BeSpoke+rew. (1179)	59.9%*	77.7%*	64.0%*	51.9%

Table 5: Recall analysis in assessing factual consistency on UNISUMEVAL. We report the sentence count for each erroneous type. We also include the number of predicted error sentences for each model, which indicates high false positives. (*) indicates a significant difference in recall compared to AlignScore.

ID	Eval System	UNISUMEVAL					
	Gen.Model	non-LLMs		open LLMs		close-LLMs	
	Eval Metric	r	ρ	r	ρ	r	ρ
Specialized fact-checkers							
1	INFUSE	0.27*	0.29*	0.19*	0.19*	0.06	0.04
2	AlignScore	0.40*	0.32*	0.30*	0.27*	-0.04	-0.06
3	MC-FT5	0.24*	0.30*	0.28*	0.28*	-0.02	-0.02
LLM-based							
4	GPT4o	0.54*	0.53*	0.43*	0.37*	0.02	0.02
5	Gemini	**0.59***	**0.60***	0.41*	**0.42***	**0.14***	**0.10***
6	BeSpoke	0.46*	0.43*	**0.47***	0.40*	0.01	0.05
7	FineSurE	0.47*	0.44*	0.42*	0.34*	0.07	0.10
Lingustic-informed							
8	MC+rew.	0.24*	0.28*	0.27*	0.28*	-0.02	-0.01
9	StructS-MC	0.28*	0.31*	0.25*	0.26*	-0.00	-0.02
10	BeSpoke+rew.	0.44*	0.41*	0.41*	0.34*	-0.00	0.02
11	StructS-BS	0.41*	0.40*	0.30*	0.29*	0.01	-0.00

Table 6: Summary level results on UNISUMEVAL separated by the types of summarizer model. We report Pearson correlation r and Spearman correlation ρ under **summary-level**, with * indicating statistical significance. The best model performance per column is **bold**. Green, orange and red indicate the performance ranking intervals of the system for each column, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively.

evaluation systems tend to generate many false positives in sentence-level predictions. For instance, in UNISUMEVAL (long split), each of the three models in Table 5 predicts over 900 error sentences, which is far more than the actual count of 392 inconsistent sentences. Notably, over 75% of model-predicted "non-factual" sentences are annotated as consistent. This calls for further endeavors in developing evaluation methods that produce fewer false positives in error prediction. Meanwhile, accurately predicting fine-grained error types remains challenging, as discussed in Song et al. (2024).

5.3 RQ4. How do factuality-consistency evaluation systems perform across datasets of different generations?

We experiment with UNISUMEVAL by splitting the dataset based on the summarization model used (splits shown in Table 1). As seen in Table 6, LLM-based models consistently obtain the highest performance across all three groups of summaries. Meanwhile, specialized fact-checkers (row 1 and row 2) show a decline in performance as summarization models evolve (i.e., in row 1, InFUSE's performance, measured by Pearson correlation, declines from 0.27 on non-LLMs summaries to 0.06 with close-sourced proprietary LLMs' output.) Lastly, most models–except for Gemini–fail to generate evaluations that are correlated with the human-annotated errors when using proprietary LLMs as the summarizer. This finding aligns with the preliminary exploration of faithfulness evaluation systems' domain-level performances in Lee et al. (2024), calling for efforts in developing detection models that can identify the more nuanced errors generated by powerful LLMs.

6 Analysis

We conducted both quantitative and qualitative analyses of model predictions across different systems, leading to the following insights and takeaways. **Takeaway 1**: The main types of factual consistency errors have shifted as summarizer models continue to improve, from surface-level and grammatical errors to those requiring a deeper understanding of context and inferential reasoning. (Findings 1.1 and 1.2 in §E.1.) **Takeaway 2**: The primary performance gains of recent LLMs lie in their ability to reduce false positive predictions on factual sentences. However, no significant improvements are observed regarding specific fine-grained error types. (Finding 1.3, 1.4 in §E.1 and examples below). In addition, two actionable insights are drawn. **Insight 1 — Evaluation Focus:** On evaluation, future work should move beyond overall summary-level correlations and focus on pinpointing specific errors. As LLM-generated text is mostly factual, the inconsistencies only take a small portion. Aggregated summary-level evaluations may thus overlook the true capabilities of

Ex 1: They were all men who paid for their ship with their lives. **Label:** SentE; **Generator:** Phi-2; All systems failed to identify the error. **Author comment:** The crew did not pay for their ship with their lives; rather, they paid for their greed when encountering the cursed derelict ship. This factual error is more nuanced and pinpoints the challenge of deriving the correct relations when using specific entities.
Ex 2: The medicos are aware of the dangers of contagious diseases from the beginning rather than being explicitly warned later. **Label:** EntE; **Generator:** Phi-2; All systems failed to identify the error. **Author comment:** The medicos are not aware of the dangers of contagious diseases from the beginning. The error is embedded in the conceptual understanding of the context, not simply on the surface level.
Ex 3: Map is a chronic debilitating disease in ruminants. **Label:** SenE; **Generator:** T5 model, GPT4o and Gemini detected the error, while all others failed. **Author comment:** In the original document, "the general characteristics of Johne's disease with respect to the pathogenesis and immune response to MAP, as well as recent advances in development of vaccines were briefly examined" suggests that MAP itself is not a disease, but the causative agent of Johne's disease.
Ex 4: The story explores themes of isolation, adaptation, and the risks and ethical dilemmas of colonization and medical experimentation. **Label:** OutE; **Generator:** GPT-4; Only linguistic-based models detect the error, and all other models fail to identify it. **Author comment:** The sentence introduces an interpretative element that is not explicitly stated in the transcript. This suggests that interpretative elements are becoming harder to detect, even for LLMs that generate the summary itself.
Ex 5: Starrett Blade, a space pirate, is trapped by the feared Devil Garrett and fights for his life. **Label:** EntE; **Generator:** GPT3.5 model; only GPT4o and StructS (BS) succeeded in detecting the errors. **Author comment:** The source text begins with, "Trapped by the most feared space pirates, Devil Garrett, Starrett Blade was fighting for his life." Further main document texts show that Starrett is, in fact, a hunter of space pirates. Here, the model messed up the entity attribution.

Table 7: Examples of system detection error analysis, including the oracle label, the generator of the corresponding summary, model prediction status, and accompanying author comments.

LLM-based systems in error detection. (Findings in §E.2.) **Insight 2 — Model Improvements**: Future factual consistency evaluation systems should leverage the reasoning capabilities of LLMs to verify the credibility of individual sentences. **Our case studies in Table 7 review that LLM-generated summaries can include errors that appear plausible on the surface yet lack factual grounding or omit critical context upon closer inspection**. These subtle errors often require deeper reasoning and contextual understanding to detect—an area where current automatic evaluation systems still fall short. Incorporating approaches such as multi-step reasoning will be a potential future direction. Further details can be found in Appendix E.

7 Discussion and Conclusion

For long document summary evaluation, the connection between summary-level and sentence-level evaluation has largely been overlooked. Most prior work directly adopted approaches from the short-text domain or proposed a new benchmark and demonstrated the superior performance of their proposed approach. Instead, we examine more fundamental challenges of selecting suitable evaluation granularity and metrics. Without such reflection, researchers risk optimizing their methods toward objectives that are misaligned. For instance, many novel systems tested on AggreFact (Tang et al., 2023) show minimal differences in scores. However, Laban et al. (2023) point out that this dataset contains at least 6% mislabeled content (flipped binary labels), casting doubt on the reliability and usefulness of the improved benchmarking results for informing future studies. Our work serves as a foundational study to clarify existing challenges and provide guidance for future research.

To sum up, we conducted a comprehensive evaluation of factual-consistency detection systems across multiple long-document datasets, analyzing their performance at different granularity levels. Our findings highlight the benefits of fine-grained continuous scoring over binary classification and reveal the transferability of sentence-level predictions to summary-level evaluation. Our analysis reveals that while advanced LLM-based systems exhibit higher recall in detecting errors compared to older benchmarks, they struggle with false positives and the fine-grained detection of errors. Our work further provides thorough analyses and useful insights into the current state of evaluation systems, calling for continuous benchmark development and system refinement to enhance factual consistency assessment in long document summarization.

Acknowledgment

This work is supported by the National Science Foundation under Grant No. 2040490, by Amazon, and in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681. We want to thank the Pitt PETAL NLP, and AI Fairness and Law groups, and reviewers, for their valuable comments in improving this work.

Limitation

Our findings and conclusions relied on human annotation efforts for the different datasets and focused on the factual consistency dimension of summary quality. Additional dimensions, such as completeness and salience, could enhance the assessment of summaries. Our selection of evaluation systems is decided based on recent literature in the long document summarization and requires a sentence-level prediction. We acknowledge that there exist recent approaches that focus on decomposing the summary into atomic facts (Scirè et al., 2024; Yang et al., 2024a) and function on making summary-level predictions. We did not include these approaches in the main result section because they break the summary into standalone claims. This is incompatible with our primary research question, which aims to uncover the connections between summary-level and sentence-level evaluations. We include the summary-level results by rerunning models using their released codebase, as presented in Appendix F, finding that these claim-decomposition-based approaches perform on par or worse than the sentence-level baselines such as MC-FT5. Another line of question answering (QA) based metrics exists for faithfulness evaluation (Deutsch et al., 2021; Fabbri et al., 2022). However, these methods operate by verifying loosely connected sentence-level annotations, rely on some older backbone models that are unable to process long contexts, and focus on noun-phrase mismatches generated by question-answering models developed in 2021. We leave more exploration of these approaches for future work.

Ethical Statement

Throughout the paper, we have referenced datasets and models used in our analyses and experiments, ensuring that they are openly available and do not pose concerns with the public release or usage of this paper. We acknowledge the use of Grammarly and ChatGPT for correcting sentences that are less fluent, but not for generating or drafting new content.

References

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

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Jennifer A. Bishop, Sophia Ananiadou, and Qianqian Xie. 2024. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10777–10789, Torino, Italia. ELRA and ICCL.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

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

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Allan H. Murphy and Robert L. Winkler. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):41–47.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang,

Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Sanjana Ramprasad, Kundan Krishna, Zachary Lipton, and Byron Wallace. 2024. Evaluating the factuality of zero-shot summarizers across varied domains. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 50–59, St. Julian's, Malta. Association for Computational Linguistics.

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

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

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

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

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

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Liyan Xu, Zhenlin Su, Mo Yu, Jin Xu, Jinho D. Choi, Jie Zhou, and Fei Liu. 2024. Identifying factual inconsistencies in summaries: Grounding LLM inference via task taxonomy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14626–14641, Miami, Florida, USA. Association for Computational Linguistics.

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

Joonho Yang, Seunghyun Yoon, Byeongjeong Kim, and Hwanhee Lee. 2024b. Fizz: Factual inconsistency detection by zoom-in summary and zoom-out document. *arXiv preprint arXiv:2404.11184*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

```
You will receive a transcript followed by a corresponding summary.
Your task is to assess the factuality of each summary sentence
across nine categories:
* no error: the statement aligns explicitly with the content of
the transcript and is factually consistent with it.
* out-of-context error: the statement contains information not
present in the transcript.
* entity error: the primary arguments (or their attributes) of
the predicate are wrong.
* predicate error: the predicate in the summary statement is
inconsistent with the transcript.
* circumstantial error: the additional information (like location
or time) specifying the circumstance around a predicate is wrong.
* grammatical error: the grammar of the sentence is so wrong that
it becomes meaningless.
* coreference error: a pronoun or reference with wrong or non-
existing antecedent.
* linking error: error in how multiple statements are linked
together in the discourse (for example temporal ordering or
causal link).
* other error: the statement contains any factuality error which
is not defined here.

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

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

Transcript:
{input text}

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

Figure 4: Zero-shot factual consistency evaluation prompt used in FineSurE (Song et al., 2024).

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *NAACL*. Association for Computational Linguistics.

A Dataset Details

A.1 Dataset Annotation Details

Table 8 provides details about the human annotations of datasets studied in the paper.

A.2 Facual Inconsistency Error Type

Adopting the typology of factual errors introduced (Pagnoni et al., 2021), Koh et al. (2022) annotated summaries on sentence level by classifying it into one of the seven categories in Table 9 or no error (NoE). UNISUMEVAL adopt a different setup of fine-grained labels with details in Table 10.

B Implementation Details

B.1 LLM Prompts

We include our prompt for zero-shot factual consistency evaluation using GPT4o and Gemini in Table 11 and the prompts adopted from FineSurE (Song et al., 2024) for fact-verification on DIVERSUMM and UNISUMEVAL ("Binary Label + Reasoning + Error Localization") in Figure 4.

B.2 Baselines

AlignScore (model size 355M) (Zha et al., 2023) is an entailment-based model that has been trained on data from a wide range of tasks such as NLI, QA, and fact verification tasks. It divides the source document into a set of sequential chunks at sentence boundaries. For a multi-sentence summary, it predicts the max scoring value of all combinations of source chunk and target sentence, then returns the unweighted average of all sentences as the summary prediction. We follow the original setting by setting chunk size at 350 tokens and use the default model alingsocre_large ckpt. The model outputs a score between 0 and 1. We conduct experiments on top of their released codebase https://github.com/yuh-zha/AlignScore.

MiniCheck-FT5 (model size 770M) (Tang et al., 2024b) is an entailment-based fact checker built on flan-t5-large. It has been further fine-tuned on 21K datapoints from the ANLI dataset (Nie et al., 2020) and 35k synthesized data points generated in (Tang et al., 2024b) on the tasks to predict whether a given claim is supported by a document. We follow the authors's setting and set the chunk size to 500 tokens using white space splitting. The output score is between 0 and 1. We use the released code repo from https://github.com/Liyan06/MiniCheck.

InfUsE (model size 60M) Zhang et al. (2024) uses a variable premise size and breaks the summary into sentences or shorter hypotheses. Instead of fixing the source context, it retrieves the best possible context to assess the faithfulness of an individual summary sentence by applying an NLI model to successive expansions of the document sentences. Similar to prior approaches, it outputs an entailment score for each summary sentence, and the summary-level score is the unweighted average. We follow their settings on INFUSE with summary sentences instead of $\text{INFUSE}_{\text{SUB}}$ as the authors only released the code for the former model. INFUSE outputs scores in the range 0-1. We use the author's released codebase from https://github.com/HJZnlp/Infuse.

GPT4o We used the API of chatgpt-4o-latest (as of Feb 1, 2025); we set max_tokens 2000, sampling temperature at 0, and top_p as 1.0. We call the OpenAI API from https://openai.com/api. Given the lengthy summary, we prompted the LLM to assign a binary label (yes/no) to assess individual

Dataset	Source	Annotators	Kappa	Annotation Scheme
DiverSumm	ArXiv GovReport	3 annotators	Fleiss κ 0.52	{NoE, EntE, PredE, CircE, CorefE, LinkE, GramE}
LongEval	SQuALITY PubMed	crowd-source	Fleiss κ 0.74 Fleiss κ 0.53	clause-level binary
Ramprasad'24	BillSum PubMed	2 expert attorneys 2 expert medical doctors	Cohen's κ 0.17 Cohan's κ 0.11	binary / {intrinsic, extrinsic, mixed, others}
UniSumEval	9 domains	crowed-source	Krippendorff's α 0.60	{OutE, EntE, RelE, SentE}

Table 8: Summary of the human-labeled datasets we used in Table 1.

	Category	Description	Example
PredE	Relation Error	The predicate in the summary statement is inconsistent with the source article.	*The Ebola vaccine **was rejected** by the FDA in 2019.*
EntE	Entity Error	The primary arguments (or their attributes) of the predicate are wrong.	*The **COVID-19 vaccine** was approved by the FDA in 2019.*
CircE	Circumstance Error	Additional information (such as location or time) specifying the circumstance around a predicate is wrong.	*The first vaccine for Ebola was approved by the FDA in **2014**.*
CorefE	Coreference Error	A pronoun/reference with a wrong or non-existent antecedent.	*The first vaccine for Ebola was approved in 2019. **They** say a vaccine for COVID-19 is unlikely to be ready this year.*
LinkE	Discourse Link Error	Error in how multiple statements are linked together in discourse (e.g., incorrect temporal ordering or causal link).	*To produce the vaccine, scientists have to show successful human trials, **then** sequence the DNA of the virus.*
OutE	Out of Article Error	The statement contains information not present in the source article.	***China** has already started clinical trials of the COVID-19 vaccine.*
GramE	Grammatical Error	The grammar of the sentence is so incorrect that it becomes meaningless.	*The Ebola vaccine **accepted have already started**.*

Table 9: Typology of factual errors in DiverSumm. The example is directly taken from (Pagnoni et al., 2021). Original text for the examples: *The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.*

summary sentences' consistency with the original article. Then, we reported the percentile of "yes" answers as the summary-level rating.

Gemini We used the API of Gemini-2.0-Flash; we set max_tokens 2000, sampling temperature at 0, and top_p as 1.0. We call the API from `https://ai.google.dev/gemini-api/`. Given the lengthy summary, we prompted the LLM to assign a binary label (yes/no) to assess individual summary sentences' consistency with the original article. Then, we reported the percentile of "yes" answers as the summary-level rating.

BeSpoke-MC-7B We harnessed the SOTA Llama-3.1-BeSpoke-MiniCheck-7B (BeSpoke-MC-7B) released by BeSpoke Labs. The model is fine-tuned from "internlm/internlm2_5-7b-chat" (Cai et al., 2024) on the combination of 35K data points following the approach in MiniCheck(Tang et al., 2024b). We use the suggested code repo from `https://huggingface.co/bespokelabs/BeSpoke-MiniCheck-7B`. To calculate the AUC score, we employed the raw probabilities returned by the code to determine sentence-level ratings. We calculated the summary-level score as the unweighted average across all sentences.

FineSurE We utilize the original factual error types used in the prompt of FineSureE (Song et al., 2024) and the modified version used in UniSumEval (Lee et al., 2024). We use the author's released codebase from `https://github.`

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

Table 10: Descriptions and examples of factual error types in UNISUMEVAL. The parts of each summary sentence that are relevant to the specific error type are highlighted in **bold**.

Determine whether the provided claims are consistent with the corresponding document. Consistency in this context implies that all information presented in the claim is substantiated by the document. If not, it should be considered inconsistent.

Document: [DOCUMENT]
Claims: [CLAIMS]
Please assess the claim's consistency with the document by responding with either "yes" or "no".
The CLAIMs are ordered in the format of a dictionary, with { index: CLAIM }. You will need to return the result in JSON format.
For instance, for a CLAIMs list of 4 items, you should return {0:yes/no, 1:yes/no, ..., 3:yes/no}.

ANSWER:

Table 11: Zero-shot factual consistency evaluation prompt for GPT4o and Gemini.

com/DISL-Lab/FineSurE-ACL24.

StructS We use the authors' released codebase[7] to make predictions. Being compatible with different sentence-level prediction models, the STRUCTS (Zhong and Litman, 2025) is composed of two parts: (1) discourse-inspired document chunking and (2) a reweighting algorithm aggregating discourse structure information to re-calibrate the sentence-level scores predicted. For more details, we refer the authors to the original paper.

B.3 Machine Configuration for Models

We use up to 2 NVIDIA L40S GPU, each equipped with 48 GB VRAM, for model inferences on our hardware. According to Lambda[8] a single NVIDIA A6000 (the closest to our setting with 48GB VRAM) GPU costs $0.8 per hour.

[7] https://github.com/cs329yangzhong/discourse-driven-summary-factuality-evaluation

[8] https://lambdalabs.com/service/gpu-cloud

C Extra Results

C.1 Different Coloring

Here we provide a different coloring method for Tables 2 and 3 in Table 12 and Table 13. For each column, we rank the scores into quantile intervals.

C.2 Summary-level Evaluation with Domain Split

Table 14 and Table 15 show the domain-split of summary-level evaluations on the experimented four datasets.

ID	Eval System	DIVERSUMM				RAMPRASAD'24				LONGEVAL				UNISUMEVAL			
	Ref. Label	Binary		Contin.		Binary		Contin.		Binary		Contin.		Binary		Contin.	
	Eval Metric	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ
Specialized fact-checkers																	
1	INFUSE	87.15	59.31	0.55	0.60	74.16	60.86	0.36	0.41	96.53	84.67	0.72	0.69	52.91	50.86	0.05	0.04
2	AlignScore	84.59	74.00	0.53	0.57	75.55	54.46	0.42	0.45	91.20	76.00	0.74	0.74	57.41	53.74	0.21	0.12
3	MC-FT5	88.05	55.27	0.57	0.62	75.34	54.27	0.48	0.45	86.93	79.33	0.80	0.77	57.31	49.96	0.10	0.12
LLM-based																	
4	GPT4o	82.45	53.19	0.59	0.59	60.24	53.97	0.44	0.33	84.40	71.33	0.88	0.87	69.67	52.65	0.51	0.41
5	Gemini	83.11	55.10	0.52	0.55	56.58	54.35	0.54	0.33	89.20	80.67	0.90	0.86	71.18	52.16	0.54	0.47
6	BeSpoke	61.92	53.21	0.35	0.30	66.91	53.58	0.49	0.33	90.13	68.67	0.94	0.92	69.61	53.29	0.41	0.32
7	FineSurE	81.36	68.37	0.49	0.53	62.63	51.86	0.32	0.28	84.93	84.67	0.86	0.85	69.24	57.84	0.41	0.34
Linguistic-informed																	
8	MC+rew.	88.62	61.27	0.59	0.64	76.94	54.23	0.49	0.47	88.53	81.33	0.81	0.76	56.05	50.44	0.08	0.10
9	StructS-MC	91.19	62.54	0.63	0.68	71.68	58.08	0.43	0.38	89.60	80.67	0.90	0.86	54.91	51.30	0.08	0.08
10	BeSpoke+rew.	60.78	55.45	0.34	0.29	71.40	53.85	0.51	0.40	90.13	70.67	0.93	0.90	64.81	55.94	0.32	0.25
11	StructS-BS	89.65	69.72	0.62	0.66	72.03	61.63	0.47	0.40	86.40	74.67	0.86	0.85	58.99	55.67	0.17	0.15
System-level Eval Metrics Ranking Correlation																	
AUC vs. bACC		$\rho = 0.45, \tau = 0.35$				$\rho = 0.41, \tau = 0.24$				ρ = -0.02, $\tau = 0.00$				$\rho = 0.50, \tau = 0.27$			
AUC vs. Pearson		$\rho = 0.89$*, $\tau = 0.84$*				ρ = -0.18, τ = -0.11				ρ = -0.09, τ = -0.04				$\rho = 0.99$*, $\tau = 0.94$*			
BACC vs. Pearson		$\rho = 0.25$, τ =0.18				ρ = -0.19, τ = -0.15				ρ = -0.57, τ = -0.45				$\rho = 0.52, \tau = 0.28$			

Table 12: Summary-level results for all summarization evaluation systems on DIVERSUMM, RAMPRASAD'24, LONGEVAL and UNISUMEVAL. We study two types of summary-level reference labels (Ref Label): (1) Binary, evaluated using ROC-AUC (AUC) and balanced accuracy (bAcc), and (2) Continuous, for which we report Pearson correlation r and Spearman correlation ρ. In each column, the system's performance is ranked into quantile intervals, indicated by specific colors: Green:top 25%, Orange: 50% to 75%, Red:25%-50% and Gray:Bottom 25%. We further examine system-level rank correlations across different metric pairs, finding relatively low correlations, which highlights the challenges in comprehensively evaluating a model's capabilities.

ID	Eval System	DIVERSUMM			RAMPRASAD'24			LONGEVAL			UNISUMEVAL		
	Eval. Task	Sent-level		Summ-level	Sent-level		Summ-level	Sent-level		Summ-level	Sent-level		Summ-level
	Eval. Metric	*AUC*	*bAcc*	*AUC*	*AUC*	*bAcc*	ρ	*AUC*	*bAcc*	r	*AUC*	*bAcc*	r
Specialized fact-checkers													
1	INFUSE	75.62	67.96	87.15	67.30	63.02	0.41	77.84	72.89	0.72	62.31	61.52	0.05
2	AlignScore	72.70	68.73	84.59	74.78	66.06	0.45	76.76	72.24	0.74	66.88	61.90	0.21
3	MC-FT5	74.45	68.40	88.05	68.33	58.82	0.45	86.47	77.88	0.80	68.94	61.19	0.10
LLM-based													
4	GPT4o	74.88	74.88	82.45	61.54	61.54	0.33	88.72	88.72	0.88	75.89	75.89	0.51
5	Gemini	70.47	70.47	83.11	60.33	60.33	0.33	88.49	88.49	0.90	70.50	70.50	0.54
6	BeSpoke	76.53	70.32	61.92	74.35	60.13	0.33	95.11	87.94	0.94	82.39	71.64	0.41
7	FineSurE	76.69	76.69	81.36	64.93	64.93	0.28	84.64	84.64	0.86	77.23	77.23	0.41
Linguistic-informed													
8	MC+rew.	75.89	69.16	88.62	73.41	60.39	0.47	85.97	79.53	0.81	67.73	61.14	0.08
9	StructS-MC	75.74	69.75	91.19	68.66	57.95	0.38	90.37	83.80	0.90	65.57	60.31	0.08
10	BeSpoke+rew.	78.58	71.34	60.78	74.46	63.46	0.40	94.41	87.80	0.93	79.81	72.69	0.32
11	StructS-BS	81.58	75.89	89.65	71.83	66.06	0.40	91.88	86.19	0.86	73.36	68.36	0.17
System-level Eval Metrics Ranking Correlation													
sent-AUC vs. sent-bAcc		$\rho = 0.54, \tau = 0.38$			$\rho = 0.25, \tau = 0.18$			$\rho = 0.70$*, τ =0.56*			$\rho = 0.81$*, $\tau = 0.60$*		
sent-AUC vs. summ-metric		ρ = -0.11, τ = -0.05			$\rho = 0.51, \tau = 0.38$			$\rho = 0.86$*, $\tau = 0.69$*			$\rho = 0.72$*, $\tau = 0.54$*		
sent-bAcc vs. summ-metric		ρ = -0.36, τ = -0.24			$\rho = 0.06, \tau = 0.02$			$\rho = 0.80$*, $\tau = 0.65$*			$\rho = 0.81$*, $\tau = 0.61$*		

Table 13: Sentence-level performance of different evaluation systems on DIVERSUMM, RAMPRASAD'24, LONGEVAL and UNISUMEVAL. We report performances on **sentence-level** (both AUC and bAcc) and also include the official summary-level metric released for each dataset, following the setup in Song et al. (2024). In each column, the system's performance is ranked into quantile intervals, indicated by specific colors: Green:top 25%, Orange: 50% to 75%, Red:25%-50% and Gray:Bottom 25%. We report the rank correlations across sentence-level and summary-level ratings, with * indicating statistical significance (p<0.05).

ID	Eval System	DIVERSUMM						RAMPRASAD'24						LONGEVAL	
		GovReport			ArXiv			BillSum			PubMed			SQuALITY	PubMed
	Eval Metric	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*r*	*r*
Specialized fact-checkers															
1	INFUSE	80.41	59.37	0.43*	71.33	53.88	0.32*	76.16	61.86	0.41*	65.91	55.14	0.09	0.69*	0.55*
2	AlignScore	81.95	73.87	0.44*	71.43	62.64	0.34*	77.12	56.33	0.50*	62.93	48.12	0.14	0.68*	0.50*
3	MC-FT5	74.61	50.66	0.42*	77.19	55.13	0.41*	74.68	53.53	0.47*	64.64	50.00	0.12	0.52*	0.41*
LLM-based															
4	GPT4o	68.37	53.46	0.47*	78.98	51.85	0.48*	61.12	53.77	0.45*	59.00	52.94	0.33*	0.78*	0.63*
5	Gemini	67.99	55.77	0.38*	71.02	51.85	0.33*	57.69	54.81	0.57*	54.82	52.94	0.35*	0.71*	0.74*
6	BeSpoke	79.64	55.14	0.47*	74.79	62.81	0.38*	53.37	52.72	0.44*	62.37	52.94	0.33*	0.84*	0.73*
7	FineSurE	68.76	60.68	0.41*	68.05	62.47	0.27	65.48	50.40	0.39*	60.74	53.40	0.24*	0.75*	0.63*
Lingustic-informed															
8	MC+rew.	76.02	57.96	0.43*	77.93	56.57	0.44*	75.44	53.45	0.48*	66.34	49.40	0.10	0.57*	0.40*
9	StructS-MC	77.83	59.50	0.47*	82.57	56.57	0.51*	69.95	56.89	0.43 *	55.63	48.19	-0.05	0.68*	0.61*
10	BeSpoke+rew.	77.60	67.72	0.50*	72.83	65.11	0.36*	58.25	51.44	0.47*	64.42	52.34	0.27*	0.78*	0.72*
11	StructS-BeSpoke	77.01	69.50	0.50*	79.08	55.31	0.48*	62.06	56.49	0.42*	61.87	54.68	0.17	0.73*	0.53*
Rank Correlation															
AUC vs. bACC		ρ = 0.47, τ = 0.31			ρ = -0.09, τ = -0.02			ρ = 0.42, τ = 0.27			ρ = -0.09, τ = -0.07			–	–
AUC vs. Pearson		ρ = 0.41, τ = 0.31			ρ = 0.99*, τ = 0.95*			ρ = -0.02, τ = 0.00			ρ = -0.47, τ = -0.44			–	–
BACC vs. Pearson		ρ = 0.35, τ = 0.27			ρ = -0.12, τ = -0.07			ρ = -0.09, τ = -0.07			ρ = 0.29, τ = 0.19			–	–

Table 14: Results for all summarization evaluation systems on DIVERSUMM, RAMPRASAD'24 and LONGEVAL. We report performances under **summary-level**. We opt for two types of summary-level labels: (1) Binary, evaluated using ROC-AUC (AUC) and balanced accuracy (bAcc), and (2) Continuous, for which we report Pearson's r correlation, with * indicating statistical significance. LONGEVAL Green, orange and red indicate the performance ranking intervals of the model for each column, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively. We analyze the system-level rank correlations across different metric pairs, revealing that they exhibit relatively low correlations, which underscores the challenges in comprehensively assessing a model's capability.

ID	Eval System	UNISUMEVAL														
		GovReport			PubMed			SQuALITY			MediaSum			MeetingBank		
	Eval Metric	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*	*AUC*	*bAcc*	*r*
Specialized fact-checkers																
1	INFUSE	47.36	46.25	-0.03	56.61	49.93	0.13	42.78	41.74	-0.15	44.44	45.66	-0.08	47.60	46.59	-0.03
2	AlignScore	58.58	50.00	0.08	60.92	50.00	0.21	52.69	54.98	0.09	53.26	49.52	0.08	57.31	55.92	0.32
3	MC-FT5	48.84	49.38	0.01	62.67	48.70	0.18	43.55	43.91	-0.09	58.81	48.65	0.20	51.31	50.13	0.06
LLM-based																
4	GPT4o	72.84	52.27	0.54	71.28	49.67	0.40	72.25	50.45	0.40	61.74	50.00	0.35	69.02	58.26	0.60
5	Gemini	70.60	51.96	0.46	74.81	51.28	0.62	72.92	51.10	0.44	60.64	49.61	0.31	77.57	55.88	0.64
6	BeSpoke	73.12	49.69	0.46	79.90	52.56	0.60	60.03	54.41	0.27	67.74	49.56	0.36	66.72	55.16	0.45
7	FineSurE	77.73	54.55	0.41	79.12	54.15	0.52	62.09	56.52	0.25	59.60	51.72	0.26	66.81	63.40	0.48
Lingustic-informed																
8	MC+rew.	45.62	47.19	-0.01	61.91	52.55	0.17	42.26	44.14	-0.11	56.41	48.10	0.15	50.04	49.03	0.03
9	StructS-MC	45.28	46.88	-0.01	61.62	57.32	0.14	43.72	44.15	-0.12	52.06	50.73	0.11	47.66	46.36	0.04
10	BeSpoke+rew.	60.68	58.47	0.27	78.49	52.56	0.49	53.98	51.81	0.13	61.35	52.07	0.19	65.14	57.58	0.43
11	StructS-BeSpoke	49.94	45.31	0.02	64.97	56.67	0.28	47.03	50.52	-0.06	51.44	50.44	0.07	60.58	57.68	0.30
Rank Correlation																
AUC vs. bACC		ρ = 0.73*, τ = 0.53*			ρ = 0.28, τ = 0.22			ρ = 0.68* , τ = 0.53*			ρ = 0.31, τ = 0.24			ρ = 0.75*, τ = 0.60*		
AUC vs. Pearson		ρ = 0.90*, τ = 0.76*			ρ = 0.89*, τ = 0.78*			ρ = 0.95* , τ = 0.85*			ρ = 0.94*, τ = 0.85*			ρ = 0.98* , τ = 0.93*		
BACC vs. Pearson		ρ = 0.73* , τ = 0.54*			ρ = 0.15, τ = 0.07			ρ =0.65*, τ = 0.45			ρ = 0.23, τ = 0.16			ρ = 0.72*, τ = 0.53*		

Table 15: Results for all summarization evaluation systems on splits of UNISUMEVAL. We report performances under **summary-level**. We opt for two types of summary-level labels: (1) Binary, evaluated using ROC-AUC (AUC) and balanced accuracy (bAcc), and (2) Continuous, for which we report Pearson's r correlation, with * indicating statistical significance. LONGEVAL Green, orange and red indicate the performance ranking intervals of the model for each column, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively. We analyze the system-level rank correlations across different metric pairs, revealing that they exhibit relatively low correlations, which underscores the challenges in comprehensively assessing a model's capability.

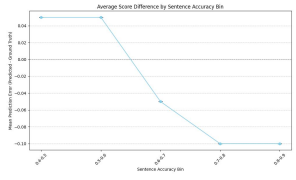

Figure 5: Sample diagram

D Examples on Applying Reliability Diagram

Here we provide a toy example showcasing how we execute the reliability diagram analysis.

Assume we have a dataset with columns in Table 16.

ID	Avg. Sent Acc	Pred. Factuality	Ref. Factuality
1	0.80	0.75	0.85
2	0.60	0.65	0.60
3	0.90	0.85	0.95
4	0.70	0.70	0.75
5	0.50	0.55	0.50

Table 16: Summary Evaluation Metrics

The binning and average is shown in Table 17 and the converted plot is shown as Fig 5.

Bin Range	Avg. Pred. Score - Avg. Ref. Score
[0.0, 0.1]	NaN
(0.1, 0.2]	NaN
(0.2, 0.3]	NaN
(0.3, 0.4]	NaN
(0.4, 0.5]	0.05
(0.5, 0.6]	0.00
(0.6, 0.7]	-0.05
(0.7, 0.8]	0.00
(0.8, 0.9]	-0.10
(0.9, 1.0]	-0.05

Table 17: Binning Result

E Analysis on Model Predictions

E.1 Study 1. Sentence-level Error Analysis

We aggregated the model predictions from 10 systems (rows 1-6, 8-11) in Table 2 and analyzed their sentence-level prediction label. We omitted row 7 as we find their prompting results sometimes do not follow the input sentence splits, thus could potentially introduce noise in the analysis. Specifically, we ranked sentences based on how many systems failed to make correct predictions, highlighting those that posed the greatest challenge across models. Below are findings we drew from analyzing DiverSumm (3143 sentences).

- Finding 1.1: Among the top 200 most frequently misclassified sentences (by at least 9 out of 10 systems), 177 were annotated as NoE. A closer inspection of these cases reveals that many errors stem from noisy annotations (i.e., the sentence is incomplete or contains grammatical errors but was mislabeled as factual). For example, the following sentence is incomplete but labeled as NoE. "*experts noted that public awareness is critical to helping people understand the implications of.*"

- Finding 1.2: We inspected the top-500 challenging sentences and our interest is on the hard-to-detect human-labeled non-factual sentences. The distribution of error types is {'CorefE': 1, 'EntE': 23, 'GramE': 22, 'LinkE': 4, 'OutE': 3, 'PredE': 1}, suggesting that nuanced entity-related (EntE) and grammatical (GramE) errors present consistent difficulties across different systems.

Empirical Comparison on DiverSumm Predictions. We group the systems in Table 2 to compare how well each model group detects different types of errors. The groups are: Specialized models (rows 1-3), LLM-based models (rows 4-6), and Linguistic-featured models (rows 8-11). We use data subsets where each sentence has only one type of error or no error (NoE). Then, we check how accurately each model can detect the presence or absence of that error using binary classification and report the average accuracy score. The result is shown in Table 18.

- Finding 1.3: We group the systems to compare how well each model group detects different types of errors. We observe that the detection improvements brought by LLMs are most notable in reducing false positives (NoE) and in detecting GramE and PredE.

Empirical Comparison on UniSumEval Predictions. On UniSumEval (5299 sentences), we have conducted similar experiments and analyses on UniSumEval on per-error performance.

- Finding 1.4: According to Table 19, LLM-based approaches have improved in reducing false positives during error detection and show better performance on SenE and EntE. However, OutE remains a challenging and persistent issue, with all models continuing to struggle to identify it accurately (below 50%).

Model	NoE	EntE	GramE	OutE	LinkE	PredE	CircE	CorefE
Specialized models	70.16%	62.39%	52.61%	88.19%	76.19%	71.11%	89.74%	55.56%
Δ (LLM vs. specialized model)	8.93%	-1.42%	3.21%	-7.64%	-9.52%	8.89%	-7.69%	-3.70%
Δ (Linguistic-featured vs. specialized)	0.02%	5.56%	9.74%	1.39%	3.81%	12.22%	2.56%	16.67%

Table 18: Comparison of error type performance across Models on DIVERSUMM

Model	NoE	OutE	SenE	EntE	RelE
Specialized models	75.12%	50.31%	45.95%	44.67%	45.68%
Δ (LLM vs. specialized)	19.58%	-4.72%	17.80%	10.00%	-9.88%
Δ (Linguistic-inspired vs. specialized)	13.75%	-0.62%	6.41%	10.83%	7.10%

Table 19: Comparison of error type performance across Models on UNISUMEVAL

E.2 Study 2. Comparing non-LLM-generated and LLM-generated Errors

In this subsection, we investigate the performance of systems in detecting errors in summaries generated by more advanced LLM-based models. This analysis is inspired by the finding that LLMs and older generative models produce different types of factuality errors (Figure 2 in §3.2).

Task Setup: We split the UniSumEval into three subsets: summaries generated by older pre-trained LMs, the open-sourced LMs, and the closed-sourced LMs (an extension of Table 6).

Finding: Our analysis shows that recent LLM-based evaluators perform well in detecting all errors in summaries produced by older models. However, as summarization models have evolved, LLM-based evaluators have shown greater improvement in reducing false positives—avoiding incorrect flags on factually consistent sentences—and in handling SentE cases. However, the LLM-based systems did not show significant performance gains when examining specific error types, and the performance of all evaluation systems is equally low.

Below we depict the model's performance on the three individual splits in Tables 20, 21, and 22, following the same setup in §E.1. We find that for summaries generated by older models, LLM-based approaches improve the detection of errors across all error types. However, as summarization models continue to evolve, these approaches significantly enhance the detection of NoE errors (reducing false positives), while other relation types remain challenging to detect across all three groups of models.

F Extra Experimental Results on Claim-level Approaches

We added experiments using two recent approaches that incorporate claim decomposition into the evaluation task: FENICE (Scirè et al., 2024) and FIZZ (Yang et al., 2024b) by rerunning models using their released codebase. We evaluated them on the same four benchmark datasets, reporting Summary-level AUC and balanced accuracy (bAcc) under the binary setting and Pearson's r under the continuous setting. Our results (Tables 23, 24, 25 and 26) show that claim decomposition-based models generally underperformed or performed comparably to other sentence-level baselines. The only exception is Fenice, which showed improved performance on UniSumEval; however, a performance gap remains between these models and LLM-based systems on that dataset (we use MC-FT5, row 3 of Tables 2 and 3).

Model	NoE	OutE	SenE	EntE	RelE
Specialized models	88.81%	50.81%	37.78%	44.44%	37.50%
Δ (LLM vs. specialized)	2.54%	13.41%	23.89%	18.89%	4.17%
Δ (Linguistic-inspired vs. specialized)	-3.32%	8.03%	12.22%	9.72%	3.12%

Table 20: Error detection performance on summaries generated by older LMs such as BART and T5. LLM-based approaches improve the detection of errors across all error types.

Model	NoE	OutE	SenE	EntE	RelE
Specialized models	76.11%	56.44%	61.90%	48.15%	53.33%
Δ (LLM vs. specialized)	18.06%	-9.85%	10.48%	-3.70%	-13.33%
Δ (Linguistic-inspired vs. specialized)	-0.17%	6.34%	13.81%	14.35%	11.67%

Table 21: Error detection performance on summaries generated by open LMs such as Phi-2. The improvements of LLM-based approaches are more profound in the detection of NoE and SenE.

Model	NoE	OutE	SenE	EntE	RelE
Specialized models	68.22%	36.51%	37.50%	16.67%	33.33%
Δ (LLM - specialized model)	28.57%	-29.37%	4.17%	0.00%	-25.00%
Δ (linguistic - specialized model)	-0.21%	3.37%	25.00%	-4.17%	-2.08%

Table 22: Error detection performance on summaries generated by closed-source LLMs such as GPT-4 and Claude 2.1. LLM-based approaches significantly improve the NoE detection (reducing false positives). However, the other relation types are hard to detect across all three groups of models.

Model	AUC	bAcc	Pearson's r
Fenice	80.00	49.95	0.36
Fizz	87.90	50.91	0.54
MC-FT5	88.05	55.27	0.57

Table 23: Summary-level evaluation results on DIVERSUMM

Model	AUC	bAcc	Pearson's r
Fenice	70.01	50.72	0.33
Fizz	67.49	50.61	0.27
MC-FT5	75.34	54.27	0.48

Table 24: Summary-level evaluation results on RAMPRASAD'24

Model	AUC	bAcc	Pearson's r
Fenice	92.00	56.00	0.82
Fizz	88.53	72.66	0.55
MC-FT5	86.93	79.33	0.80

Table 25: Summary-level evaluation results on LONGEVAL

Model	AUC	bAcc	Pearson's r
Fenice	59.11	50.09	0.19
Fizz	54.55	50.29	0.08
MC-FT5	57.31	49.96	0.10

Table 26: Summary-level evaluation results on UNISUMEVAL