PlainQAFact: Retrieval-augmented Factual Consistency Evaluation Metric for Biomedical Plain Language Summarization

Zhiwen You[a], Yue Guo[a]

[a]*School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, 61820, IL, United States*

Abstract

Hallucinated outputs from large language models (LLMs) pose risks in the medical domain, especially for lay audiences making health-related decisions. Existing automatic factual consistency evaluation methods, such as entailment- and question-answering (QA) -based, struggle with plain language summarization (PLS) due to *elaborative explanation* phenomenon, which introduces external content (e.g., definitions, background, examples) absent from the scientific abstract to enhance comprehension. To address this, we introduce PlainQAFact, an automatic factual consistency evaluation metric trained on a fine-grained, human-annotated dataset PlainFact, for evaluating factual consistency of both source-simplified and elaborately explained sentences. PlainQAFact first classifies sentence type, then applies a retrieval-augmented QA scoring method. Empirical results show that existing evaluation metrics fail to evaluate the factual consistency in PLS, especially for elaborative explanations, whereas PlainQAFact consistently outperforms them across all evaluation settings. We further analyze PlainQAFact's effectiveness across external knowledge sources, answer extraction strategies, answer overlap measures, and document granularity levels, refining its overall factual consistency assessment. Taken together, our work presents the first evaluation metric designed for PLS factual consistency evaluation, providing the community with both a robust benchmark and a practical tool to advance reliable and safe plain language communication in the medical domain. PlainQAFact and PlainFact are available at: `https://github.com/zhiwenyou103/PlainQAFact`.

Keywords:
Plain language summarization, Factual consistency evaluation, Retrieval-augmented generation, Hallucination, Large language models

1. Introduction

Communicating biomedical scientific knowledge in plain language is essential for improving health information accessibility and health literacy [1, 2]. Recent advances in large language models (LLMs) have made significant progress in plain language summarization (PLS) of biomedical texts [3, 4, 5, 6]. However, ensuring the factual consistency of these summaries remains a major challenge. A key source of inconsistency stems from *elaborative explanations*: content such as definitions, background information, and examples that enhance comprehension but are not explicitly present in the original scientific abstracts (i.e., source) [7, 8, 9]. While such elaborations are critical for effective communication, they introduce external content that cannot be directly verified against the source, complicating automatic factual consistency evaluation.

Factual consistency in PLS is typically assessed through a combination of human evaluation and automated metrics [3, 10]. While human evaluation is reliable [11], it is costly and difficult to scale, particularly in biomedical contexts where domain expertise is required. Commonly used factuality metrics can effectively verify content supported by the source but fail to assess factual consistency of added information [12]. However, these metrics depend heavily on high-quality reference summaries, which are often unavailable in plain language summaries. Recent prompt-based evaluation techniques show promise [13, 14], but their sensitivity to factual perturbations in elaborative content remains limited [12].

The lack of suitable benchmark datasets further hinders progress. Many existing datasets are constructed from LLM-generated summaries or apply rule-based perturbations to simulate non-factual content. For example, FactPICO provides expert annotations for plain language summaries of randomized controlled trial abstracts, focusing on PICO elements and evidence inference [9]. However, it includes factuality labels only for added content, leaving simplified sentences unannotated, which are generated by LLMs and potentially inaccurate. In contrast, APPLS perturbs human-written summaries using rule-based transformations [12], but cannot ensure that the resulting outputs remain coherent or factually plausible. These limitations underscore the need for high-quality, sentence-level annotations grounded in human-authored summaries.

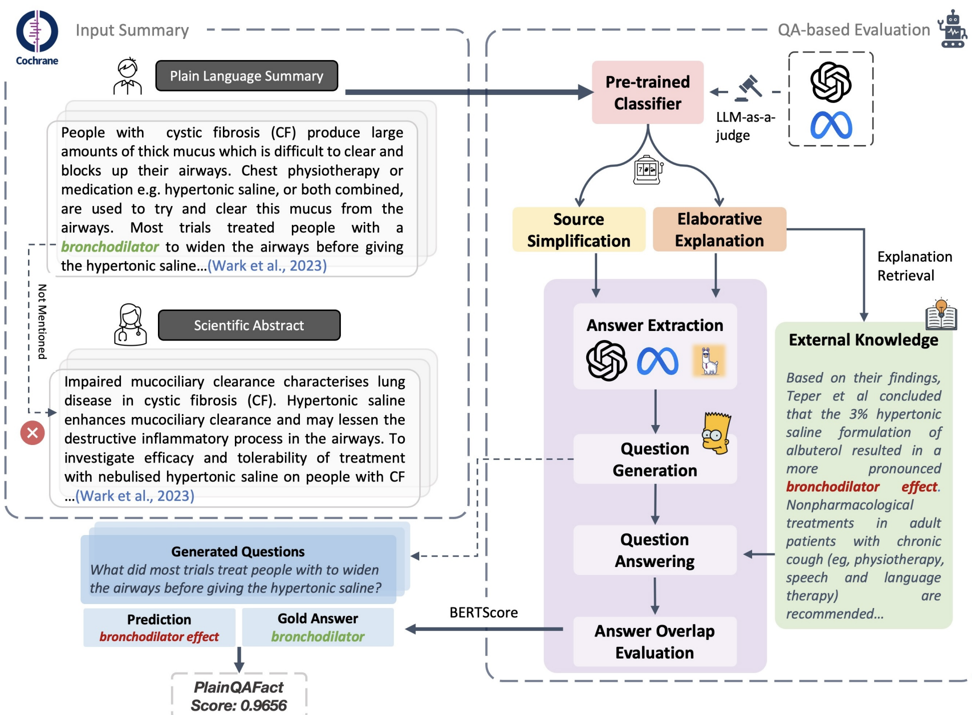

Figure 1: Overview of PlainQAFact. A fine-tuned classifier first identifies the sentence type, involving either source simplification or elaborative explanation. Then, a QA-based evaluation pipeline performs answer extraction, question generation, question answering, and answer overlap evaluation. For elaborative content not present in the scientific abstract, PlainQAFact retrieves external knowledge to verify factual consistency. The illustrated example shows an elaborative explanation involving a "bronchodilator" not mentioned in the source abstract but verifiable through external evidence. PlainQAFact assigns a high score, reflecting strong alignment between the extracted and gold answers.

To address these challenges, we propose a targeted retrieval-based metric for factual consistency evaluation in PLS (Figure 1). We introduce a new expert-annotated dataset, PlainFact, consisting of human-written plain language summaries aligned with scientific abstracts. Each sentence is labeled with its type (elaborative explanation vs. source simplification), functional role, and alignment to the source (§3.1). Building on this, we present PlainQAFact, a dual-stage QA-based evaluation metric that selectively applies retrieval only to elaborative explanations (§3.2). This design ensures both efficiency and fidelity in evaluating factual consistency. Experiment results on several PLS datasets demonstrate the effectiveness of our approach, particularly in assessing complex, elaborative content (4, §5). Taken together, our work presents the first evaluation metric designed for PLS factual consistency evaluation, providing the community with both a robust benchmark and a practical tool to advance reliable and safe plain language communication in the medical domain.

2. Related Work

Limitations of Existing Factuality Evaluation. The primary approach for evaluating plain language generation combines automated metrics with human evaluation [3, 10]. While human assessment provides a thorough analysis [11], its high cost and time demands make it impractical for large-scale datasets. Evaluating factual consistency in biomedical PLS is particularly challenging, as it requires domain expertise. Entailment- [15, 16], similarity- [17, 18], model- [19] and QA-based [20, 21] metrics are commonly used for factual consistency assessment but rely heavily on high-quality reference summaries, which are often unavailable or difficult to obtain for PLS. Recent advancements in prompt-based evaluation show promise [13]; however, their sensitivity to factuality perturbations in PLS remains limited [12]. To address these limitations, we propose a reference-free solution for factual consistency evaluation of PLS that effectively assesses factual consistency with external information retrieval to augment the reference summary.

Retrieval-Augmented Generation. Retrieval-augmented methods enhance text generation by extracting relevant information from external sources to supplement input queries [22]. These methods have been shown to be effective in open-domain QA [23, 24], knowledge-based QA [25], and multi-step reasoning [26, 27]. In the context of PLS, retrieval from structured knowledge bases (KBs) has been shown to improve factual accuracy compared to language models alone [28]. However, retrieval-augmented approaches have not been extensively explored for factual consistency evaluation in PLS, despite their potential for addressing elaborative explanations. In this work, we investigate retrieval-augmented QA to enhance PLS factual consistency assessment while also examining its limitations.

3. Methods

We first introduce the protocol of PlainFact dataset curation (§3.1). Based on this dataset, we propose PlainQAFact, a two-stage retrieval-augmented QA framework for factual consistency evaluation in PLS tasks (§3.2).

3.1. PlainFact Benchmark

To develop a high-quality factual consistency evaluation benchmark in PLS tasks, we collect a subset from the largest human-authored CELLS [28] dataset (§3.1.1) and hire domain experts to provide fine-grained sentence-level annotations (§3.1.2).

3.1.1. Human-Authored PLS Dataset

Rather than relying on LLM-generated plain language summaries, we construct our benchmark using human-authored summaries. CELLS [28], the largest parallel corpus of scientific abstracts and their corresponding plain language summaries, is written by the original authors and sourced from 12 biomedical journals. We primarily select data from the Cochrane Database of Systematic Reviews (CDSR)[1] within CELLS, as CDSR contains systematic reviews that support evidence-based medical decision-making across healthcare domains [29]. Since systematic reviews represent the highest level of scientific evidence, this selection enhances the factual rigor of our dataset. To ensure readability, we filter the 200 most readable plain language summaries based on average scores from three standard readability metrics: Flesch-Kincaid Grade Level (FKGL) [30], Dale-Chall Readability Score (DCRS) [31], and Coleman-Liau Index (CLI) [32]. Additionally, given the collected plain language summaries from CELLS are all factual, we further conduct sentence-level perturbation to transform each plain language sentence into incorrect ones. Specifically, we use `GPT-4o`[2] to perturb each plain language sentence based on prompts introduced in Guo et al. [12]. Therefore, for PlainFact, we have 400 summary-abstract pairs in total, where 200 factual pairs and 200 non-factual pairs.

[1] `http:///www.cochranelibrary.com`

[2] We use `gpt-4o-2024-11-20` version for all experiments of `GPT-4o` in this paper.

	Elaborative Explanation	Source Simplification
# of Sentences	1,213	1,527
Average Length (token)	29	28
Vocabulary Size	4,230	4,046
Has Reference	417	1,527
# of Background	533	329
# of Definition	82	44
# of Method/Result	512	1,107
# of Example	10	3
# of Other	76	44
FKGL ↓	12.5	12.4
DCRS ↓	11.3	11.6
CLI ↓	13.5	13.9

Table 1: Overview of the PlainFact benchmark. Medical experts annotated 200 pairs of plain language summaries and their corresponding scientific abstracts from the CELLS dataset [28], categorizing each plain language sentence as either a source simplification or an elaborative explanation related to the abstract. Lower scores of FKGL, DCRS, and CLI indicate better readability.

3.1.2. Expert Annotation

Since each summary-abstract pair is authored by the same individual, we assume all information to be factual. The annotation aims to capture how plain language sentences relate to their scientific abstracts. Annotators analyze each plain language sentence across three dimensions: (1) **Factuality type**: identify whether a sentence is a source simplification (derived from the abstract) or an elaborative explanation (introducing new content); (2) **Functional role**: categorize the sentence as background, definition, example, method/result, or other; and (3) **Sentence alignment**: map each plain language sentence to its corresponding sentence(s) in the scientific abstract. Details of annotation guidelines are provided in Appendix A.

Annotations are conducted by four independent annotators, each with at least a bachelor's degree in biomedical sciences and prior fact-checking experience. Annotators are recruited via Upwork and compensated from $15 to $20 per hour. Each summary-abstract pair is annotated by two independent annotators, with disagreements resolved by a third. Inter-rater agreement, measured by Cohen's Kappa for factuality type, functional role, and sentence alignment are 0.43, 0.60, and 0.55 respectively, indicated moderate agreement for all tasks [33].

Table 3.1.1 summarizes the dataset characteristics. Notably, 44% of plain language sentences are elaborative explanations, highlighting their role in enhancing the readability of plain language summaries. 66% of these cannot be directly verified against the source abstract, which underscores the need for factuality evaluation methods that account for such phenomena. Moreover, elaborative explanations include more background, definitions, and examples than source simplifications. For annotation examples, see Appendix B.

3.2. PlainQAFact Framework

PlainQAFact conducts fine-grained factual consistency evaluation for plain language summaries by first segmenting each summary into sentences, classifying as either simplification or explanation, and retrieving tailored external knowledge within a retrieval-augmented QA framework. Figure 1 provides an overview of its three key components: sentence-level classification (§3.2.1), domain knowledge retrieval (§3.2.2), and dual-stage QA-based factual consistency evaluation (§3.2.3).

3.2.1. Learned Factuality Type Classifier

As elaborative explanations are prevalent in plain language generation and existing metrics struggle to capture added information [12], we first fine-tune a pre-trained language model to categorize factuality types of plain language sentences. Based on the factuality type annotations (source simplification vs. elaborative explanation) in PlainFact, we fine-tune the `PubMedBERT-base` model[3] as a classifier. Our PlainFact is split 8:1:1 for training, validation, and testing. Additionally, we compare the pre-trained classifier with GPT-4o [34] as a zero-shot classifier (prompts are provided in Appendix C) to explore extending our factual consistency evaluation metric to domains and tasks lacking human-annotated data.

3.2.2. Domain Knowledge Retrieval

Retrieval-augmented methods have proven effective for explanation generation [28, 35], making them a natural fit for evaluating elaborative explanations. Since our dataset is in the medical domain, we employ MedCPT [36], a retriever pre-trained on large-scale PubMed search logs to generate biomedical text embeddings. For external resources, we incorporate StatPearls [37]

[3] `https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext`

for clinical decision support and medical textbooks [38] for domain-specific knowledge.

3.2.3. QA Evaluation Components

QA-based metrics have proved to be effective than other factual consistency evaluation metrics [12] in general summarization tasks and align more closely with human annotations [20, 39]. In this study, we adopt a QA-based approach as a backbone of our evaluation metric to verify factual consistency, incorporating answer extraction, question generation, question filtering, question answering, and answer overlap calculation modules.

Gold Answer Extraction. The first step in QA-based factual consistency evaluation is to extract answer entities (keyphrases) from plain language summaries as gold answer, then verify factuality by comparing them with answers generated by a QA model for the same questions. If the generated answer is correct or relevant, the summary is considered factual. The previous study used PromptRank [40], a keyphrase generation method based on the `T5` model. To compare answer extraction strategies, we use PromptRank as the baseline, and evaluate LLM-based extractors, including an open-source `Llama 3.1 8B Instruct` model (`Llama 3.1`) [41] and close-source `GPT-4o` (`GPT-4o`) [34]. Prompts for `Llama 3.1` and `GPT-4o` are in Appendix C.

Question Generation (QG). Given an input plain language summary, the QG model generates questions based on the extracted answers (§3.2.3) and the input plain language summary. Following previous studies [20, 21][4], we fine-tune `BART-large` model on standard QG datasets, including SQuAD [42] and QA2D [43], for use as the QG model in our evaluation metric. The QG model will generate multiple relevant questions based on the input, which will increase the probability of verifying the factual consistency by answering these questions in the following stages.

Question Filtering (QF). Questions generated by the QG model (§3.2.3) may not always be answerable. For example, the QG model generates "<What can occur over several days?>," which is vague and hard to predict the correct answer. To prevent these unanswerable questions from impacting the final

[4]These studies use allennlp, an open-source NLP research library built on PyTorch, which is no longer actively maintained.

evaluation performance, we follow QAFactEval [39] and remove the unanswerable questions using a pre-trained `Electra-large` model [44]. During QF, the filtering model receives only the plain language summary and its corresponding questions. In the subsequent QA stage (§3.2.3), answers are extracted from the source for answer overlap evaluation, while QF only determines whether the questions can be answered by the plain language summary.

Question Answering. The QA model extracts answers to answer the filtered questions from the source document. To prevent hallucinated output, we use an extractive QA model, a pre-trained `Electra-large`, which was the best performing QA model in QAFactEval [39].

Answer Overlap Evaluation (AOE). We evaluate the alignment between the gold and generated answers from the QA model using the baseline Learned Evaluation metric for Reading Comprehension (LERC) score [45] and BERTScore [46]. The final step in PlainQAFact is weighting the factual consistency scores for source simplification and elaborative explanation sentences. Specifically, the simplification score s is computed using only abstracts as the source in QA, while the explanation score e is calculated by incorporating both abstracts and retrieved knowledge as source for QA-based evaluation.

$$\text{PlainQAFact} = \frac{s_{\text{Avg.}} \cdot n_s + e_{\text{Avg.}} \cdot n_e}{n_s + n_e}, \tag{1}$$

where n_s is the number of simplification sentences, and n_e is the number labeled as explanation. $s_{\text{Avg.}}$ and $e_{\text{Avg.}}$ denote the average PlainQAFact scores computed for instances classified as simplification and explanation sentences, respectively.

4. Experiments

We conduct main experiments on three publicly available datasets, including PlainFact, CELLS [28], and FareBio [47]. To verify the effectiveness of our proposed PlainQAFact on PLS tasks, we compare with five widely used factual consistency evaluation metrics and two LLM-based evaluators.

4.1. Datasets

As introduced in Sec 3.1, we create PlainFact with fine-grained sentence-level annotations regarding the factuality types. Additionally, to test the generalizability of our pre-trained classifier, we also collect another subset from the test set of CELLS dataset [28], gathering 200 summary-abstract pairs. Similar to PlainFact (§3.1.1), we also conduct sentence-level perturbation for each plain language summary in our collected CELLS dataset. Therefore, we have 200 factual and 200 non-factual summary-abstract pairs for both PlainFact and CELLS datasets. We also compare PlainQAFact on the FareBio dataset, which contains 175 plain language summaries generated by seven different LLMs, together with the source articles [47]. Specifically, FareBio includes 25 full scientific article examples, each source article is summarized by seven LLMs, resulting 175 plain language summaries in total. We sample plain language sentences that are not both faithful and factual as non-factual sentences, and the rest as factual sentences, resulting in 33 and 174 summaries respectively. We also extract sentences that are labeled as not faithful but factual in their "factual hallucination" label, which represents the plain language sentence that is not from the source article, but still conveying factual information, resulting in 53 valid summaries in total. These sentences are equal to our factual elaborative explanation sentences. Note in FareBio dataset, there are no non-factual simplification sentences. Therefore, we use the same set of non-factual sentences (33 sentences) for main experiment (Table 5) and explanation-only (Figure 2) evaluation.

4.2. Experiment Settings

The classifier used in PlainQAFact is fine-tuned on PlainFact with a batch size of 32 for 10 epochs. We apply early stopping during training based on validation loss. The temperature is set to 0 for `GPT-4o` and 0.01 for `Llama 3.1`. The `Llama 3.1` model used in AE also uses a temperature of 0.01. For the QG model, we fine-tune `BART-large` with a batch size of 16, a learning rate of 3e-5, and two training epochs. The maximum input length is set to 512 for all models in PlainQAFact. More details of prompts and model hyper-parameters for `GPT-4o` and `Llama 3.1` are provided in Appendix C and Appendix D.

4.3. Existing Factuality Metrics

To the best of our knowledge, no prior work has efficiently evaluated factuality metrics for detecting elaborative explanations in PLS tasks. Moreover, most factuality metrics are designed for general-domain applications, largely due to limited quality and annotation of existing PLS datasets. To address this gap, we incorporate the following metrics in our experiment, informed by prior work [9]: (1) **Dependency-Arc Entailment (DAE)** [48] is an entailment-based method that evaluates summary factuality by breaking it into smaller entailment tasks at the arc level. The model independently determines whether each arc's relationship is supported by the input. (2) **AlignScore** [19] is a model-based factuality metric using `RoBERTa` [49]. It extracts claims from the summary and calculates the alignment scores with all the context chunks from the input document. The final score is an average of all highest alignment probabilities. (3) **SummaC** [16] is an NLI-based inconsistency detection method designed for summarization tasks. It segments documents into sentences and aggregates scores between sentence pairs using a trained NLI model. (4) **QAFactEval** [39] is a QA-based factuality evaluation metric that assesses summary consistency by generating questions from the source document and comparing the model-generated answers with expected answers. (5) **QuestEval** [50] is a reference-less evaluation metric employs a `T5` model to generate questions from the source document and verifies whether the summary can correctly answer them.

5. Results and Analysis

We first conduct a pilot study on the FactPICO dataset [9] to investigate the performance of existing factual consistency evaluation metrics on PLS tasks. Our study (Appendix G) reveals two major limitations. First, many LLM-generated plain language summaries in FactPICO contain incomplete or incoherent sentences, highlighting the need for a benchmark based on high-quality, human-authored plain language summaries. Second, existing automatic factual consistency metrics are not sensitive to evaluate the factuality of "added information", underscoring the need for a more domain-related factual consistency evaluation metric, especially for those summaries with additional explanations that are not mentioned by the original source texts. To address these challenges, we introduce a new expert-annotated dataset, PlainFact, and propose PlainQAFact, a QA-based, retrieval-augmented metric for evaluating factual consistency in plain language summaries.

In this section, we first benchmark PlainQAFact against five widely used automatic factual consistency evaluation metrics on three PLS datasets (§5.1). We then evaluate its performance in an explanation-only setting using PlainFact (§5.2), followed by an ablation study to assess the contribution of each component (§5.3). Next, we conduct an error analysis to examine failure cases and highlight directions for further improvement (§5.4).

Datasets	Metrics	Kendall's τ	Pearson	AUC-ROC
CELLS	Llama 3.1	57.5	42.1	83.4
	GPT-4o	70.0	75.4	99.3
	QAFactEval	61.6	70.6	93.5
	QuestEval	23.2	28.3	66.4*
	SummaC	24.8	29.8	67.5*
	AlignScore	56.0	66.2	89.6
	DAE	14.4	14.4	55.0*
	PlainQAFact	62.1	75.2	93.8
FareBio	Llama 3.1	40.5	38.9	83.0
	GPT-4o	27.3	35.7	75.7
	QAFactEval	32.8	43.9	81.6
	QuestEval	37.8	45.2	86.4
	SummaC	33.8	32.1	82.5
	AlignScore	32.2	47.2	81.0
	DAE	8.8	8.8	53.4
	PlainQAFact	16.8	35.6	66.1
PlainFact	Llama 3.1	60.7	52.6	85.3
	GPT-4o	69.8	80.3	99.2
	QAFactEval	65.8	79.3	96.4
	QuestEval	28.4	34.7	70.1*
	SummaC	34.1	42.8	74.1*
	AlignScore	60.2	72.6	92.5
	DAE	5.3	5.3	51.8*
	PlainQAFact	65.7	81.3	96.4

Table 2: Evaluation results of automatic metrics on the CELLS [28], PlainFact, and FareBio [47] datasets. Results are evaluated using Kendall's τ, Pearson correlation, and AUC-ROC, with eight metric scores compared against human-labeled factuality. The standard deviations (std.) over five runs are: 0.2 (PlainQAFact), 0.5 (`Llama 3.1`), and 6.0 (`GPT-4o`) on PlainFact; 0.1 (PlainQAFact), 0.2 (`Llama 3.1`), and 6.3 (`GPT-4o`) on CELLS; and 0.1 (PlainQAFact), 2.9 (`Llama 3.1`), and 17.0 (`GPT-4o`) on FareBio. * indicates an improvement of our metric over prior work with 95% confidence interval (details in Appendix E). **Note the plain language summaries in FareBio are generated by seven LLMs, while CELLS and PlainFact are written by the authors of original articles.**

5.1. Main Results

We report rank correlation (Kendall's τ), linear correlation (Pearson), and discrimination (AUC-ROC [51]) for three datasets (PlainFact, CELLS [28], and FareBio [47]) following previous studies [16, 39, 49] in Table 5. Overall, `GPT-4o` performs the best on CELLS and PlainFact, but its performance drops on FareBio. DAE is consistently the weakest across all datasets and metrics, suggesting that a dependency-only metric is insufficient for capturing the factual consistency required in PLS tasks.

On CELLS, PlainQAFact outperforms all other automatic metrics on all three criteria (Kendall's τ=62.1, Pearson=75.2, AUC-ROC=93.8). This shows that our approach, first detecting elaborative content and then verifying it with retrieval, works better than existing QA-based metrics, which ask all questions from the source only, and also better than NLI-based evaluation metrics. Similarly, on PlainFact, PlainQAFact achieves the highest Pearson (81.3), surpassing `GPT-4o` by 1.0 point. It ties AUC-ROC (96.4) score with QAFactEval and has a competitive Kendall's τ (65.7). These results on both datasets show that PlainQAFact is reliable for evaluating both factual and non-factual plain language summaries.

On FareBio, no single method performs the best on all criteria. QuestEval achieves the best AUC-ROC (86.4), AlignScore has the best Pearson (47.2), while `Llama 3.1` reaches the best Kendall's τ (40.5). However, FareBio differs from the other two datasets in several ways: (1) the original labels in FareBio are defined at the sentence level, and a summary is marked as non-factual if any sentence is non-factual; (2) their summaries are generated from full scientific articles rather than abstracts, making it harder for metrics to locate matching source text; (3) the dataset is imbalanced (33 non-factual vs. 174 factual); and (4) All plain language summaries are generated by LLMs and no elaborative explanation annotations are provided for each sentence or summary. These factors likely increase evaluation difficulty on FareBio and impact the performance of multiple metrics, including PlainQAFact and `GPT-4o`.

In summary, PlainQAFact is a more consistent and effective metric for factual consistency compared to existing automatic approaches. It clearly improves over prior QA-based metrics on CELLS and matches or exceeds them on PlainFact. Although `GPT-4o` performs better in most of the settings, our PlainQAFact is built entirely on the open-source LLM (i.e., `Llama 3.1`), ensuring transparency, reproducibility, and accessibility. Its two-step evaluation process, detecting explanations and verifying them with retrieval, helps align metric scores more closely with human judgments in PLS evaluation tasks.

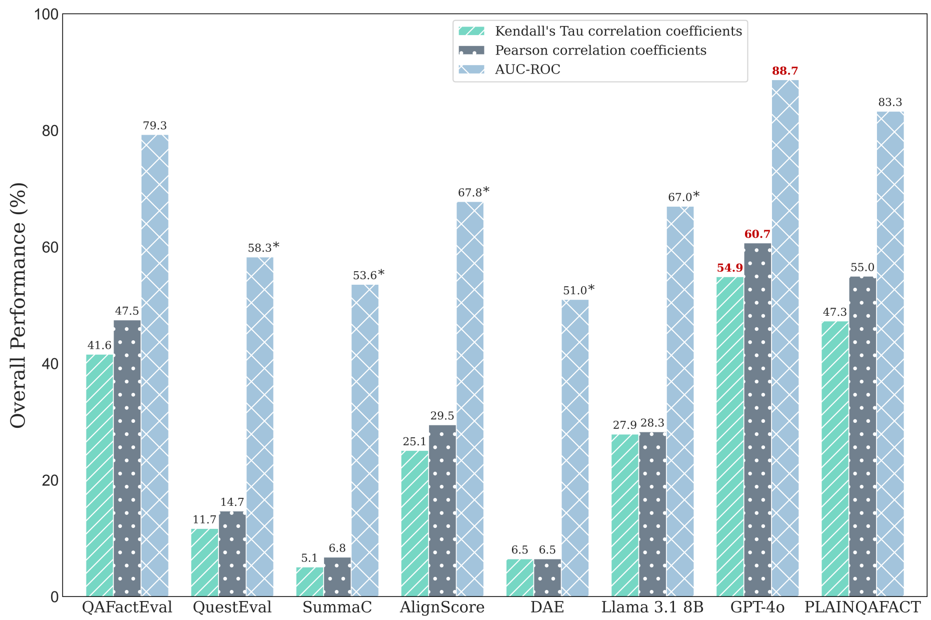

Figure 2: Overall performance on human-annotated elaborative explanation summaries from PlainFact (392 summaries). The std. of PlainQAFact, `Llama 3.1`, and `GPT-4o` are 0.1, 1.0, and 7.7, respectively based on five runs for each metric. * indicates an improvement of our metric over prior work with 95% confidence interval (details in Appendix E). PlainQAFact significantly outperforms most of the automatic factual consistency evaluation metrics in AUC-ROC. Note that the CELLS dataset does not contain annotations for elaborative explanations. Results of explanation-only evaluation on FactPICO and FareBio are reported in Appendix F.

5.2. Explanation-Only Evaluation

To test whether existing evaluation metrics are limited in assessing elaborative explanations in PLS, we evaluate PlainQAFact on human-annotated explanation-only summaries from PlainFact. We select only the sentences labeled as "elaborative explanation" and group them into plain language summaries, resulting in 392 summary–abstract pairs for PlainFact. Note that CELLS does not provide sentence-level annotations for elaborative explanations, so we evaluate it only as a general PLS dataset (§5.1).

As shown in Figure 2, PlainQAFact outperforms all existing automatic metrics across all three criteria. It is 5.7 points higher in Kendall's τ, 7.5

points higher in Pearson, and 4.0 points higher in AUC-ROC compared with the second-best metric, QAFactEval. These results indicate that PlainQAFact captures the factual consistency of elaborative explanations more effectively than prior automatic metrics. While `GPT-4o` achieves higher overall scores (Kendall's τ=54.9, Pearson=60.7, AUC-ROC=88.7), it shows much higher variance (std. 7.7) compared to PlainQAFact (std. 0.1), making it less stable for factual consistency evaluation.

Compared to the results in Table 5, AlignScore is competitive on general PLS tasks but drops on evaluating explanation-only sentences. In contrast, PlainQAFact remains consistently strong when we isolate explanation-only content. PlainQAFact detects added elaborative explanations and checks them with domain knowledge retrieval, which helps when factual consistency depends on facts beyond simple restatement from the source. Since 44% of sentences in our curated benchmark are labeled as elaborative by human annotators, we believe that PlainQAFact is the more suitable and robust factual consistency evaluation metric in this scenario. PlainQAFact also outperforms `Llama 3.1` on all three criteria (27.9/28.3/67.0), showing that instruction-tuned LLM judges alone are not enough for reliable evaluation of elaborative content.

5.3. Ablation Study

PlainQAFact consists of several modules, including a fine-tuned classifier, QA modules, a retrieval function, and an answer overlap evaluation process. To understand the contribution of each component, we conduct an ablation study on PlainFact (§5.3). Table 5.2 summarizes the ablation results, where each component of PlainQAFact is individually modified to measure its impact on overall performance.

Fine-tuned Classifier Table 5.2 shows that removing the classifier and retrieving for every sentence **does not** improve factual consistency assessment. Performance drops across all criteria (Kendall's τ from 65.7 to 61.2; Pearson from 81.3 to 74.2; AUC-ROC from 96.4 to 93.2), and the std. increases to 0.7. This approach also requires notably more computation. Using `GPT-4o` as a simple classifier provides reasonable results but still underperforms compared to our fine-tuned classifier. These findings support the use of a lightweight fine-tuned classifier to trigger retrieval only when necessary.

Answer Extraction For extracting gold answers from plain language summaries, the LLM-based extractors are better than PromptRank [40]. `GPT-4o`

Component	Method Choice	Kendall's τ	Pearson	AUC	std.
PlainQAFact		65.7	81.3	96.4	0.2
Classifier	**Fine-tuned classifier**	-	-	-	-
	GPT-4o	62.8	77.1	94.3*	0.1
	No (retrieve for all)	61.2	74.2	93.2*	0.7
Answer Extraction	**Llama 3.1**	-	-	-	-
	GPT-4o	66.9	81.9	97.2	0.6
	PromptRank	66.3	80.2	96.0	-
Retrieval Source	**Abs + TB + SP**	-	-	-	-
	TB + SP	63.4	76.8	94.8	0.1
	Abs (no retrieval)	62.8	77.1	94.3*	0.1
Answer Overlap	**BERTScore**	-	-	-	-
	LERC	64.5	75.9	95.5	0.1
Granular Level	**Sentence**	-	-	-	-
	Summary	49.3	59.6	84.7*	0.5

Table 3: Ablation study of PlainQAFact on PlainFact, analyzing the impact of individual components in PlainQAFact. Abs: abstract; TB: Textbooks; SP: StatPearls. AUC represents AUC-ROC. The first row of each component setting represents our best combination. * indicates an improvement of our metric over prior work with 95% confidence interval. We run each setting for five times and report the standard deviation (std.). Note that only the settings using LLMs may produce fluctuated scores.

achieves the best overall scores (Kendall's τ=66.9, Pearson=81.9, AUC-ROC=97.2), with small gains over our Llama-based method. PromptRank performs better on Kendall's τ (increases 0.6), but fails on other two criteria compared to our best combination. Although `GPT-4o` achieves slightly higher performance, the marginal gain does not justify the increased cost, making `Llama 3.1` a more practical and cost-effective choice for large-scale factual consistency evaluation.

Combined Domain Resources for Retrieval The highest performance is achieved when combining all three sources (Abs + TB + SP). This suggests that comprehensive retrieval, which includes the original source abstracts, improves overall factual consistency assessment. Compared to using only external sources (TB + SP), performance drops by 2.3/4.5/1.6 across the three criteria. Using only abstracts (i.e., no elaborative retrieval) also lowers performance (2.9/4.2/2.1). To further examine the effectiveness of retrieval for source simplification and elaborative explanation, Table 5.3 breaks down

Component	**Method Choice**	**Kendall's τ**	**Pearson**	**AUC**	**std.**
Source Simplification	Abs	63.5	79.0	94.8	0.2
Elaborative Explanation	Abs	32.8	38.2	72.8*	0.2
	TB	31.2	37.2	70.6*	0.1
	SP	33.3	37.8	72.5	1.7
	TB + SP	36.7	42.0	75.4	0.1
	Abs + TB	40.3	47.0	78.2	0.2
	Abs + SP	40.5	47.4	78.3	0.2
	Abs + TB + SP	**44.2**	**51.3**	**81.0**	0.1
Full Dataset	Abs	62.8	77.1	94.3*	0.1
	TB	60.8	73.8	93.0*	0.1
	SP	61.3	74.2	93.3*	0.1
	TB + SP	63.4	76.8	94.8	0.1
	Abs + TB	64.3	79.7	95.4	0.2
	Abs + SP	65.0	80.3	95.9	0.2
	Abs + TB + SP	**65.7**	**81.3**	**96.4**	0.2

Table 4: Ablation study results on the PLAINFACT, evaluating how different retrieval sources affect various information types using PLAINQAFACT. The fine-tuned classifier categorizes input sentences as **source simplification** or **elaborative explanation**. Overall, the numbers of factual summaries that only include source simplification and elaborative explanation sentences are 198 and 191 respectively, and 198 and 194 for non-factual summaries. Abs: abstracts; TB: Textbooks; SP: StatPearls. AUC represents AUC-ROC. * indicates an improvement of Abs + TB + SP over other settings with 95% confidence interval. We run each setting for five times and report the standard deviation (std.) in the brackets.

the results as follows:

(1) *Source simplification*: These are sentences in plain language summaries classified as "source simplification." We evaluate summaries containing only simplification sentences, using abstracts as the source. Using abstracts as sole source texts achieves strong performance (63.5/79.0/94.8) across all three criteria.

(2) *Elaborative explanation*: For summaries containing only explanation sentences, abstracts alone are not sufficient (32.8/38.2/72.8), performing much worse than in simplification cases. Adding abstracts with external sources improves performance, with the best results obtained by combining all three sources (Abs + TB + SP).

(3)*Full dataset*: Retrieval from StatPearls provides better results than retrieval from Textbooks, highlighting the importance of using high-quality, domain-specific knowledge bases for PLS evaluation. Overall, the best combination across all settings is **Abs + TB + SP** (65.7/81.3/96.4).

These findings show that effective retrieval for factual consistency should include both the source abstract and reliable external medical references, especially when summaries contain elaborative explanations.

Answer Overlap Evaluation BERTScore outperforms the LERC approach introduced in QAFactEval [21], with gains of 1.2/5.4/0.9 across the three evaluation criteria. This performance gap can be attributed to the fundamental differences between the two metrics. LERC is a learned model trained to predict human factual consistency scores, which may reduce its generalizability to new domains or sentence types. In contrast, BERTScore computes embedding-based semantic similarity, allowing it to capture fine-grained semantic overlap between generated and gold answers. Since PlainQAFact evaluates factual consistency by comparing model-generated answers to those extracted from the source, BERTScore's sensitivity to semantic alignment makes it a more effective and robust choice for this step of the evaluation.

Sentence-level vs. Summary-level Evaluation To further examine the effect of input granularity in factual consistency evaluation, we deactivate the sentence-splitting function and instead pair each plain language summary with an abstract to evaluate factual consistency at the summary level. Our best sentence-level evaluation achieves an AUC-ROC of 96.4, while the summary-level setting performs worse across all three criteria (Kendall's τ=49.3, Pearson=59.6, AUC-ROC=84.7; std. 0.5). Evaluating at the summary level introduces broader contextual dependencies, which can omit sentence-specific factual errors or create inaccurate entailments. These results suggest that processing one sentence at a time makes QA-based question generation and answering more focused and reduces noise from unrelated context.

5.4. Error Analysis

To analyze cases where PlainQAFact fails on the PlainFact benchmark, we categorize one correct example and three types of errors in Table 5.3. For each case, we present the original plain language sentence, the corresponding scientific abstract with retrieved knowledge (from medical textbooks and StatPearls), the model-generated questions based on extracted answers, and the QA model's final responses. Each question is generated based on the extracted answer and its corresponding plain language sentence.

Correct Case	**Plain Language Sentence:** Limits to the availability of SSB in schools (e.g. replacing SSBs with water in school cafeterias). [52] **Scientific Abstract:** Frequent consumption of excess amounts of sugar-sweetened beverages (SSB) is a risk factor for obesity, type 2 diabetes, cardiovascular disease and dental caries... [52] **Generated Question:** Limits to the availability of what in schools? **Retrieved Knowledge:** Even though additional data is required to determine the impact of limiting the availability of nutrient-poor or high-sugar goods in schools on obesity, some study results have shown a net-positive result... **Final Answer:** nutrient-poor or high-sugar goods
Misclassification	**Plain Language Sentence:** This review looked at how well the methods worked to prevent pregnancy, if they caused bleeding problems, if women used them as prescribed, and how safe they were. [53] **Scientific Abstract:** To compare the contraceptive effectiveness, cycle control, compliance (adherence), and safety of the contraceptive patch or the vaginal ring versus combination oral contraceptives (COCs)... [53] **Generated Question:** This review looked at how well the methods worked to prevent what? **Retrieved Knowledge:** Appropriate treatment for the underlying etiology should start as soon as possible, and the patients and family members should receive appropriately targeted education... **Final Answer:** cycle control
Retrieval Quality	**Plain Language Sentence:** The patch is a small, thin, adhesive square that is applied to the skin. [53] **Scientific Abstract:** Users of the norelgestromin-containing patch reported more breast discomfort, dysmenorrhea, nausea, and vomiting. In the levonorgestrel-containing patch trial, patch users reported less vomiting, headaches, and fatigue...[53] **Generated Question:** The patch is a small, thin, what kind of square applied to the skin? **Retrieved Knowledge:** Nonstick dressing Petrolatum-infused gauze strip or other material to form a bolster over the graft site. This may be sutured or taped securely in place to provide some pressure and to keep graft immobilized. **Final Answer:** Nonstick dressing Petrolatum-infused gauze strip
Unanswerable Question	**Plain Language Sentence:** Government officials, business people and health professionals implementing such measures should work together with researchers to find out more about their effects in the short and long term. [52] **Scientific Abstract:** To assess the effects of environmental interventions (excluding taxation) on the consumption of sugar-sweetened beverages and sugar-sweetened milk, diet-related anthropometric measures and health outcomes, and on any reported unintended consequences or adverse outcomes...[52] **Generated Question:** What officials, business people and health professionals implementing such measures should work together with researchers to find out more about their effects in the short and long term? **Retrieved Knowledge:** Implementation should be accompanied by high-quality evaluations using appropriate study designs, with a particular focus on the long-ter effects of approaches suitable for large-scale implementation. **Final Answer:** long-ter effects

Table 5: Error analysis of PlainQAFact with intermediate metric outputs. "Retrieved Knowledge" refers to the source texts retrieved for each plain language sentence. We present correct and failure cases sampled only from elaborative explanation examples (i.e., sentences classified as explanations during evaluation). The correct case has a PlainQAFact score above 0.6, while the failure cases have scores below 0.5.
Color legend: extracted answer, answer origin, correct answer, incorrect answer.

In the correct case, the QG model generates a question from the plain language sentence and the extracted answer "SSB." The QA model then provides the correct answer using the retrieved knowledge rather than the original abstract, demonstrating PlainQAFact 's effectiveness in evaluating explanation sentences.

In the first failure case, the extracted answer is "pregnancy," but the QA model returns "cycle control" from the abstract because the correct term is missing in the retrieved content. Since this sentence is classified as an "elaborative explanation," which requires external knowledge for verification, this error points to a potential misclassification between *simplification* and *explanation* by the fine-tuned classifier. The second failure illustrates how noisy retrieved knowledge can impair evaluation. The QA model provides an irrelevant answer from the retrieved content that does not match the extracted gold answer. The third case shows an unanswerable question generated by the QG model (e.g., regarding "Government"), highlighting a limitation of QA-based factual consistency evaluation. Some questions remain unanswered even with retrieval. Additionally, we acknowledge that some plain language sentences may not generate any questions, resulting in empty question sets. Addressing these challenges requires further improvements in classification, retrieval, and domain-specific question generation.

6. Discussion and Conclusion

Our study advances the assessment of hallucinations in the field of PLS by supplying both the first domain expert-annotated biomedical PLS benchmark PlainFact and a novel retrieval-augmented QA-based factual consistency evaluation metric PlainQAFact. Unlike existing biomedical domain corpora that either lack fine-grained labels for elaborative content [54] or focus solely on text simplification (e.g., FactPICO[9]), PlainFact captures sentence-level distinctions, including simplification versus explanation, functional roles, and explicit alignment to source sentences. This granularity not only enables precise error analysis of hallucinations introduced for clarity, but also provides a reusable framework for researchers in other disciplines (e.g., legal, technical) to replicate our annotation protocol and build high-quality factual consistency datasets tailored to their specialized texts. By offering these data and detailed annotation protocol (Appendix A), we anticipate community-driven extensions, such as multilingual adaptations or integration with domain-specific ontologies that will democratize development of trustworthy summarization models.

Building on this resource, PLAINQAFACT differentiates from prior QA-based metrics (e.g., QAFactEval[21], QuestEval[50]) by first classifying whether a sentence requires external verification given elaborative explanations are often ignored by metrics relying solely on source text, and improve the evaluation efficiency without retrieving for every plain language summary instance. Then, retrieving targeted domain knowledge for each elaborative explanation before posing and answering fact-checking questions. The ablations demonstrate that selective retrieval yields substantial gains in Kendall's τ, Pearson correlation, and AUC-ROC over both end-to-end LLM judges[41] and alignment-based scorers[19], highlighting that judicious incorporation of external evidence is essential for robust factual consistency assessment. Beyond its empirical strengths, PLAINQAFACT serves as a plug-and-play evaluation metric for future summarization systems. Developers can leverage our classifier to classify sentences, apply retrieval only where necessary, thus containing computational costs and obtain interpretable question–answer pairs that manifest non-factual hallucinations. More importantly, PLAINQAFACT is developed with the open-source backbone model (i.e., `Llama 3.1`), ensuring broad use with transparency, reproducibility, and fewer computational costs. Looking forward, extending our proposed evaluation method to interactive human–AI workflows, where clinicians or subject-matter experts verify QA outputs on-the-fly, could further enhance trust and adoption in high-stakes domains, including medicine, healthcare, or law.

Despite these promising results, we consider the following limitations of this study. (1) Our error analysis reveals that misclassifications, particularly between source simplifications and elaborative explanations, can lead to retrieval mismatches and unanswerable QA prompts. These issues underscore the need for more nuanced classification and question generation modules. Although we provide substitute solutions for classifying input summaries or sentences, our fine-tuned classifier is limited to the biomedical domain and specifically designed for PLS tasks, fine-tuned through PLAINFACT. We acknowledge that the dataset used for model fine-tuning is insufficient, resulting in only moderate classification accuracy. Given the scarcity of high-quality human-annotated data for classifying source simplification and elaborative explanation information, future efforts are needed on developing domain-specific (e.g., law, healthcare, social science, etc.) classifiers [55] for factual consistency evaluation. (2) While our experiments indicate that our retrieval-augmented evaluation metric can improve factual consistency assessment in most of the settings, especially in elaborative explanation evaluation, the computational time increases compared to NLI-based evaluation metrics and LLM-based evaluators. The trade-off between evaluation precision and efficiency suggests that further optimization will be beneficial. We suggest to balance wisely based on the trade-offs we report in this study regarding the evaluation time and accuracy.

Acknowledgment

This work used Delta GPU at NCSA through allocation [CIS240504] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Appendix A. Dataset Annotation Protocol

We developed a comprehensive annotation procedure for freelancers on Upwork to conduct fact-checking annotations.

Our annotation procedure involves two stages, starting with thorough training using detailed examples to ensure consistent understanding of the task between annotators. Annotators receive a spreadsheet where each row contains a pair of data: a sentence extracted from the plain language summary and its corresponding scientific abstract. For every sentence-abstract pair, the annotators are required to label three features: External, Category, and Relation, with the appropriate labels.

Step 1: Sentence Annotation: Compared to the scientific abstract, analyze each sentence of the plain language summary across three dimensions: external information, category, and relation.

1. **External**: Determine whether the sentence includes information does not present in the scientific abstract.

   1) `Yes`: The sentence contains external information that is not explicitly mentioned, paraphrased, or implied in the scientific abstract.
   2) `No`: The sentence contains information that is explicitly stated or closely paraphrased from the scientific abstract.

2. **Category**: Classify the sentence of the plain language summary into one of the following categories (you can only choose one category per annotation)

   1) `Definition`: Provides a fundamental explanation of a term.
   2) `Background`: Information that helps understand the term within the context of the abstract, such as relevance, significance, or motivation.
   3) `Example`: Specific instances that illustrate the use of the term in the scientific abstract.
   4) `Method/result`: Details about the methodology or results described in the scientific abstract.
   5) `Other`: For sentences that do not fit into the categories above, please indicate the category

3. **Relation**: Identify the sentence(s) in the scientific abstract that the sentence of the plain language summary is related to. Use indices of sentences from the scientific abstract to link the sentence of the plain language summary. You can select one or more sentences from the scientific abstract. List the indices like s1_1,s2_3,s3_6. If no relation is found, mark it as "external."

In Step 2, we provide explanations of the existing sentence-level indexes for plain language sentences and scientific abstracts.

> **Step 2:** Using the Annotation Spreadsheet: You will work within a structured spreadsheet containing the segmented sentences from both the summaries and the scientific abstracts.
>
> 1. **Target_Summary_ID**: A unique identifier for each plain language summary. Sentences from the same plain language summary share the same ID.
>
> 2. **Target_Sentence_Index**: An identifier for each sentence within a summary, forming as tx_y, where 'tx' is the same as its 'Target_Summary_ID', and 'y' represents the index of each sentence in the plain language summary, starting from 1. e.g., t0_1 refers to the first sentence of the first summary)
>
> 3. **Target_Sentence**: The plain language sentence you are annotating.
>
> 4. **Original_Abstract**: The abstract corresponding to each summary, with each sentence indexed for easy reference.

Each annotator will annotate 40 summary-abstract pairs to ensure each sentence of the plain language summary has two sets of annotations from different people. For each row of the spreadsheet, they need to annotate three columns: "External," "Category," and "Relation."

> <TO BE ANNOTATED> **External**: Mark "Yes" or "No" to indicate if the sentence in the 'Target_Sentence' column contains external information.
>
> <TO BE ANNOTATED> **Category**: Choose the most fitting category of the sentence from the list (Definition, Background, Example, Method/Result, Other).
>
> <TO BE ANNOTATED> **Relation**: List the relevant sentence indices from the abstract in the 'Original_Abstract' column that relate to the plain language sentence (e.g., s10_1,s10_5). For example, filling in 's10_1,s10_5,s10_9' if you think these three sentences from the abstract are relevant to sentence t1_3. Use commas to separate multiple indices.

To ensure annotators fully understand the context and task requirements, we provide comprehensive annotation training and a screening test prior to the annotation process. We select candidates from freelancers with a medical education background, and only those who pass the screening test are finalized as annotators.

	Feature	Annotation
Example 1	Plain Language Sentence	Gout caused by crystal formation in the joints due to high uric acid levels in the blood.
	Need External Information?	yes
	Category	Background
	Relation	external
	Corresponding Abstract	None
Example 2	Plain Language Sentence	Reducing blood pressure with drugs has been a strategy used in patients suffering from an acute event in the heart or in the brain, such as heart attack or stroke.
	Need External Information?	yes
	Category	Background
	Relation	s10_1,s10_2
	Corresponding Abstract	<s10_1>Acute cardiovascular events represent a therapeutic challenge. <s10_2>Blood pressure lowering drugs are commonly used and recommended in the early phase of these settings.
Example 3	Plain Language Sentence	We looked at whether choice of antibiotic made a difference in the number of people who experienced failed treatment, and we determined the proportions who had resolution of fever at 48 hours.
	Need External Information?	no
	Category	Method/Result
	Relation	s15_16,s15_17,s15_20
	Corresponding Abstract	<s15_16>For treatment failure, the difference between doxycycline and tetracycline is uncertain (very low-certainty evidence). <s15_17>Doxycycline compared to tetracycline may make little or no difference in resolution of fever within 48 hours (risk ratio (RR) 1.14, 95% confidence interval (CI) 0.90 to 1.44, 55 participants; one trial; low-certainty evidence) and in time to defervescence (116 participants; one trial; low-certainty evidence). <s15_20>For most outcomes, including treatment failure, resolution of fever within 48 hours, time to defervescence, and serious adverse events, we are uncertain whether study results show a difference between doxycycline and macrolides (very low-certainty evidence).

Table A.6: Examples of our curated dataset. Need External Information feature represents whether a plain language sentence is a simplification or an explanation. A label of "yes" indicates that the sentence is an explanation and requires additional elaborative information beyond the source abstract to verify its factual consistency. Conversely, a label of "no" shows that the sentence is a simplification that can be validated using only the source abstract.

Appendix B. Dataset Examples

In Table Appendix A, we presents three representative examples from PlainFact. Each example is annotated with five features: a plain language sentence, an indicator of whether the sentence is simplification or explanation, its category, its relation, and the corresponding abstract. All plain language sentences are **factual**, as they were written by the authors from the Cochrane database. "Need External Information?" feature specifies whether a sentence can be validated solely by the abstract. A "Yes" label indicates that the sentence includes information not explicitly mentioned in the abstract, and vice versa. The "Relation" feature identifies the sentence(s) in the scientific abstract most relevant to the plain language summaries; if no corresponding content exists in the abstract, the relation is marked as "external." We randomly sample three examples from the dataset to to illustrate the dataset's structure. Additionally, indexes have been created for both the plain language sentences and the abstract sentences to facilitate annotation.

Appendix C. LLM Prompts

We utilize two types of prompts to guide LLMs through two stages of FC evaluation: classification and answer extraction. In accordance with the benchmark annotation protocol (Appendix A), we employ `GPT-4o` as a classifier to determine whether a given sentence or summary requires elaborative explanations for factual verification. For both stages, we set the `max_tokens` parameter to 512 and configure the temperature to 0 for `GPT-4o` and 0.01 for the `Llama 3.1 8B Instruc` model.

```
Developer
Annotate whether a sentence or summary includes information not present in the original
abstract.

The sentence or summary contains external information that is not explicitly mentioned,
paraphrased, or implied in the original abstract will be labeled as 'Yes'.

The sentence or summary contains information that is explicitly stated or closely
paraphrased from the original abstract will be labeled as 'No'.

User
Sentence or summary: <input>
Original abstract: <abstract>
```

Example 1: Prompt of `GPT-4o` as the Classifier.

For the AE stage, we explore both `GPT-4o` and `Llama 3.1 8B Instruct` as backbone models to assess the FC evaluation performance of PLAINQAFACT. Following the task description outlined in QAFactEval [39], we instruct both LLMs to extract potential answer entities from the input PLS.

```
Developer
QA-based metrics compare information units between the summary and source, so it is thus
necessary to first extract such units, or answers, from the given summary. Please extract
answers or information units from a plain language summary.

User
Extract a comma-separated list of the most important keywords from the following text:
<input>
```

Example 2: Prompt of `GPT-4o` as the Answer Extractor.

```
System
QA-based metrics compare information units between the summary and source, so it is thus
necessary to first extract such units, or answers, from the given summary. Please extract
answers or information units from a plain language summary.

User
Extract a comma-separated list of the most important keywords from the following text:
<input>
```

Example 3: Prompt of `Llama 3.1 8B Instruct` as the Answer Extractor.

Additionally, we report the performance of using `Llama 3.1 8B Instruct` as a judge to evaluate the FC of a given PLS based on its source scientific abstract.

```
System
Rate the factuality of the given plain language sentence or summary compared with the
scientific abstract. Output a numeric score from 0 to 100, with 100 meaning the sentence is
 completely factually consistent with the abstract and 0 meaning the sentence is completely
 non-factual with the abstract.

User
Sentence or summary: <input>
Original abstract: <abstract>
Factuality score (only output a numeric score): <score>
```

Example 4: Prompt of `Llama 3.1 8B Instruct` as a FC judge.

For the LLM-based perturbation of PLAINFACT and CELLS datasets, we follow the protocol of APPLS [12] on faithfulness criteria.

```
System
You are a data transformation assistant. You will receive a sentence from a biomedical
literature. You will generate a new version of the given sentence based on the following
rules for faithfulness perturbations:
```

```
1. Number Swap
Locate any numeric value(s) in the sentence and swap them with different numeric value(s).
Example: "infected more than 59 million people" -> "infected more than 64 million people"

2. Entity Swap
Locate a key entity (e.g., virus name, drug name, organization) in the sentence and swap it
 with a different entity.
Example: "coronavirus 2 (SARS-CoV-2)" -> "canine adenovirus (CAV-2)"

3. Synonym Verb Swap
Identify a key verb in the sentence and replace it with a near-synonym or related verb that
 changes the nuance or meaning slightly.
Example: "killed more than one of them" -> "stamped out more than one of them"

4. Hypernym/Antonym Swap
Select a word and replace it with either a hypernym (a more general term) or an antonym (
opposite meaning), as appropriate.
Example (antonym): "killed more than one of them" -> "saved more than one of them"
Example (hypernym): "dog" -> "animal" (if relevant)

5. Negation
Negate a key part of the sentence to flip its meaning.
Example: "has infected more than 59 million people" -> "hasn't infected more than 59
million people" or "has not infected more than 59 million people"

Your task:
Read each sentence, generate a new sentence based on one of the five types of perturbation
stratigies (Number Swap, Entity Swap, Synonym Verb Swap, Hypernym/Antonym Swap, Negation)
above.
Return the perturbation sentence.

Do not change the rest parts of the sentence except for the perturbation content. For
example, the original sentence is "The skin patch and the vaginal (birth canal) ring are
two methods of birth control." The perturbation sentence should be "The skin patch and the
vaginal (birth canal) ring are five methods of birth control."

User
Sentence: <input>
Perturbation sentence: <sentence>
```

Example 5: Prompt of GPT-4o for faithfulness perturbation

Appendix D. Detailed Experiment Settings

All experiments we conduct are under one NVIDIA A100 GPU with 40 GB GPU memory. We employ the Natural Language Toolkit (NLTK)[5] to split the plain language summaries into sentences using the punkt package. For the classifier fine-tuning, we set the random seed to 42 to split PlainFact into training, validation, and test sets. We tune the PubMedBERT-base

[5] https://www.nltk.org/

model for three epochs with early stopping under the validation set. The final fine-tuning accuracy of the classifier on the test set is 0.77. The same seed (42) is also used to sample 200 summary-abstract pairs from the CELLS test set for the comparison in Table 5.

The QA model used in our PlainQAFact is downloaded from QAFactEval [39]. However, we set its maximum input length to 512 (from 364) tokens to incorporate more source context. Similarly, to ensure that the answers extracted during the AE stage are valid (i.e., within the 512-token limit to maintain consistency with the subsequent QA model), we configure the LLMs' input lengths to 512 tokens. In MedCPT retrieval, we set `k` (the number of retrieved snippets) to 3, ensuring the retrieved information remains within a short context window for the subsequent QA process (Appendix D).

We apply the default setting (i.e., `RoBERTa-base`) for AlignScore [19] and SummaC-Conv for SummaC [16]. For `GPT-4o` experiments as an factual consistency evaluator, we set the temperature as 1.0 in this paper.

Appendix E. Statistical Testing

Following SummaC [16] and QAFactEval [39], we test whether our proposed metric achieves statistically significant improvements over other methods. Given Kendall's τ and Pearson correlations do not show the discriminative feature of continuous scores in various evaluation metric, we run a systematic evaluation on four datasets (CELLS, FareBio, FactPICO, and PlainFact) through bootstrap resampling [56] on the AUC-ROC results. We compare our best metric to other methods using the confidence intervals as significance level of 0.05 and apply Bonferroni correction [57] following SummaC [16]. Our results show that PlainQAFact achieves statistically significant improvements over QuestEval [50], SummaC [16], and DAE [48] on the CELLS and PlainFact datasets. As shown in Figure 2, PlainQAFact also outperforms other automatic metrics in evaluating elaborative explanation sentences, except for QAFactEval [39]. This highlights the effectiveness of PlainQAFact for biomedical PLS tasks that include external explanations. On the other hand, when evaluating elaborative content in the FactPICO and FareBIO datasets (Figures E.3 and E.4), we do not find statistically significant differences across metrics. While some existing metrics perform similarly in certain settings, our metric consistently shows better performance across all three evaluation criteria.

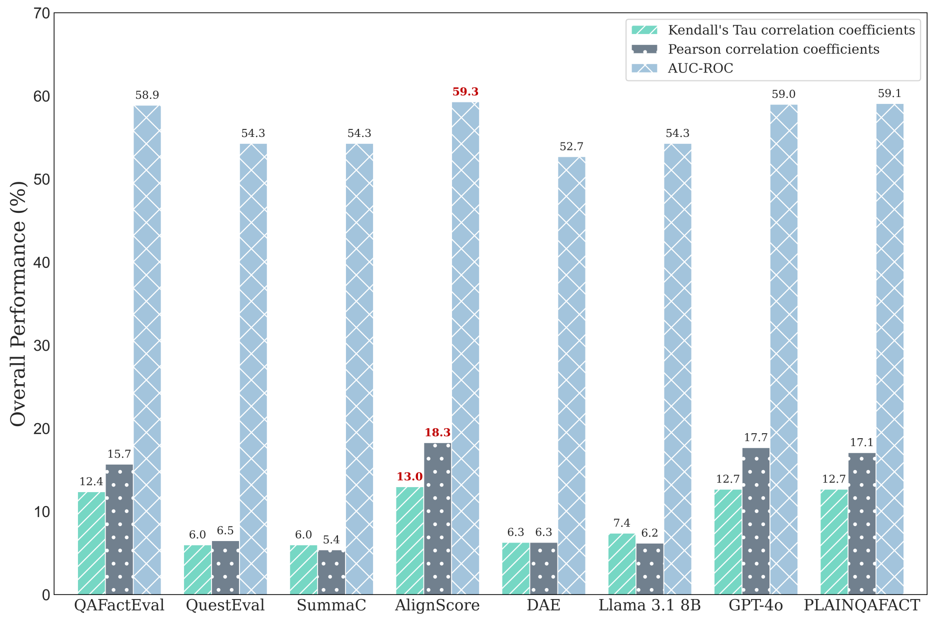

Figure E.3: Overall performance on summaries containing added information (i.e., elaborative explanations) from FactPICO [9]. The std. of PlainQAFact, `Llama 3.1`, and `GPT-4o` are 0.2, 0.2, and 3.3, respectively, based on five runs of each metric (details in Appendix E).

Appendix F. Summary-Level Explanation-Only Evaluation

Similar to our sentence-level explanation annotations in PlainFact, FactPICO [9] and FareBio [47] datasets also provide human-annotated explanation information in plain language summaries. According to the results shown in Section 5.2, we also assess the performance of five factual consistency evaluation metrics under FactPICO and FareBio datasets.

In the FactPICO, the elaborative explanation is defined as "added information." We first select summaries that include added information, and then we label a summary as "non-factual" if it contains any non-factual added information as determined by annotators. We generate a summary-level explanation-only FactPICO dataset consisting of 231 summary-abstract pairs. As shown in Figure E.3, AlignScore achieves the best performance across all three evaluation criteria. However, since the summaries in the FactPICO dataset are generated using LLMs, the factual consistency between the generated summaries and the original source abstracts is not guaranteed. For

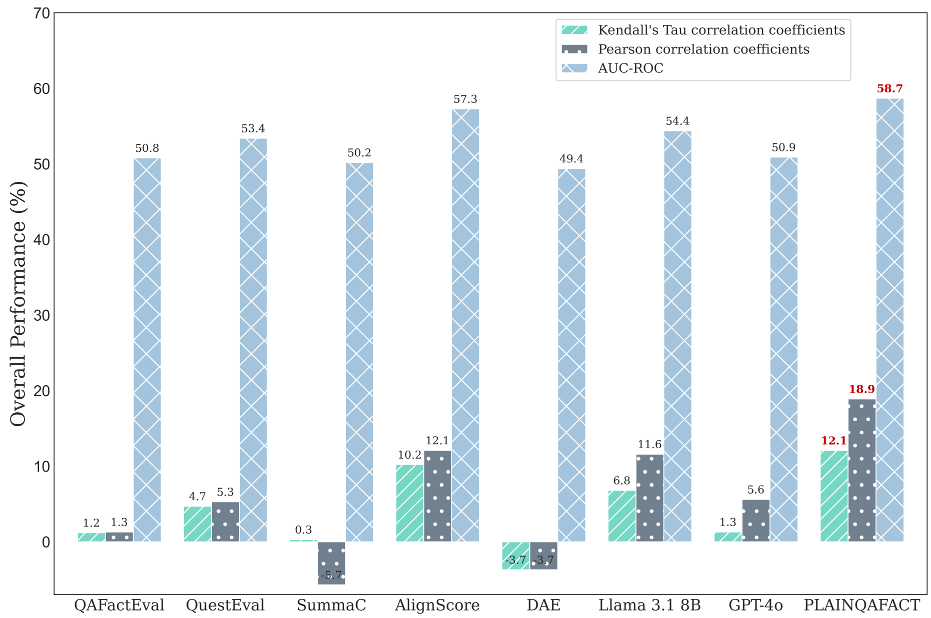

Figure E.4: Overall performance on summaries with elaborative explanations from FareBio [47]. The std. for PLAINQAFACT, `Llama 3.1`, and `GPT-4o` are 0.03, 0.3, and 5.7, respectively, computed over five runs (details in Appendix E.)

instance, we treat a summary as "factual" indicates that all the added information is factual, but it does not necessarily reflect the factual consistency of other content. Therefore, the results presented in Figure E.3 may become biased.

For FareBio [47], we collect explanation sentences based on both "faithfulness" and "factual hallucination" labels in the original dataset. As shown in Figure E.4, PLAINQAFACT achieves the best performance compared with all other metrics, showing a clear advantage over `GPT-4o`. Since the FareBio dataset provides sentence-level annotations of factual consistency for each sentence in the plain language summaries, the results are more reliable than those from FactPICO.

Appendix G. Pilot Study on FactPICO

To investigate the performance of existing automatic factual consistency evaluation metrics on plain language generation tasks, we employ the recently introduced FactPICO dataset [9]. This dataset comprises human-labeled plain language summaries of Randomized Controlled Trials (RCTs) that address several critical elements: Populations, Interventions, Comparators, Outcomes (PICO), and additional information. All summaries are generated by various LLMs based on medical literature and include added information (i.e., extensive explanations) not present in the original abstracts. We hypothesize that existing factuality evaluation metrics for text summarization may struggle to accurately assess the factuality of this added information. To validate this assumption, we conduct pilot studies using four factual consistency evaluation metrics from FactPICO alongside one NLI-based metric: DAE [48], AlignScore [19], SummaC [16], QAFactEval [39], and QuestEval [50].

Appendix G.1. Dataset Pre-processing

FactPICO provides span-level annotations for LLM-generated summaries, assessing whether the additional information is present and determining its factuality, labeled as either "yes" or "no." We first remove all special identification tags from the abstracts, such as "ABSTRACT.BACKGROUND," "ABSTRACT.RESULTS," and "ABSTRACT.CONCLUSIONS." We then deduplicate the generated summaries, while retaining duplicates within abstracts, as each abstract has three summaries generated by different LLMs, and each abstract may have varying numbers of generated summaries, from none to multiple. This pilot study focuses on two key research questions: evaluating the effectiveness of existing factuality metrics in assessing added information and non-factual added information.

Appendix G.2. Experiments and Results

We conduct experiments using the processed FactPICO dataset to conduct pilot studies that assess the performance of existing factual consistency evaluation metrics in detecting added information in plain language summarization tasks. It is important to note that some outliers in the plain language summaries of the FactPICO dataset contain extraneous content, which causes the input lengths to exceed the limitations of both the DAE and QAFactEval metrics. Furthermore, the factuality of the LLM-generated plain language summaries is not guaranteed.

RQ1: Do existing metrics perform well when external information is added to plain language summaries compared with no added information?

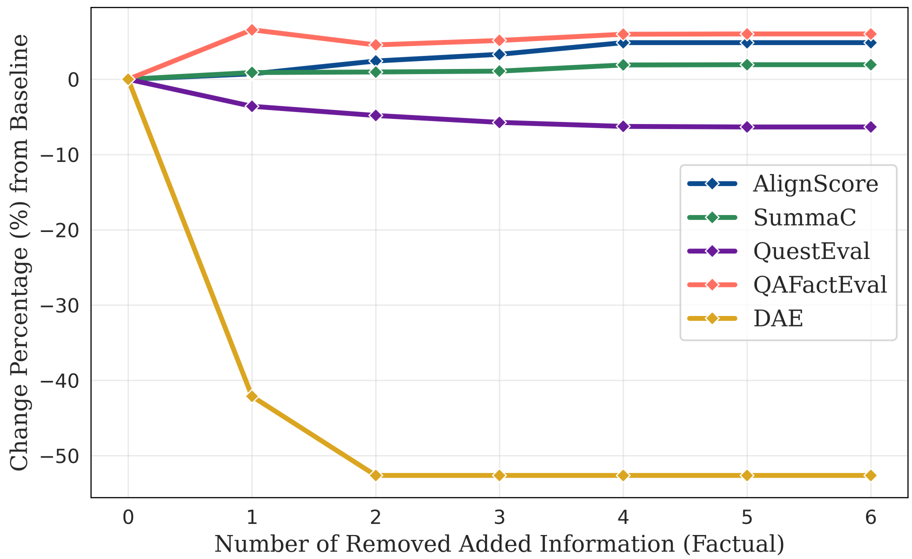

Figure G.5: Score change percentage from baselines over five metrics on the FactPICO dataset in removing factual added information. We expect each metric stays unchanged even when more added factual information is removed. The evaluation dataset contains 88 valid summary-abstract pairs.

In this study, we evaluate the ability of existing factual consistency evaluation metrics to detect added information in plain language summaries. We focus on summaries where all added information is annotated as factual, resulting in a dataset of 88 summary-abstract pairs. To assess metric sensitivity to the added information, we iteratively remove sentences from each plain language summary that contain added spans. For example, if a span such as “of a medicine called haloperidol” is labeled as factual (“yes”), we remove the entire sentence in the original plain language summaries containing that span through exact matching, continuing this process until no added information remains. This procedure enables us to determine how effectively current metrics handle added information that is absent from the original abstracts. As shown in Figure G.5, we report performance changes relative to baseline scores. For example, the AlignScore evaluation increases by 3.3% (on a 100-point scale) when six spans of added information are removed, resulting in an overall change of approximately 4.9% compared to the setting in which no added information is removed. These findings indicate that added information affects model-based factuality evaluation metrics such as

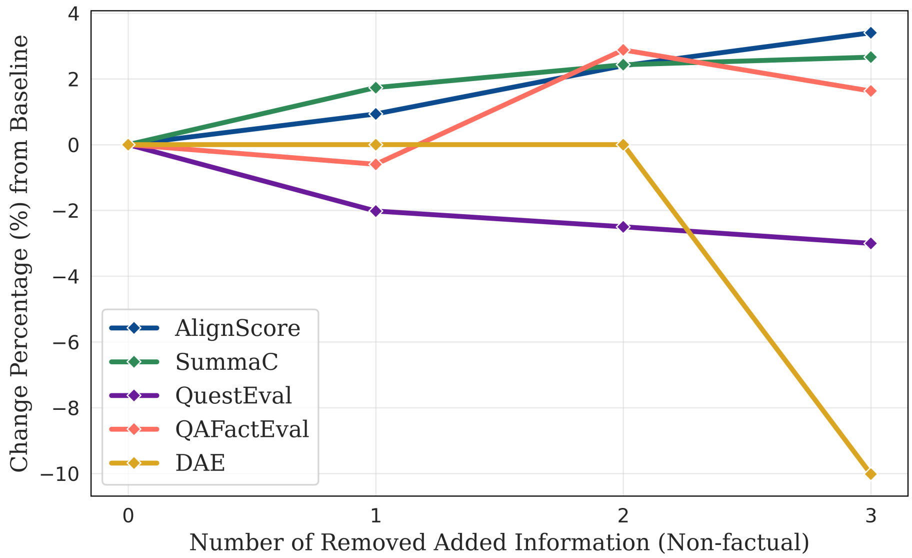

Figure G.6: Score change percentage from baselines over five metrics on the FactPICO dataset in removing non-factual added information (60 pairs). We expect the change percentage from baseline increases when more added non-factual information is removed.

QAFactEval (6.0%), AlignScore (4.9%), and SummaC (1.9%), with scores improving as more added information is removed. In contrast, both QuestEval and DAE scores decline with the removal of added information, and notably, DAE exhibits the most significant performance drop. Overall, these findings suggest that all the five metrics have difficulty accurately assessing added factual information, as evidenced by both increases and decreases in their performance.

RQ2: Can existing metrics distinguish between non-factual and factual added information?

In this research question, our goal is to evaluate the sensitivity of factuality evaluation metrics in detecting non-factual information within plain language summaries. The FactPICO dataset labels added information as either “yes” (factual) or “no” (non-factual). In this scenario, we sample only those plain language summaries that contain both “yes” and “no” labels for added spans. As with RQ1, we iteratively remove sentences containing non-factual added spans from each plain language summary until no non-factual sentences remain. Overall, as illustrated in Figure G.6, only AlignScore, SummaC, and QAFactEval show improved performance as more non-factual information is removed, indicating that these metrics are sensitive in added non-factual information. Nevertheless, based on the results of RQ 1 and 2, our findings suggest that existing factual consistency evaluation metrics have limited capacity to accurately distinguish between factual and non-factual added information in plain language summarization tasks.

References

[1] Stoll, M., Kerwer, M., Lieb, K., Chasiotis, A.: Plain language summaries: A systematic review of theory, guidelines and empirical research. Plos one **17**(6), 0268789 (2022)

[2] Kuehne, L.M., Olden, J.D.: Lay summaries needed to enhance science communication. Proceedings of the National Academy of Sciences **112**(12), 3585–3586 (2015)

[3] Ondov, B., Attal, K., Demner-Fushman, D.: A survey of automated methods for biomedical text simplification. Journal of the American Medical Informatics Association 1976–1988 (2022)

[4] Guo, Y., Chang, J.C., Antoniak, M., Bransom, E., Cohen, T., Wang, L.L., August, T.: Personalized jargon identification for enhanced interdisciplinary communication. Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting 4535–4550 (2023)

[5] Goldsack, T., Zhang, Z., Lin, C., Scarton, C.: Domain-driven and discourse-guided scientific summarisation. In: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, pp. 361–376 (2023). Springer

[6] You, Z., Radhakrishna, S., Ming, S., Kilicoglu, H.: UIUC_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In: Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., Tsujii, J. (eds.) Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pp. 132–143. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.bionlp-1.11 . https://aclanthology.org/2024.bionlp-1.11/

[7] Guo, Y., Qiu, W., Wang, Y., Cohen, T.: Automated lay language summarization of biomedical scientific reviews. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 160–168 (2021)

[8] Srikanth, N., Li, J.J.: Elaborative simplification: Content addition and explanation generation in text simplification. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 5123–5137 (2021)

[9] Joseph, S., Chen, L., Trienes, J., Göke, H., Coers, M., Xu, W., Wallace, B., Li, J.J.: FactPICO: Factuality evaluation for plain language summarization of medical evidence. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8437–8464. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.459 . https://aclanthology.org/2024.acl-long.459

[10] Jain, D., Borah, M.D., Biswas, A.: Summarization of legal documents: Where are we now and the way forward. Computer Science Review **40**, 100388 (2021)

[11] Hardy, H., Narayan, S., Vlachos, A.: Highres: Highlight-based reference-less evaluation of summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3381–3392 (2019)

[12] Guo, Y., August, T., Leroy, G., Cohen, T., Wang, L.L.: APPLS: Evaluating evaluation metrics for plain language summarization. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 9194–9211. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.519 . https://aclanthology.org/2024.emnlp-main.519/

[13] Luo, Z., Xie, Q., Ananiadou, S.: Chatgpt as a factual inconsistency evaluator for abstractive text summarization. arXiv preprint arXiv:2303.15621 (2023)

[14] You, Z., Lee, H., Mishra, S., Jeoung, S., Mishra, A., Kim, J., Diesner, J.: Beyond binary gender labels: Revealing gender bias in LLMs through gender-neutral name predictions. In: Faleńska, A., Basta, C., Costa-jussà, M., Goldfarb-Tarrant, S., Nozza, D. (eds.) Proceedings

of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pp. 255–268. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.gebnlp-1.16 . https://aclanthology.org/2024.gebnlp-1.16/

[15] Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P.N., Shoeybi, M., Catanzaro, B.: Factuality enhanced language models for open-ended text generation. Advances in Neural Information Processing Systems **35**, 34586–34599 (2022)

[16] Laban, P., Schnabel, T., Bennett, P., Hearst, M.A.: Summac: Revisiting nli-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics **10**, 163–177 (2022)

[17] Wan, D., Bansal, M.: Evaluating and improving factuality in multimodal abstractive summarization. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 9632–9648 (2022)

[18] Ye, Y., Simpson, E., Rodriguez, R.S.: Using similarity to evaluate factual consistency in summaries. arXiv preprint arXiv:2409.15090 (2024)

[19] Zha, Y., Yang, Y., Li, R., Hu, Z.: 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), pp. 11328–11348 (2023)

[20] Deutsch, D., Bedrax-Weiss, T., Roth, D.: 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 (2021)

[21] Fabbri, A.R., Wu, C.-S., Liu, W., Xiong, C.: 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, pp. 2587–2601 (2022)

[22] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)

[23] Ren, R., Wang, Y., Qu, Y., Zhao, W.X., Liu, J., Tian, H., Wu, H., Wen, J.-r., Wang, H.: Investigating the factual knowledge boundary of large language models with retrieval augmentation. In: International Conference on Computational Linguistics (2023). https://api.semanticscholar.org/CorpusID:259991467

[24] Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., Chen, W.: Generation-augmented retrieval for open-domain question answering. In: Annual Meeting of the Association for Computational Linguistics (2020). https://api.semanticscholar.org/CorpusID:221802772

[25] Kim, S.: Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation. (2025). https://api.semanticscholar.org/CorpusID:276116319

[26] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. ArXiv **abs/2312.10997** (2023)

[27] Tang, Y., Yang, Y.: Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. ArXiv **abs/2401.15391** (2024)

[28] Guo, Y., Qiu, W., Leroy, G., Wang, S., Cohen, T.: Retrieval augmentation of large language models for lay language generation. Journal of Biomedical Informatics **149**, 104580 (2024)

[29] Murad, M.H., Asi, N., Alsawas, M., Alahdab, F.: New evidence pyramid. BMJ Evidence-Based Medicine **21**(4), 125–127 (2016)

[30] Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)

[31] Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin, 37–54 (1948)

[32] Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)

[33] Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational linguistics **34**(4), 555–596 (2008)

[34] OpenAI: GPT-4o. `https://platform.openai.com/docs/models#gpt-4o`. Accessed: 2025-01-20 (2024)

[35] Xu, Z., Cruz, M.J., Guevara, M., Wang, T., Deshpande, M., Wang, X., Li, Z.: Retrieval-augmented generation with knowledge graphs for customer service question answering. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2905–2909 (2024)

[36] Jin, Q., Kim, W., Chen, Q., Comeau, D.C., Yeganova, L., Wilbur, W.J., Lu, Z.: Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics **39**(11), 651 (2023)

[37] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)

[38] Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024, pp. 6233–6251. Association for Computational Linguistics, Bangkok, Thailand (2024). `https://doi.org/10.18653/v1/2024.findings-acl.372` . https://aclanthology.org/2024.findings-acl.372/

[39] Fabbri, A., Wu, C.-S., Liu, W., Xiong, C.: QAFactEval: Improved QA-based factual consistency evaluation for summarization. In: Carpuat, M., Marneffe, M.-C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2587–2601. Association for Computational Linguistics, Seattle, United States (2022). `https://doi.org/10.18653/v1/2022.naacl-main.187` . https://aclanthology.org/2022.naacl-main.187

[40] Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Bai, X.: Promptrank: Unsupervised keyphrase extraction using prompt. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9788–9801 (2023)

[41] Meta: Llama 3.1: Open Foundation and Instruction-Tuned Models (2024). https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[42] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/D16-1264 . https://aclanthology.org/D16-1264

[43] Demszky, D., Guu, K., Liang, P.: Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922 (2018)

[44] Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pretraining text encoders as discriminators rather than generators. In: International Conference on Learning Representations (2020)

[45] Chen, A., Stanovsky, G., Singh, S., Gardner, M.: MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6521–6532. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.528 . https://aclanthology.org/2020.emnlp-main.528

[46] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations

[47] Fang, B., Dai, X., Karimi, S.: Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 9890–9911 (2024)

[48] Goyal, T., Durrett, G.: 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, pp. 1449–1462 (2021)

[49] Liu, Y.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 **364** (2019)

[50] Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., Gallinari, P.: QuestEval: Summarization asks for fact-based evaluation. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6594–6604. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.529 . https://aclanthology.org/2021.emnlp-main.529

[51] Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition **30**(7), 1145–1159 (1997)

[52] Philipsborn, P., Stratil, J.M., Burns, J., Busert, L.K., Pfadenhauer, L.M., Polus, S., Holzapfel, C., Hauner, H., Rehfuess, E.: Environmental interventions to reduce the consumption of sugar-sweetened beverages and their effects on health. Cochrane Database of Systematic Reviews (6) (2019) https://doi.org/10.1002/14651858.CD012292.pub2

[53] Lopez, L.M., Grimes, D.A., Gallo, M.F., Stockton, L.L., Schulz, K.F.: Skin patch and vaginal ring versus combined oral contraceptives for contraception. Cochrane Database of Systematic Reviews (3) (2010) https://doi.org/10.1002/14651858.CD003552.pub3

[54] Luo, Z., Xie, Q., Ananiadou, S.: Factual consistency evaluation of summarization in the era of large language models. Expert Systems with Applications **254**, 124456 (2024)

[55] You, Z., Han, K., Zhu, H., Ludaescher, B., Diesner, J.: SciPrompt: Knowledge-augmented prompting for fine-grained categorization of scientific topics. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Lan-

guage Processing, pp. 6087–6104. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.350 . https://aclanthology.org/2024.emnlp-main.350/

[56] Efron, B.: The jackknife, the bootstrap and other resampling plans. In: CBMS-NSF Regional Conference Series in Applied Mathematics (1982)

[57] Bonferroni, C.E.: Il calcolo delle assicurazioni su gruppi di teste. Studi in onore del professore salvatore ortu carboni, 13–60 (1935)