

Understanding Faithfulness and Reasoning of Large Language Models on Plain Biomedical Summaries

Anonymous ACL submission

Abstract

Plain biomedical summaries generation with Large Language Models (LLMs) can enhance the accessibility of biomedical knowledge to the public. However, how faithful the generated summaries are remains an open yet critical question. To address this, we propose FAREBIO, a benchmark dataset with expert-annotated **Faithfulness** and **Reasoning** on plain **Biomedical Summaries**. This dataset consists of 175 plain summaries, including 1445 sentences generated by 7 different LLMs, paired with PubMed articles. Based on our dataset, we identify the performance gap of LLMs in generating faithful plain biomedical summaries and show the impact of abstractiveness on faithfulness. We show that current faithfulness metrics do not transfer well in the biomedical domain. To better understand the faithfulness judgements, we further benchmark LLMs in retrieving supporting evidence. Going beyond the binary faithfulness labels, coupled with the annotation of supporting sentences, our dataset could further contribute to the understanding of faithfulness evaluation and reasoning.

1 Introduction

Generating plain text summaries—summarizing technical articles in plain language—helps facilitate public access to biomedical knowledge and has been an important topic in the biomedical domain (Goldsack et al., 2022, 2023; Guo et al., 2021). Despite the overall promising performance achieved by LLMs (Jahan et al., 2024; Guo et al., 2024; Sim et al., 2023), the faithfulness of the generated summary, i.e., to what extent the generated text is consistent with the source articles, has been a known problem of LLMs (Pagnoni et al., 2021; Ji et al., 2023) and it has not been well-studied in the biomedical domain (Joseph et al., 2024).

Apart from labor-intensive and costly manual examination, prior work (Scialom et al., 2021; Laban et al., 2022; Zha et al., 2023) has proposed

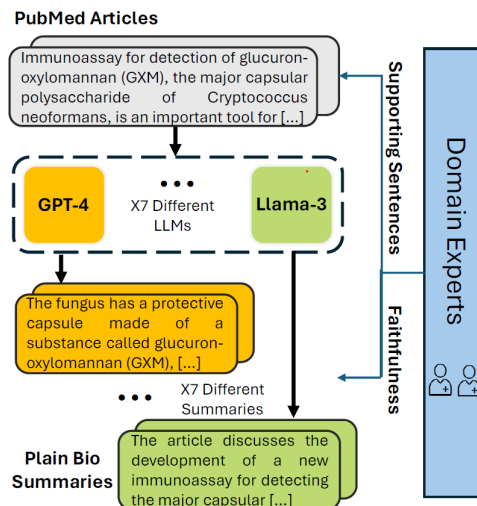


Figure 1: Faithfulness and reasoning annotations on plain biomedical summaries generated from 7 LLMs.

various metrics to automatically evaluate the faithfulness of generated text. However, these metrics are designed to measure faithfulness in the general domain, e.g., news. To what extent it can be used in domain-specific areas, e.g., the biomedical domain, remains an open question (Ramprasad et al., 2024).

Additionally, current research (Chiang and Lee, 2023b) has shown that, although LLM-based evaluators achieve promising alignment with human judgment, they do not always provide correct reasoning for their decisions. Examining to what extent LLMs can provide correct reasoning for their choices could help better understand the reasoning behind LLMs, especially in the biomedical domain where it relies on accurate evidence.

To address these problems, we propose a benchmark dataset, FAREBIO, on evaluating the **Faithfulness** and **Reasoning** of LLMs on plain **Biomedical** summaries in Section 3. Specifically, as shown in Figure 1, we enlist medical doctors to manually evaluate the faithfulness of plain summaries from seven representative zero-shot summarization systems and highlight the corresponding

supporting sentences from the source articles.

In Section 4, we answer the four major research questions on faithfulness evaluation and LLMs on plain biomedical summaries: (1) How faithful are generated summaries across current LLMs? (2) How abstractive and readable are plain biomedical summaries, and how do they relate to faithfulness? (3) How do current faithfulness evaluators align with human judgment? (4) Do LLMs consider their generation more faithful than others?

We further evaluate the LLMs’ capability of providing correct reasoning for their judgment in Section 5. Specifically, we address the three major research questions: (5) Can LLMs identify the supporting sentences from the source article? (6) Does the abstractiveness of the summary impact the identification of supporting sentences? (7) Do LLMs perform better when identifying supporting sentences for their own generated summaries?

To the best of our knowledge, our study is the first publicly available benchmark dataset investigating faithfulness and identification of supporting sentences for plain biomedical summaries.¹ We find that the generated summaries from current LLMs demonstrate a high degree of hallucination and the level of abstractiveness shows a positive impact on the faithfulness of summaries. Current faithfulness evaluators trained in the general domain do not directly transfer well to the biomedical domain. We also observe a tendency where LLMs, as evaluators, favor their generation when evaluating faithfulness. However, the construction of the prompt could also impact such a tendency. Additionally, LLMs show the potential to identify the supporting sentences from the source articles, either with high or low abstractiveness summaries.

2 Related Work

Faithfulness, where the generated text is factually consistent with the source (Maynez et al., 2020), is a known challenge in text generation (Ji et al., 2023; Huang et al., 2023). Current faithfulness research on LLMs mostly focuses on the general domain, with a particular interest in news articles (Pagnoni et al., 2021; Fabbri et al., 2021; Tang et al., 2023; Cao and Wang, 2021). Some studies evaluate faithfulness and factuality—factual consistency with enteral knowledge—in the biomedical domain. For instance, Ramprasad et al. (2024) measured the

¹Our dataset will be publicly available at [link withhold for anonymous submission].

factuality of zero-shot summaries from GPT-3.5 (Brown et al., 2020) and Flan-T5-XL (Chung et al., 2022). FACTPICO (Joseph et al., 2024) was proposed to measure the factuality of GPT-4 (Achiam et al., 2023), Llama-2-Chat (Touvron et al., 2023), and Alpaca (Taori et al., 2023) under the PICO framework (Lehman et al., 2019).

Current research has proposed various metrics based on different frameworks to evaluate the faithfulness of generated text for the general domain: (1) QA-based metrics (Scialom et al., 2021; Fabbri et al., 2022; Durmus et al., 2020), utilizing QA systems to measure the correctness of answering the questions based on the source and summaries, as a proxy of faithfulness; (2) NLI-based metrics (Laban et al., 2022; Falke et al., 2019), measuring the entailment of the summary (hypothesis) from the source (premise) by employing models that are trained on NLI datasets (Kryscinski et al., 2020; Nie et al., 2020); (3) Faithfulness classification-based (Zha et al., 2023; Zhou et al., 2021), training evaluators to directly predict faithfulness; and (4) LLM-based metrics (Min et al., 2023; Sottana et al., 2023; Chiang and Lee, 2023b), prompting LLMs as the faithfulness evaluator.

Apart from solely evaluating the binary faithfulness label of the generated summary, a natural question to ask is to provide the reasoning for the judgment, e.g., supporting sentences from the source. For faithfulness reasoning, the FEVER dataset (Thorne et al., 2018) annotated the factuality of the claims based on Wikipedia articles and provided extracted facts from the corresponding sources. Wadden et al. (2020) created SciFact, a dataset of 1.4K expert-written scientific claims paired with the abstracts from S2ORC (Lo et al., 2020), annotating with labels and rationales. Ghosal et al. (2024) proposed a shared task in identifying all grounding context from the scholarly paper discussing methodological details in the claim.² However, it only contains 109 test claims and the dataset is not publicly available.

3 Dataset Creation

3.1 Model Selection

To investigate how faithful current LLMs are in generating plain biomedical summaries, we evaluate the following representative summarization systems across various settings: (1) open-source vs., close-source and (2) pretrained vs., fine-tuned:

²<https://github.com/oasisresearchlab/context24>

GPT-4 (Achiam et al., 2023), a large close-source multimodal model developed by Open AI. The GPT family is adopted in various NLP tasks including summarization (Zhang et al., 2023; Adams et al., 2023a; Shaib et al., 2023). We use `gpt-4-turbo` to generate plain biomedical summaries.

Claude-3 (Anthropic, 2024), a close-source model developed by Anthropic. It has been seen to outperform the GPT family in certain tasks, e.g., open-domain conversation (Lin and Chen, 2023) and reading comprehension test (Kuo et al., 2023). We use `claude-3-sonnet` for our experiment.

Gemini-1.5 (Reid et al., 2024), a close-source model developed by Google DeepMind. It claims the capabilities of understanding complex medical context (Saab et al., 2024). We include `Gemini-1.5-Flash` as a closed-source model.

Llama-3 (Meta, 2024), a open-source model released by Meta. Compared to the previous models, one major difference is that this model is open-sourced and available for both research and commercialization purposes. We consider the newly-released version, `Llama-3-8B-Instruction`, for our experiment.

Flan-T5 (Chung et al., 2022; Longpre et al., 2023), one other popular open-source model released by Google. It is an enhanced version of T5 models (Raffel et al., 2020) and has been used for various summarization tasks (Sim et al., 2023; Alqahtani et al., 2023). we investigate `Flan-T5-XL` for our plain biomedical summarization task.

Finetuned-Llama-3 To investigate the impact of customizing the plain summaries for a specific type, we further fine-tune `Llama-3-8B-Instruction` on the PLOS dataset (Goldsack et al., 2022), a corpus for generating layman summaries based on science and medicinal peer-reviewed journals.³

Finetuned-Flan-T5 Similarly, we fine-tune `Flan-T5-XL` on the PLOS dataset and investigate the faithfulness of the generated summaries from the fine-tuned model.

For selected models, we ask LLMs to generate a plain summary based on the source article provided in Section 3.2, with the input of titles, authors, abstract, and first section of the content.⁴

³Details of fine-tuning `Llama-3-8B-Instruction` and `Flan-T5-XL` on the PLOS dataset in Appendix A.

⁴Detailed prompt constructions are in Appendix B.

	Number
Source Articles	25
Avg. Sentences per Source Article	26.08
Generated Summaries	175
Total Sentences in Generated Summaries	1445
Avg. Sentences per Generated Summary	8.31
GPT-4	8.92
Claude-3	8.32
Gemini-1.5	11.88
Llama-3	7.64
Flan-T5	5.80
Finetuned-Llama-3	7.24
Finetuned-Flan-T5	8.00

Table 1: Statistics of our annotated dataset.

3.2 Annotation Data

To generate plain biomedical summaries, we obtained English PubMed articles from S2ORC (Lo et al., 2020),⁵ an open-source scholarly dataset based on Semantic Scholar containing more than 205M publications across various resources. We randomly selected 25 articles that (1) were published in PubMed; (2) were published no later than 2010; and (3) contained metadata of title, authors, abstract, and full content. As shown in Figure 1, for each article, we then generate 7 different plain summaries based on various types of LLMs (Section 3.1), resulting in 175 summaries.

To provide a more fine-grained level of faithfulness analysis, we tokenize the generated summaries into sentences and ask annotators to annotate the faithfulness of the generated summaries at the sentence level (Section 3.3). Note that due to the imperfection of the off-the-shelf tokenizer tool, sentence tokenization could result in the segment of text, instead of the correct sentence. To address this, we filter out the tokenized sentences that are less than 5 characters, resulting in 1445 sentences. The statistics of our dataset are shown in Table 1.

3.3 Annotation Collection

As discussed in Section 3.2, we annotate the faithfulness at the sentence level. Aligned with the summary generation in Section 3.1, we provide annotators with the article title, author, abstract, and the first section of the content. The annotation includes four parts: (1) annotate whether the summary sentence is faithful given the source article; (2) provide a brief rationale of the annotation choice; (3) if it is faithful, highlight the supporting evidence from the source article; and (4) if it is not

⁵<https://api.semanticscholar.org/api-docs/graph>

243 faithful, highlight the part that is not consistent in
244 the summary sentence.⁶

245 In line with the literature (Maynez et al., 2020;
246 Ramprasad et al., 2024), we also ask the annotators
247 to flag the sentences that are factually hallucinated.
248 That is, the generated sentence is supported by
249 external knowledge but not by the source article.
250 This helps better understand to what extent external
251 knowledge is injected to generate plain summaries
252 as it requires plain explanations of technical terms.

253 We recruit two medical doctors via Upwork.⁷
254 Specifically, before the annotators started the an-
255 notation separately, we started with the annotation
256 training by giving two annotators 6 summaries (34
257 sentences) generated from different LLMs based
258 on different source articles. We consider the inter-
259 annotator agreement (IAA) at the sentence level,
260 i.e., binary faithfulness labels, and at the summary
261 level, considering the summary as faithful if all
262 sentences are annotated as faithful. We achieve
263 a percentage agreement of 0.94 and 0.83 and Co-
264 hen’s Kappa (McHugh, 2012) of 0.48 and 0.57 at
265 the sentence and summary level, respectively. Simi-
266 lar to observations from previous work (Ramprasad
267 et al., 2024; Joseph et al., 2024), faithfulness an-
268 notation is imbalanced, e.g., on average 2 out of 34
269 sentences in those 6 summaries are hallucinated, re-
270 sulting in expected a higher percentage agreement
271 and lower Cohen’s kappa score.

272 We further calculate the IAA on annotated sup-
273 porting sentences based on the subset that both
274 annotators consider to be faithful. Specifically, we
275 consider the agreement where both annotators high-
276 light the same supporting sentences, resulting in
277 Precision, Recall, and F1 of 0.47, 0.56, and 0.51,
278 respectively. Despite the challenge of finding sup-
279 porting sentences, one possible reason for such
280 agreement is that we did not ask annotators to high-
281 light all related supporting sentences. Multiple sen-
282 tences from the source article could solely support
283 the summary sentence. Annotators might overlook
284 other supporting sentences once they find one.⁸

285 4 Faithfulness Evaluation and Analysis 286 on Plain Biomedical Summaries

287 **RQ1. How faithful are generated summaries
288 across different LLMs?** Figure 2 shows the
289 faithfulness annotations across selected LLMs.

⁶Detailed annotation interface is provided in Appendix F.

⁷<https://www.upwork.com/>

⁸More detailed analysis of the annotation of supporting sentences are provided in Appendix H.

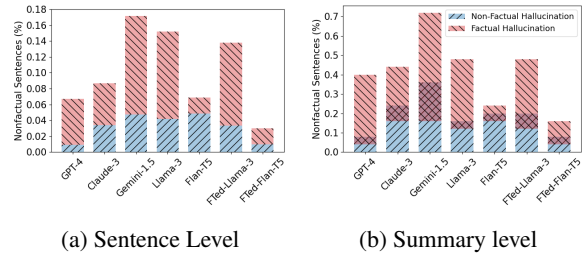


Figure 2: Faithfulness annotation across models at sentence and summary levels. At the summary level, we aggregate the annotations on sentences and consider the summary as faithful if all sentences are faithful.

290 Considering non-factual hallucination (blue and
291 forward-slashed bars), i.e., neither faithful nor fact-
292 ual, we observe a small hallucination rate across
293 all models at the sentence level, i.e., less than 5%
294 sentences that are hallucinated. However, we ob-
295 serve a higher rate at the summary level, with at
296 least 8% summaries containing hallucination (i.e.,
297 at least 2 out of 25 summaries), indicating the per-
298 formance gap in generating faithful and factual
299 plain biomedical summaries.

300 Additionally, as we ask summarizers to gener-
301 ate plain summaries, this might introduce external
302 knowledge to explain technique concepts in simple
303 terms. We label the information where it is correct
304 yet not in the source article as factual hallucination
305 (Cao et al., 2022; Li et al., 2024). As shown in
306 Figure 2 (red and backward-slashed bars), all mod-
307 els, except the Flan-T5 family, show a high rate of
308 factual hallucination at both sentence and summary
309 levels, indicating that models would introduce ex-
310 ternal knowledge to generate plain summaries.

311 Interestingly, Flan-T5 and its finetuned version,
312 show a low non-factual hallucination rate. One
313 possible reason is due to the low abstractiveness of
314 the generated summaries, that is, the model will be
315 inherently more factual if only extracting sentences
316 from the source article. We further investigate this
317 in the following paragraph.

318 **RQ2. How abstractive and readable are the
319 plain biomedical summaries, and how do they
320 relate to faithfulness?** Plain summaries from
321 biomedical articles might incorporate external
322 knowledge, e.g., explaining jargon, to make it more
323 readable for general audiences (Goldsack et al.,
324 2022). This could affect the summary’s abstractive-
325 ness and potentially introduce more hallucinations.

326 To measure the abstractiveness of the summary,
327 we compare the n -gram novelty (See et al., 2017;

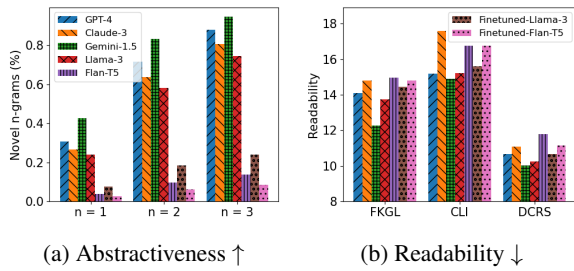


Figure 3: Abstractiveness and readability on generated plain biomedical summaries across different LLMs.

Sharma et al., 2019) between the summary and corresponding source article, i.e., the percentage of non-overlapping n -grams.

As shown in Figure 3a, summaries from closed-source models, i.e., GPT-4, Claude-3, and Gemini-1.5, and open-source model Llama-3 demonstrate high abstractiveness. Interestingly, although layman summaries from the PLOS dataset have shown to be abstractive (Goldsack et al., 2022), models finetuned on this dataset, i.e., Finetuned-Llama-3 and Finetuned-Flan-T5, show a decrease in abstractiveness, compared to the off-the-shelf models.

We calculate the Spearman r correlation between n -grams novelty and the ratio of hallucination in summary, i.e., the percentage of hallucinated sentences in summary. We separate the correlation into two groups: (1) non-factual hallucination and (2) factual hallucination.⁹ We observe Spearman r correlations ranging from 0.2 to 0.24 ($p < 0.05$) between abstractiveness and non-factual hallucination, indicating that the level of abstractiveness could be one factor impacting the non-factual hallucination. Also, strong correlations with factual hallucination, ranging from 0.37 to 0.43 ($p < 0.05$), echo our hypothesis where the generation of plain summaries could introduce external knowledge, i.e., the level of abstractiveness as a proxy, and in turn will impact the faithfulness of summaries.

To evaluate the readability of the summary, we use the standard metrics: Flesch-Kincaid Grade Level (FKGL; Kincaid et al. (1975)), Coleman-Liau Index (CLI; Coleman and Liau (1975)), Dale-Chall Readability Score (DCRS; Dale and Chall (1948)). These metrics measure the approximate (US) grade level of education required to read a given text, by employing experimental formulas on the number of characters, words, and sentences.

As shown in Figure 3b, We observe that GPT-4,

⁹Detail Spearman r correlations between abstractiveness and faithfulness are shown in Appendix D.

Gemini-1.5, and Llama-3 show a lower readability score, i.e., generate more readable summaries across the three metrics, while finetuned models show the opposite. We investigate the correlation between abstractness and readability and observe a Spearman r correlation of at least -0.31 ($p < 0.05$)¹⁰ among the score of n -gram novelty and readability, indicating a negative correlation in these two dimensions. We also observe a negative correlation between readability and faithfulness, where the factual hallucination ratio has a higher negative correlation, with minimum -0.21 ($p < 0.05$) correlation scores. This again indicates that generating more readable plain summaries in the biomedical domain could introduce more factual hallucinations.

RQ3. To what extent do current faithfulness evaluators align with human judgment in plain biomedical summarization?

We compare the human annotations with different types of automatic faithfulness evaluators. We consider two QA-based faithfulness metrics, Questeval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2022) which utilize T5-based models to generate questions and answers based on summaries and source articles. We also compare with Summac (Laban et al., 2022), an entailment-based metric trained on the NLI dataset (FactCC; Kryscinski et al. (2020)), and AlignScore (Zha et al., 2023), an alignment metric measuring the information alignment between two arbitrary text pieces.

Furthermore, following past studies (Wang et al., 2023; Chiang and Lee, 2023a; Liu et al., 2023), we investigate the capability of LLMs as faithfulness evaluators in the biomedical text. Prior work (Chiang and Lee, 2023b) has also shown that prompting LLMs for additional reasoning can boost the faithfulness evaluation. To study this, we construct two types of prompts: (1) *only label*, promoting LLMs to provide faithfulness labels; and, (2) *label & sentences*, prompting LLMs to provide faithfulness labels and supporting sentences from the source (Section 5).¹¹ We exclude Flan-T5 as it cannot produce meaningful results from our prompt.

In Table 2, we measure faithfulness evaluation agreement on automatic evaluation metrics and human judgment at the sentence level, considering both factual and non-factual hallucination in Figure 2 as hallucination. We observe a performance

¹⁰Detail correlations between readability with abstractions and faithfulness are provided in Appendix D.

¹¹Prompt construct is provided in Appendix C.

	Agreement		Prediction Performance	
	Cohen Kappa \uparrow	P. Agreement (%) \uparrow	Recall \uparrow	Pred. Non-faithful (%)
All labeled as faithful	0.00	0.89	-	0.00
All labeled as non-faithful	0.00	0.11	1.00	1.00
GPT-4 (only label)	0.29	0.86	0.40	0.12
GPT-4 (label&sentences)	0.23	0.88	0.23	0.06
Claude-3 (only label)	0.35	0.89	0.38	0.09
Claude-3 (label&sentences)	0.33	0.91	0.25	0.04
Gemini-1.5 (only label)	0.19	0.88	0.17	0.04
Gemini-1.5 (label&sentences)	0.22	0.89	0.19	0.05
Llama-3 (only label)	0.04	0.85	0.09	0.06
Llama-3 (label&sentences)	0.17	0.88	0.17	0.05
QAFactEval	0.11	0.48	0.91	0.61
QuestEval	0.01	0.14	0.99	0.97
SummaCZS	0.09	0.42	0.94	0.68
SummaCConv	0.13	0.49	0.95	0.60
AlignScore	0.16	0.86	0.21	0.08

Table 2: Performance of faithfulness evaluators at the sentence level. “P. Agreement (%)” represents the percentage agreement. “Pred. Non-faithful (%)” represents the percentage of non-faithful instances predicted by evaluators.

gap in improving faithfulness evaluation agreements with human annotation across all faithfulness metrics. Specifically, similar to the prior work (Ramprasad et al., 2024), current metrics trained in the general domain (Questeval, QAFactEval, Summac, and AlignScore), do not achieve strong agreement with human annotation, indicating the difficulty directly transferring those metrics to the biomedical domain. LLM-based evaluators achieve better results compared to traditional metrics. Interestingly, prompting LLMs to additionally provide supporting sentences improves the performance of Gemini-1.5 and Llama-3, but it does not show further improvement for GPT-4 and Claude-3.

We further investigate the capability of evaluators in identifying all hallucinated sentences (Adams et al., 2023b), i.e., Recall. Although QaFactEval, Questeval, and Summac achieved high recall (Table 2, Column 4), they predicted more than 60% sentences as hallucinated, making it impractical considering only 11% sentences contain hallucination. Among other metrics, We observe a low recall in identifying hallucinated sentences, indicating the gap in this direction.

We also aggregate the sentence level annotations to the summary level and evaluate the Pearson and Spearman r correlation (Appendix E). We observe a similar performance gap in the alignment of current faithfulness evaluators and human judgments.

RQ4. Do LLMs consider their generation more faithful? Following previous work (Tam et al., 2023; Panickssery et al., 2024), where they showed that LLMs tend to be over-confident with their gen-

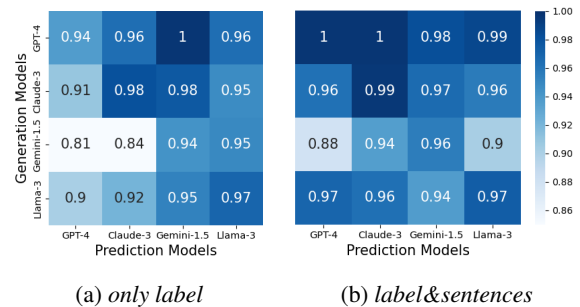


Figure 4: Heatmaps of predicted faithfulness percentage across selected LLM. “only label” and “label&sentences” represent the prompting setting where only responding with the label and with additional supporting sentences, respectively.

erated text, we investigate where this holds in our faithfulness evaluation of plain biomedical summaries. Specifically, we consider the subset of the annotations, only focusing on the summaries from GPT-4, Claude-3, Gemini-1.5, and Llama-3.

Figure 4 shows the heatmap of predicted faithfulness percentage across selected LLMs. We find that LLMs favor their generation with the prompts with returning *label&sentences* (Figure 4b). For the experiment in *only label* (Figure 4a), LLMs tend to consider their generation as more faithful but this is not consistent across models, e.g., GPT-4 results. This is different from the observation from the prior work, where we find that the different construction of prompts would also impact the model tendency regarding their faithfulness preference.

	Number
Instances	1,305
Support Sentences	1,713
Avg. Support Sentences per Instance	1.31
Avg. Support Sentences per Summary	2.10
GPT-4	2.34
Claude-3	2.24
Gemini-1.5	1.87
Llama-3	1.94

Table 3: Statistics of the experiment dataset on supporting sentences identification.

5 Supporting Sentences Identification

In Section 4, we investigate the capability of LLMs in predicting faithful sentences. One following question is whether the models have correct reasoning to support their judgment. In our annotation, we ask the annotators to highlight the supporting evidence from the source article. This enables us to understand if the LLMs can identify the evidence from scientific literature. Specifically, we consider the subset of the dataset where sentences are labeled as faithful and support evidence is provided. The statistic of this subset is shown in Table 3.¹²

For the baseline, we consider Okapi BM25 (Robertson et al., 1995; Trotman et al., 2014), a ranking model based on the term and document frequency. We select the most relevant sentence from the document as the supporting evidence.

It is worth mentioning that, as discussed in Section 3.3, annotators might oversee the support evidence. Although we provide Precision, Recall, and F1 on extract sentence matching, we focus on Recall, i.e., the coverage of the annotated sentence for our analysis, to understand if models can retrieve comprehensive evidence from the source.

RQ5. Can LLMs identify the supporting sentences from the source article? Table 4 (Overall) shows the results of LLMs in identifying supporting sentences. We observe that Okapi BM25 achieves strong performance, i.e., the highest Precision and F1. The low recall might be due to the selection of only one relevant sentence. Across LLMs, GPT-4 achieves the best performance in Recall, i.e., 0.76, indicating the capability of LLMs in identifying supporting evidence from the source articles. Additionally, we observe that models have higher Recall compared to Precision. This might

¹²The post-process of the identified sentences are shown in Appendix G.

	Precision	Recall	F1
Overall			
Okapi BM25	0.73	0.56	0.63
GPT-4	0.43	0.76	0.55
Claude-3	0.41	0.70	0.51
Gemini-1.5	0.48	0.69	0.57
Llama-3	0.38	0.56	0.45
High Abtractiveness			
Okapi BM25	0.61	0.41	0.49
GPT-4	0.41	0.72	0.52
Claude-3	0.41	0.69	0.51
Gemini-1.5	0.46	0.65	0.54
Llama-3	0.37	0.51	0.43
Low Abtractiveness			
Okapi BM25	0.94	0.90	0.92
GPT-4	0.47	0.85	0.60
Claude-3	0.41	0.72	0.52
Gemini-1.5	0.53	0.76	0.63
Llama-3	0.40	0.69	0.50

Table 4: Supporting sentence identification.

be due to the incomprehensive annotation of the support evidence (Section 3.3).

We further investigate the LLMs identified support evidence. We randomly sample 50 summaries from our dataset. As shown in Table 5, we found that errors mostly exist in (1) Annotator Overlooks, (2) Usage of Abbreviation, (3) Copy from Summary Sentence, and (4) Irrelevant Sentences.

RQ6. Does abtractiveness impact the identification of supporting evidence? As shown in Figure 3a, summaries from different LLMs demonstrate different levels of abtractiveness. High abstractive sentences might require a deeper understanding of the text in order to identify the supporting sentences. To study the impact, we further separate the generated summaries into two groups based on our observation: (1) High abtractiveness, i.e., GPT-4, Claude-3, Gemini-1.5, and Llama-3; and (2) Low abtractiveness, i.e., Flan-T5, Finetuned-LLama-3, and Finetuned-Flan-T5.

Table 4 shows the performance on different levels of abtractiveness. Compared to low abtractiveness, we observe a consistent performance drop in the high abtractiveness subset across all models, indicating the impact of abtractiveness and the difficulty in identifying support evidence from high abstractive summaries. Okapi BM25 achieves is the best in low abtractiveness summaries but it suffers when abtractiveness of summaries increases. LLMs achieve high recall in identifying supporting

Source Article	[...] Student participants reported the IEC was relevant (98% agreement) and motivated them to apply theoretical knowledge to a clinical context (97% agreement). The themes identified through qualitative analysis were: factors inherent to the virtual simulation that enabled learning through VSIP, the VSIP supported cognitive apprenticeship, VSIP enabled clinical learning for optometric education, VSIP' role in cross-cultural professional identity development in optometry students. Conclusion The study found that the VSIP platform helped to motivate students to learn and improve their clinical skills. The VSIP was considered a potential supplement to physical clinical placements and could revolutionize global optometric education by offering co-learning across cultures. [...] The International Eyecare Community (IEC) was created with the purpose to incorporate the inherent advantages of virtual simulation and deliver collaborative global education by offering flexible, diverse, personalised, accessible and equal learning opportunities [4,5]. This platform was not created to replace face-to-face teaching; [...].
Summary	It has potential to enhance optometry training by offering flexible, accessible international learning experiences.
Extraction #1	Error: Annotator Overlook: The International Eyecare Community (IEC) was created with the purpose to incorporate the inherent advantages of virtual simulation and deliver collaborative global education by offering flexible, diverse, personalised, accessible and equal learning opportunities [4,5]
Extraction #2	Error: Usage of Abbreviation The IEC was created with the purpose to incorporate the inherent advantages of virtual simulation and deliver collaborative global education by offering flexible, diverse, personalised, accessible and equal learning opportunities
Extraction #3	Error: Copy from Summary Sentence It has potential to enhance optometry training by offering flexible, accessible international learning experiences.
Extraction #4	Error: Irrelevant Sentences Student participants reported the IEC was relevant (98% agreement) and motivated them to apply theoretical knowledge to a clinical context (97% agreement).

Table 5: Error examples of extracted supporting sentences from LLMs. Expert annotations are highlighted (blue) in the source article. Note that the illustrated example does not contain all four types of errors for supporting sentence extraction. For illustration purposes, we adapt the errors from other predictions.

sentences, with high or low abstractiveness. Specifically, in the low abstractiveness setting, GPT-4 achieves a Recall of 0.85, demonstrating its potential to identify supporting evidence.

RQ7. Do LLMs perform better when extracting evidence for their generated summaries?

In Section 4, we observe that LLMs tend to favor their generation in faithfulness evaluation (Tam et al., 2023; Panickssery et al., 2024). Whether this observation holds in identifying supporting sentences is yet to be explored. That is, assuming LLMs generate summaries based on their reasoning, would LLMs retrieve evidence for its generation, as it might follow a similar reasoning process? We plot the heatmap of LLMs' retrieval performance across different subsets of summaries generated by different LLMs in Figure 5. Overall, Gemini-1.5 archives higher precision across all summary subsets (Figure 5a, Third Column). GPT-4 consistently achieves the highest recall among the generated summaries (Figure 5b, First Column). Among the generated text, LLMs achieve higher precision based on the summaries from GPT-4 (Figure 5a, First Row) and higher recall from Claude-3 (Figure 5b, Second Row). However, we do not observe that the model outperforms the others when identifying the supporting evidence from its generation, i.e., LLMs do not necessarily outperform other models when reasoning its generation over others. The retrieval performance might be affected by other factors, e.g., abstractiveness.

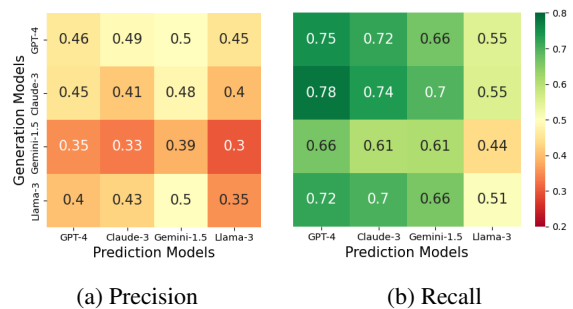


Figure 5: Heatmaps of LLMs' retrieval across different subsets of summaries generated by different LLMs.

6 Conclusions

We create a benchmark dataset, FAREBIO, with expert-annotated faithfulness evaluation and reasoning for plain biomedical summaries, consisting of 175 summaries and 1445 sentences from 7 different LLMs. We use this dataset to evaluate the faithfulness of prevalent LLMs and measure the transferability of current faithfulness metrics to the biomedical domain. We also observe a positive correlation in abstractiveness and faithfulness and find that the construction of prompts could also affect the faithfulness prediction preferences. We further benchmark the capability of LLMs in retrieving supporting sentences for the plain summaries.

By going beyond the binary faithfulness labels, equipped with annotations of faithfulness and reasoning, our dataset could further deepen the study of faithfulness in better understanding the reasoning behind LLMs for their faithfulness judgment.

7 Limitations

One of the main challenges in benchmarking the faithfulness of plain biomedical summaries is the cost involved in hiring domain experts. For our annotation, we hire two medical doctors at \$50 USD/hr. Although we facilitated the annotation via providing the annotators the summaries that are from the same source article, minimizing the time in understanding the source article, the total annotation of 1445 sentences still required approximately 110 human hours, i.e., \$5,500 USD, making scalability of the annotations challenging.

Another challenge of faithfulness annotation in the biomedical domain is to understand the generation hallucination. Our IAA of faithfulness evaluation aligns with the previous work (Ramprasad et al., 2024). We ask the annotators to highlight the inconsistent part from the summary and provide a brief rationale for their judgement. This could be used for further categorizing and analyzing the hallucination errors of LLMs in generating plain biomedical summaries.

For our supporting sentences identification task, we used exact matching for sentence evaluation. As discussed in Section 3.3, this can not capture semantically similar sentences, e.g., paraphrased sentences or omitted sentences due to the overlooking annotations from the annotators. Other evaluation metrics, e.g., ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020), would be worth investigating in complementing the evaluation on supporting sentences identification.

Our work aims to benchmark the faithfulness analysis of current LLMs, investigate the faithfulness alignment of off-the-shelf evaluators with human judgment, and the capability of LLMs in identifying support evidence. We select a subset of the representative LLMs. We do not cover all available LLMs across different variances (e.g., GPT-3.5, Llama-2-70B, and Llama-3-70B), nor is it possible to do so. Further analysis can be enhanced by including other types and variances of LLMs.

Additionally, our work proposes a benchmark faithfulness dataset in the biomedical domain and our models serve as baselines for investigating the capability of LLMs. Throughout our experiment, we follow the general prompt from prior works. One promising direction for improving model performance in the generation of plain biomedical summaries and the utilization as a faithfulness evaluator and identifier of the support evidence could

potentially be employing more advanced prompt engineering methods, e.g., automatic prompt generation (Ha et al., 2023; Zhou et al., 2022; Li and Liang, 2021) and Chain-of-Thought (CoT) prompts (Kojima et al., 2022; Yu et al., 2023).

There are also other approaches to improve the performance of LLMs as evaluators. For instance, FactScore (Min et al., 2023) extracts atomic facts from the text from LLMs and compares the consistency of the extracted facts; Lattimer et al. (2023) directly use the prediction probability “yes” and “no” from open-source models, i.e., T5, to infer the faithfulness. How to use LLMs as faithfulness evaluators in the biomedical domain would be a promising direction.

8 Ethical Discussion

For our annotation, we hired two native English-speaking annotators via Upwork and we recruited the annotators based on their expertise. We did not record any personal information of the annotators. We paid the annotator at an hourly rate of \$50 USD, which far exceeds the local minimum pay rate.

For copyright, we obtained the PubMed article from S2ORC (Lo et al., 2020), which is under the licence ODC-By 1.0.¹³ OpenAI provides Terms of Use¹⁴ for the usage of GPT-4. Anthropic provides the Consumer Terms of Service for Claude-3.¹⁵ Gemini-1.5 follows the Google Generative AI terms.¹⁶ Llama-3 is under licence “META LLAMA 3 COMMUNITY LICENSE AGREEMENT”.¹⁷ Flan-T5 is under licence “Apache License 2.0”.¹⁸

For the choice of LLMs, we surveyed the current available LLMs and selected the representative and prevalent LLMs from different categories for our study. We aim to explore the faithfulness and reasoning of current LLMs and we make no attempt to target any particular LLMs.

We randomly select the PubMed articles from the publicly available scholarly dataset. Our dataset, along with the generated content, should be only for research purposes. We do not encourage other usage. Additionally, The PubMed articles might contain authors’ information and associated affiliations. We are against any usage of these in-

¹³<https://opendatacommons.org/licenses/by/1-0/>

¹⁴<https://openai.com/policies/terms-of-use/>

¹⁵<https://www.anthropic.com/legal/consumer-terms>

¹⁶<https://policies.google.com/terms/generative-ai/use-policy>

¹⁷<https://llama.meta.com/llama3/license/>

¹⁸<https://choosealicense.com/licenses/apache-2.0/>

674	formation to target individuals.	
675	References	
676	Josh Achiam, Steven Adler, Sandhini Agarwal, Lama	
677	Ahmad, Ilge Akkaya, Florencia Leoni Aleman,	
678	Diogo Almeida, Janko Altenschmidt, Sam Altman,	
679	Shyamal Anadkat, et al. 2023. GPT-4 Technical Re-	
680	port. <i>arXiv preprint arXiv:2303.08774</i> .	
681	Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric	
682	Lehman, and Noémie Elhadad. 2023a. From sparse	
683	to dense: GPT-4 summarization with chain of density	
684	prompting . In <i>Proceedings of the 4th New Frontiers</i>	
685	<i>in Summarization Workshop</i> , pages 68–74, Singapore.	
686	Association for Computational Linguistics.	
687	Griffin Adams, Jason Zuckerg, and Noémie Elhadad.	
688	2023b. A Meta-Evaluation of Faithfulness Metrics	
689	for Long-Form Hospital-Course Summarization. In	
690	<i>Machine Learning for Healthcare Conference</i> , pages	
691	2–30. PMLR.	
692	Amal Alqahtani, Rana Salama, Mona Diab, and Abdou	
693	Youssef. 2023. Care4Lang at MEDIQA-chat 2023:	
694	Fine-tuning language models for classifying and sum-	
695	marizing clinical dialogues . In <i>Proceedings of the</i>	
696	<i>5th Clinical Natural Language Processing Workshop</i> ,	
697	pages 524–528, Toronto, Canada. Association for	
698	Computational Linguistics.	
699	Anthropic. 2024. Meet Claude . Accessed on 05 31,	
700	2024.	
701	Tom Brown, Benjamin Mann, Nick Ryder, Melanie	
702	Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind	
703	Neelakantan, Pranav Shyam, Girish Sastry, Amanda	
704	Askell, Sandhini Agarwal, Ariel Herbert-Voss,	
705	Gretchen Krueger, Tom Henighan, Rewon Child,	
706	Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens	
707	Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma-	
708	teusz Litwin, Scott Gray, Benjamin Chess, Jack	
709	Clark, Christopher Berner, Sam McCandlish, Alec	
710	Radford, Ilya Sutskever, and Dario Amodei. 2020.	
711	Language models are few-shot learners . In <i>Ad-</i>	
712	<i>vances in Neural Information Processing Systems</i> ,	
713	volume 33, pages 1877–1901. Curran Associates, Inc.	
714		
715	Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hal-	
716	lucinated but factual! inspecting the factuality of	
717	hallucinations in abstractive summarization . In <i>Pro-</i>	
718	<i>ceedings of the 60th Annual Meeting of the Associa-</i>	
719	<i>tion for Computational Linguistics (Volume 1: Long</i>	
720	<i>Papers)</i> , pages 3340–3354, Dublin, Ireland. Associa-	
721	tion for Computational Linguistics.	
722	Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive	
723	learning for improving faithfulness and factuality in	
724	abstractive summarization . In <i>Proceedings of the</i>	
725	<i>2021 Conference on Empirical Methods in Natural</i>	
726	<i>Language Processing</i> , pages 6633–6649, Online and	
727	Punta Cana, Dominican Republic. Association for	
728	Computational Linguistics.	
	Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large	729
	language models be an alternative to human evalua-	730
	tions? In <i>Proceedings of the 61st Annual Meeting of</i>	731
	<i>the Association for Computational Linguistics (Vol-</i>	732
	<i>ume 1: Long Papers)</i> , pages 15607–15631, Toronto,	733
	Canada. Association for Computational Linguistics.	734
	Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer	735
	look into using large language models for automatic	736
	evaluation . In <i>Findings of the Association for Com-</i>	737
	<i>putational Linguistics: EMNLP 2023</i> , pages 8928–	738
	8942, Singapore. Association for Computational Lin-	739
	guistics.	740
	Hyung Won Chung, Le Hou, Shayne Longpre, Barret	741
	Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang,	742
	Mostafa Dehghani, Siddhartha Brahma, Albert Web-	743
	son, Shixiang Shane Gu, Zhuyun Dai, Mirac Suz-	744
	gun, Xinyun Chen, Aakanksha Chowdhery, Sharan	745
	Narang, Gaurav Mishra, Adams Yu, Vincent Zhao,	746
	Yanping Huang, Andrew Dai, Hongkun Yu, Slav	747
	Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam	748
	Roberts, Denny Zhou, Quoc V. Le, and Jason Wei.	749
	2022. Scaling Instruction-Finetuned Language Mod-	750
	els . <i>arXiv preprint</i> .	751
	Meri Coleman and Ta Lin Liau. 1975. A computer	752
	readability formula designed for machine scoring.	753
	<i>Journal of Applied Psychology</i> , 60(2):283.	754
	Edgar Dale and Jeanne S Chall. 1948. A formula for	755
	predicting readability: Instructions. <i>Educational re-</i>	756
	<i>search bulletin</i> , pages 37–54.	757
	Esin Durmus, He He, and Mona Diab. 2020. FEQA: A	758
	question answering evaluation framework for faith-	759
	fulness assessment in abstractive summarization . In	760
	<i>Proceedings of the 58th Annual Meeting of the Asso-</i>	761
	<i>ciation for Computational Linguistics</i> , pages 5055–	762
	5070, Online. Association for Computational Lin-	763
	guistics.	764
	Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and	765
	Caiming Xiong. 2022. QAFactEval: Improved QA-	766
	based factual consistency evaluation for summariza-	767
	tion . In <i>Proceedings of the 2022 Conference of the</i>	768
	<i>North American Chapter of the Association for Com-</i>	769
	<i>putational Linguistics: Human Language Technolo-</i>	770
	<i>gies</i> , pages 2587–2601, Seattle, United States. Asso-	771
	ciation for Computational Linguistics.	772
	Alexander R. Fabbri, Wojciech Kryściński, Bryan Mc-	773
	Cann, Caiming Xiong, Richard Socher, and Dragomir	774
	Radev. 2021. SummEval: Re-evaluating summariza-	775
	tion evaluation . <i>Transactions of the Association for</i>	776
	<i>Computational Linguistics</i> , 9:391–409.	777
	Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie	778
	Utama, Ido Dagan, and Iryna Gurevych. 2019. Rank-	779
	ing generated summaries by correctness: An interest-	780
	ing but challenging application for natural language	781
	inference . In <i>Proceedings of the 57th Annual Meet-</i>	782
	<i>ing of the Association for Computational Linguistics</i> ,	783
	pages 2214–2220, Florence, Italy. Association for	784
	Computational Linguistics.	785

786	Zorik Gekhman, Jonathan Herzig, Roei Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 2053–2070, Singapore. Association for Computational Linguistics.	843
787		844
788		845
789		
790		846
791		847
792		848
793	Tirthankar Ghosal, Philipp Mayr, Anita de Waard, Aakanksha Naik, Shannon Shen, Amanpreet Singh, Orion Weller, Yanxia Qin, and Yoonjoo Lee, editors. 2024. <i>The 4th Workshop on Scholarly Document Processing</i> . Association for Computational Linguistics, Bangkok, Thailand.	849
794		850
795		851
796		852
797		
798		
799	Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles . In <i>The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks</i> , pages 468–477, Toronto, Canada. Association for Computational Linguistics.	853
800		854
801		855
802		856
803		857
804		
805		
806		
807	Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	858
808		859
809		860
810		861
811		862
812		
813		
814	Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. <i>Journal of Biomedical Informatics</i> , 149:104580.	863
815		864
816		865
817		866
818	Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 35, pages 160–168.	867
819		868
820		869
821		
822		
823	Hyeonmin Ha, Jihye Lee, Wookje Han, and Byung-Gon Chun. 2023. Meta-learning of prompt generation for lightweight prompt engineering on language-model-as-a-service . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 2433–2445, Singapore. Association for Computational Linguistics.	870
824		871
825		872
826		873
827		
828		
829		
830	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. <i>arXiv preprint arXiv:2311.05232</i> .	874
831		875
832		876
833		877
834		878
835		
836	Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks. <i>Computers in Biology and Medicine</i> , page 108189.	879
837		880
838		881
839		882
840		883
841	Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea	884
842		885
		886
		887
		888
		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900
		901
		902
		903
		904
		905
		906
		907
		908
		909
		910
		911
		912
		913
		914
		915
		916
		917
		918
		919
		920
		921
		922
		923
		924
		925
		926
		927
		928
		929
		930
		931
		932
		933
		934
		935
		936
		937
		938
		939
		940
		941
		942
		943
		944
		945
		946
		947
		948
		949
		950
		951
		952
		953
		954
		955
		956
		957
		958
		959
		960
		961
		962
		963
		964
		965
		966
		967
		968
		969
		970
		971
		972
		973
		974
		975
		976
		977
		978
		979
		980
		981
		982
		983
		984
		985
		986
		987
		988
		989
		990
		991
		992
		993
		994
		995
		996
		997
		998
		999
		1000

899	Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation . In <i>Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)</i> , pages 4582–4597, Online. Association for Computational Linguistics.	957
900		958
901		959
902		960
903		961
904		962
905		963
906		
907	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	964
908		965
909		966
910		967
911	Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models . In <i>Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)</i> , pages 47–58, Toronto, Canada. Association for Computational Linguistics.	968
912		969
913		970
914		971
915		972
916		973
917		974
918	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 2511–2522, Singapore. Association for Computational Linguistics.	975
919		976
920		977
921		978
922		979
923		980
924		
925	Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 4969–4983, Online. Association for Computational Linguistics.	981
926		982
927		983
928		984
929		985
930		986
931	Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In <i>Proceedings of the 40th International Conference on Machine Learning, ICML’23</i> . JMLR.org.	987
932		988
933		989
934		990
935		991
936		992
937		993
938	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 1906–1919, Online. Association for Computational Linguistics.	994
939		995
940		996
941		997
942		
943		
944	Mary L McHugh. 2012. Interrater reliability: the kappa statistic. <i>Biochemia medica</i> , 22(3):276–282.	998
945		999
946	Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date . Accessed on 05/31/2024.	1000
947		1001
948		1002
949	Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 12076–12100, Singapore. Association for Computational Linguistics.	1003
950		1004
951		1005
952		1006
953		1007
954		1008
955		1009
956		1010
	Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 4885–4901, Online. Association for Computational Linguistics.	1011
		1012
		1013
		1014
	Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 4812–4829, Online. Association for Computational Linguistics.	957
		958
		959
		960
		961
		962
		963
	Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations . <i>arXiv preprint arXiv:2404.13076</i> .	972
		973
		974
	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>Journal of machine learning research</i> , 21(140):1–67.	975
		976
		977
		978
		979
		980
	Sanjana Ramprasad, Kundan Krishna, Zachary Lipton, and Byron Wallace. 2024. Evaluating the factuality of zero-shot summarizers across varied domains . In <i>Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 50–59, St. Julian’s, Malta. Association for Computational Linguistics.	981
		982
		983
		984
		985
		986
		987
		988
	Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soriccut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context . <i>arXiv preprint arXiv:2403.05530</i> .	989
		990
		991
		992
		993
		994
	Stephen Robertson, Steve Walker, Susan Jones, Michelle Hancock-Beaulieu, and Mike Gafford. 1995. Okapi at TREC-3 . In <i>TREC</i> , Gaithersburg, MD, US.	995
		996
		997
	Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine . <i>arXiv preprint arXiv:2404.18416</i> .	998
		999
		1000
		1001
		1002
	Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	1003
		1004
		1005
		1006
		1007
		1008
		1009
		1010
	Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks . In <i>Proceedings of the 55th Annual Meeting of the Association for Computational</i>	1011
		1012
		1013
		1014

1015	<i>Linguistics (Volume 1: Long Papers)</i> , pages 1073–			
1016	1083, Vancouver, Canada. Association for Computa-			
1017	tional Linguistics.			
1018	Chantal Shaib, Millicent Li, Sebastian Joseph, Iain			
1019	Marshall, Junyi Jessy Li, and Byron Wallace. 2023.			
1020	Summarizing, simplifying, and synthesizing medical			
1021	evidence using GPT-3 (with varying success) . In			
1022	<i>Proceedings of the 61st Annual Meeting of the As-</i>			
1023	<i>sociation for Computational Linguistics (Volume 2:</i>			
1024	<i>Short Papers)</i> , pages 1387–1407, Toronto, Canada.			
1025	Association for Computational Linguistics.			
1026	Eva Sharma, Chen Li, and Lu Wang. 2019. BIG-			
1027	PATENT: A large-scale dataset for abstractive and			
1028	coherent summarization . In <i>Proceedings of the 57th</i>			
1029	<i>Annual Meeting of the Association for Computational</i>			
1030	<i>Linguistics</i> , pages 2204–2213, Florence, Italy. Asso-			
1031	ciation for Computational Linguistics.			
1032	Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and			
1033	Sarvnaz Karimi. 2023. CSIRO Data61 team at Bio-			
1034	LaySumm task 1: Lay summarisation of biomedical			
1035	research articles using generative models . In <i>The</i>			
1036	<i>22nd Workshop on Biomedical Natural Language</i>			
1037	<i>Processing and BioNLP Shared Tasks</i> , pages 629–			
1038	635, Toronto, Canada. Association for Computational			
1039	Linguistics.			
1040	Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan.			
1041	2023. Evaluation metrics in the era of GPT-4: Reli-			
1042	ably evaluating large language models on sequence			
1043	to sequence tasks . In <i>Proceedings of the 2023 Con-</i>			
1044	<i>ference on Empirical Methods in Natural Language</i>			
1045	<i>Processing</i> , pages 8776–8788, Singapore. Associa-			
1046	tion for Computational Linguistics.			
1047	Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah			
1048	Kwan, Mohit Bansal, and Colin Raffel. 2023. Evalu-			
1049	ating the factual consistency of large language mod-			
1050	els through news summarization . In <i>Findings of</i>			
1051	<i>the Association for Computational Linguistics: ACL</i>			
1052	<i>2023</i> , pages 5220–5255, Toronto, Canada. Associa-			
1053	tion for Computational Linguistics.			
1054	Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe La-			
1055	ban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscin-			
1056	ski, Justin Rousseau, and Greg Durrett. 2023. Un-			
1057	derstanding factual errors in summarization: Errors,			
1058	summarizers, datasets, error detectors . In <i>Proceed-</i>			
1059	<i>ings of the 61st Annual Meeting of the Association for</i>			
1060	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,			
1061	pages 11626–11644, Toronto, Canada. Association			
1062	for Computational Linguistics.			
1063	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann			
1064	Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,			
1065	and Tatsunori B. Hashimoto. 2023. Stanford Alpaca:			
1066	An Instruction-following LLaMA model. https://github.com/tatsu-			
1067	lab/stanford_alpaca .			
1068	James Thorne, Andreas Vlachos, Christos			
1069	Christodoulopoulos, and Arpit Mittal. 2018. FEVER:			
1070	a large-scale dataset for fact extraction			
1071	and VERification . In <i>Proceedings of the 2018</i>			
	<i>Conference of the North American Chapter of</i>			
	<i>the Association for Computational Linguistics:</i>			
	<i>Human Language Technologies, Volume 1 (Long</i>			
	<i>Papers)</i> , pages 809–819, New Orleans, Louisiana.			
	Association for Computational Linguistics.			
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-			
	bert, Amjad Almahairi, Yasmine Babaei, Nikolay			
	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti			
	Bhosale, et al. 2023. Llama 2: Open Founda-			
	tion and Fine-Tuned Chat Models. <i>arXiv preprint</i>			
	<i>arXiv:2307.09288</i> .			
	Andrew Trotman, Antti Puurula, and Blake Burgess.			
	2014. Improvements to BM25 and language mod-			
	els examined . In <i>Proceedings of the 2014 Aus-</i>			
	<i>tralasian Document Computing Symposium, ADCS</i>			
	<i>2014, Melbourne, VIC, Australia, November 27-28,</i>			
	<i>2014</i> , page 58. ACM.			
	David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu			
	Wang, Madeleine van Zuylen, Arman Cohan, and			
	Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying			
	scientific claims . In <i>Proceedings of the 2020 Con-</i>			
	<i>ference on Empirical Methods in Natural Language</i>			
	<i>Processing (EMNLP)</i> , pages 7534–7550, Online. As-			
	sociation for Computational Linguistics.			
	Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui			
	Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu,			
	and Jie Zhou. 2023. Is ChatGPT a good NLG evalua-			
	tor? a preliminary study . In <i>Proceedings of the 4th</i>			
	<i>New Frontiers in Summarization Workshop</i> , pages			
	1–11, Singapore. Association for Computational Lin-			
	guistics.			
	Fangyi Yu, Lee Quartey, and Frank Schilder. 2023.			
	Exploring the effectiveness of prompt engineering			
	for legal reasoning tasks . In <i>Findings of the Asso-</i>			
	<i>ciation for Computational Linguistics: ACL 2023</i> ,			
	pages 13582–13596, Toronto, Canada. Association			
	for Computational Linguistics.			
	Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu.			
	2023. AlignScore: Evaluating factual consistency			
	with a unified alignment function . In <i>Proceedings</i>			
	<i>of the 61st Annual Meeting of the Association for</i>			
	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,			
	pages 11328–11348, Toronto, Canada. Association			
	for Computational Linguistics.			
	Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023.			
	Extractive summarization via ChatGPT for faithful			
	summary generation . In <i>Findings of the Associa-</i>			
	<i>tion for Computational Linguistics: EMNLP 2023</i> ,			
	pages 3270–3278, Singapore. Association for Com-			
	putational Linguistics.			
	Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q.			
	Weinberger, and Yoav Artzi. 2020. BERTScore:			
	Evaluating Text Generation with BERT . In <i>Inter-</i>			
	<i>national Conference on Learning Representations</i> .			
	Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab,			
	Francisco Guzmán, Luke Zettlemoyer, and Marjan			
	Ghazvininejad. 2021. Detecting hallucinated content			

1129 in conditional neural sequence generation. In *Find-*
1130 *ings of the Association for Computational Linguis-*
1131 *tics: ACL-IJCNLP 2021*, pages 1393–1404, Online.
1132 Association for Computational Linguistics.

1133 Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han,
1134 Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy
1135 Ba. 2022. Large Language Models Are Human-Level
1136 Prompt Engineers.

1137 A Finetuning on the PLOS dataset

1138 We follow the instructions on Appendix B to fin-
1139 tune Llama-3 and Flan-T5 on the PLOS training
1140 dataset (Goldsack et al., 2022).

1141 For Llama-3, we follow the hyper-parameters
1142 from *llama-recipes*¹⁹ and finetune Llama-3-8B-
1143 Instruction on the PLOS training set for 10
1144 epochs. The model was finetuned on 3 H100
1145 GPUs for 5 hours. We selected the checkpoint that
1146 has the best perplexity performance on the PLOS
1147 dev set.

1148 For Flan-T5, we follow the approach in (Sim
1149 et al., 2023) to fine-tune the model We fine-tune
1150 the Flan-T5-xl (3B) model on the PLOS train-
1151 ing set for 5 epochs and use a beam search decoder
1152 during inference—a beam width of four—to gener-
1153 ate up to 386 tokens.

1154 Table 6 shows the performance of fine-tuned
1155 models on the PLOS test set.

1156 B Prompt Construction for Plain 1157 Summary Generation.

1158 We adopt a general prompt similar to prior work
1159 (Sottana et al., 2023) for generating plain biomed-
1160 ical summaries across different models. Specifically,
1161 we construct the prompt where the instruction is
1162 provided first and followed by the text. using the
1163 template as follows for all models:

1164 *Summarize this article for non-experts:*

1165 *Article:*

1166 *Title: [Title]*

1167 *Authors: [Authors]*

1168 *Abstract: [Abstract]*

1169 *[First section Name]: [First section context]*

1170 *Summary:*

1171 where *[Title]*, *[Authors]*, and *[Abstract]* repre-
1172 sent the content of the title, authors, and abstract,
1173 respectively. *[First section Name]* and *[First sec-*
1174 *tion context]* denote the name of the first section
1175 of the source article (e.g., *Introduction*) and the
1176 corresponding content, respectively.

¹⁹<https://github.com/meta-llama/llama-recipes/tree/main/recipes/finetuning>

1177 C Prompt Construction for Faithfulness 1178 Evaluation.

1179 To utilize LLMs as faithfulness evaluators, we
1180 adopt the evaluation prompt from prior work
1181 (Gekhman et al., 2023). Specifically, we use the fol-
1182 lowing template for GPT-4, Claude-3, Gemini-1.5,
1183 and Llama-3:

1184 *Source:*

1185 *Title: [Title]*

1186 *Authors: [Authors]*

1187 *Abstract: [Abstract]*

1188 *[First section Name]: [First section context]*

1189 *Summary: [Summary sentence]*

1190 *[Evaluation prompt]*

1191 where *[Title]*, *[Authors]*, *[Abstract]*, *[First sec-*
1192 *tion Name]*, and *[First section context]* are denoted
1193 as in Appendix B. *[Summary sentence]* represents
1194 the sentence from the generated summary. For
1195 cases where only prompting LLMs to return the
1196 faithfulness label, *[Evaluation prompt]* represent *Is*
1197 *the Summary supported by the Source? Answer us-*
1198 *ing "Yes" or "No" only.*; For cases where prompting
1199 LLMs to return the faithfulness label and support-
1200 ing sentences from the source, *[Evaluation prompt]*
1201 represent *Is the Summary supported by the Source?*
1202 *Answer using "Yes" or "No" and extract the sup-*
1203 *porting sentences from the Source.*

1204 D Correlations between Abstractiveness 1205 and Readability with Faithfulness

1206 Table 7 shows the Spearman r correlations between
1207 N-gram novelty and readability scores with faith-
1208 fulness.

1209 Table 8 shows the Spearman r correlations be-
1210 tween N-gram novelty and readability.

1211 E Performance of Faithfulness 1212 Evaluators at the Summary Level

1213 Table 9 shows the performance of faithfulness eval-
1214 uators aligning with human judgments.

1215 F Annotation Guidelines

1216 We include screenshots of the annotation interface
1217 of our task in Figure 6, Figure 7, and Figure 8.

1218 G Post Porcess of Extract Sentences

1219 We use the prompt from Appendix C to evaluate
1220 the capability of LLMs in extracting supporting
1221 sentences. As shown in the examples in Table 10,
1222 based on our observation, we find that almost all

	R-1	R-2	R-L	FKGL	DCRS
Finetuned-Llama-3	0.46	0.16	0.25	13.81	11.01
Finetuned-Flan-T5	0.45	0.17	0.26	14.67	11.29

Table 6: Performance of fine-tuned models on the PLOS test set (R = average ROUGE F1-score).

	Non-factual Hallucination	Factual Hallucination
Abtractiveness		
N=1	0.24 (p=0.00)	0.43 (p=0.00)
N=2	0.21 (p=0.01)	0.40 (p=0.00)
N=3	0.20 (p=0.01)	0.37 (p=0.00)
Readability		
FKGL	-0.28 (p=0.00)	-0.21 (p=0.01)
CLI	-0.13 (p=0.08)	-0.24 (p=0.00)
DCRS	-0.20 (p=0.01)	-0.33 (p=0.00)

Table 7: The Spearman r correlation between abstractiveness (n -grams) and Readability (i.e., FKGL, CLI, and DCRS) with the ratio of hallucination (i.e., non-factual and factual hallucination) in summary.

Target of the summary:
 This summary is for the non-expert audience. It is supposed to deliver the important information of the source document, using simple words and explaining technical jargon if any (e.g., metaphors or analogies). Additionally, the sentence structure and logic are purposefully simple as it is for the general purpose.

Annotation task:
Is the annotating sentence from the summary supported/verified by the source document?

Yes. Highlight the evidence on the source document.^[1]
 No. Please highlight the part that is not factual on the annotating sentence.^[2]
 Not Judgable: This sentence is not understandable due to sever linguistic errors (e.g., grammar)^[3]

Please provide a brief rationale of your choice.

Source Document

Title:
Fine spectral tuning of a flavin-binding fluorescent protein for multicolor imaging

Authors:
Andrey Nikolae, From the Research Center for Molecular Mechanisms of Aging and Age-Related Diseases, Advanced Optical Microscopy Centre and Biomedical Research Institute, Moscow Institute of Physics and Technology, DolgoprudnyRussia|Anna Yudenko, Employment at the Moscow Institute of Physics, Hasselt University, DiepenbeekBelgium

Abstract:
Reviewed by members of the JBC Editorial Board Edited by Ruma BanerjeeFlavin-binding fluorescent proteins are promising genetically encoded tags for microscopy. However, spectral properties of their chromophores (riboflavin, flavin mononucleotide, and flavin adenine dinucleotide) are notoriously similar even between different protein families, which limits applications of flavoproteins in multicolor imaging. Here, we present a palette of 22

Summary for non-experts

Preceding sentences

Annotate this sentence

Researchers have developed a set of fluorescent proteins that can emit light in different colors, which is useful for microscopy.

inconsistent part 5

Following sentences

These proteins are based on a type of protein that binds to a compound called flavin, found in many organisms. Flavin-based proteins typically emit light at similar wavelengths, making it hard to distinguish them in multicolor imaging. To overcome this limitation, the researchers modified the structure of the protein to finely tune the emitted light to different colors. They achieved this by making

⏪ ⏩ ✕ 🗨
Submit

Figure 6: An example of the annotation interface.

Target of the summary:

This summary is for the non-expert audience. It is supposed to deliver the important information of the source document, using simple words and explaining technical jargon if any (e.g., metaphors or analogies). Additionally, the sentence structure and logic are purposefully simple as it is for the general purpose.

Annotation task:

Is the annotating sentence from the summary supported/verified by the source document?

Yes. Highlight the evidence on the source document.^[1]
 No. Please highlight the part that is not factual on the annotating sentence.^[2]
 Not judgable: This sentence is not understandable due to sever linguistic errors (e.g., grammar)^[3]

Please provide a brief rationale of your choice.

Range of 486 to 512 shows different colors [↗](#) [🗒](#)

Title:
Fine spectral tuning of a flavin-binding fluorescent protein for multicolor imaging

Authors:
Andrey Nikolaev, From the Research Center for Molecular Mechanisms of Aging and Age-Related Diseases, Advanced Optical Microscopy Centre and Biomedical Research Institute, Moscow Institute of Physics and Technology, DolgoprudnyRussia/Anna Yudenko, Employment at the Moscow Institute of Physics, Hasselt University, DiepenbeekBelgium

Abstract:
Reviewed by members of the JBC Editorial Board Edited by Ruma BanerjeeFlavin-binding fluorescent proteins are promising genetically encoded tags for microscopy. However, spectral properties of their chromophores (riboflavin, flavin mononucleotide, and flavin adenine dinucleotide) are notoriously similar even between different protein families, which limits applications of flavoproteins in multicolor imaging. Here, we present a palette of 22 finely tuned fluorescent tags based on the thermostable LOV domain from Chloroflexus aggregans. We performed site saturation mutagenesis of three amino acid positions in the flavin-binding pocket, including the photoactive cysteine, to obtain variants with fluorescence emission maxima uniformly covering the wavelength range from 486 to 512 nm. We demonstrate three-color imaging based on spectral separation and two-color fluorescence

Summary for non-experts

Preceding sentences

Annotate this sentence

Researchers have developed a set of fluorescent proteins that can emit light in different colors, which is useful for microscopy.

Following sentences

These proteins are based on a type of protein that binds to a compound called flavin, found in many organisms. Flavin-based proteins typically emit light at similar wavelengths, making it hard to distinguish them in multicolor imaging. To overcome this limitation, the researchers modified the structure of the protein to finely tune the emitted light to different colors. They achieved this by making specific changes to the protein's structure at certain locations. By doing this, they created a range of proteins that emit light in colors

[Update](#)

Figure 7: An example of the faithful annotation.

Target of the summary:

This summary is for the non-expert audience. It is supposed to deliver the important information of the source document, using simple words and explaining technical jargon if any (e.g., metaphors or analogies). Additionally, the sentence structure and logic are purposefully simple as it is for the general purpose.

Annotation task:

Is the annotating sentence from the summary supported/verified by the source document?

Yes^[1]
 No. Please highlight the part that is not factual on the annotating sentence.^[2]
 Not judgable: This sentence is not understandable due to sever linguistic errors (e.g., grammar)^[3]

Please provide a brief rationale of your choice.

Changes in topo II include reduced activity which makes the drug less effective. If the word 'activity' is used alone, it implies that there is more which may confuse readers. [↗](#) [🗒](#)

play an important role in resistance to clinically active drugs. However, in clinical trials, modulating the multidrug-resistant phenotype with agents that inhibit the efflux pump has not had an impact. Since reduced drug accumulation per se is insufficient to explain tumor cell resistance to topo II inhibitors several studies have focused on characterizing mechanisms that impact on DNA damage mediated by drugs that target the enzyme. Mammalian topo I α and topo I β isozymes exhibit similar catalytic, but different biologic, activities. Whereas topo I α is associated with cell division, topo I β is involved in differentiation. In addition to site specific mutations that can affect drug-induced topo I mediated DNA damage, post-translation modification of topo II primarily by phosphorylation can potentially affect enzyme-mediated DNA damage and the downstream cytotoxic response of drugs targeting topo II. Signaling pathways that can affect phosphorylation and changes in intracellular calcium levels/calcium dependent signaling that can regulate site-specific phosphorylation of topoisomerase have an impact on downstream cytotoxic effects of topo II inhibitors. Overall, tumor cell resistance to inhibitors of topo II is a complex process that is orchestrated not only by cellular pharmacokinetics but more importantly by enzymatic alterations that govern the intrinsic drug sensitivity.

Summary for non-experts

Preceding sentences

Cancer cells can develop resistance to certain chemotherapy drugs, such as topoisomerase II inhibitors, which are used to treat various types of cancer. Researchers have found that this resistance is not just due to the cancer cells pumping the drug out, but also due to changes within the cells that affect how the drug works.

Annotate this sentence

Specifically, changes in an enzyme called topoisomerase II and its activity can make the drug less effective.

Following sentences

[Update](#)

Figure 8: an example of the hallucinated annotation.

	FKGL	CLI	DCRS
N=1	-0.44 (p=0.0)	-0.34 (p=0.0)	-0.54 (p=0.0)
N=2	-0.41 (p=0.0)	-0.33 (p=0.0)	-0.51 (p=0.0)
N=3	-0.40 (p=0.0)	-0.31 (p=0.0)	-0.49 (p=0.0)

Table 8: Spearman r correlations between N-gram novelty and readability.

	Pearson	Spearman
GPT-4 (only label)	0.41	0.33
GPT-4 (label&sentences)	0.46	0.30
Claude-3 (only label)	0.36	0.36
Claude-3 (label&sentences)	0.43	0.42
Gemini-1.5 (only label)	0.31	0.33
Gemini-1.5 (label&sentences)	0.40	0.34
Llama-3 (only label)	0.16	0.04
Llama-3 (label&sentences)	0.34	0.27
QAFactEval	0.32	0.36
QuestEval	0.13	0.22
SummaCZS	0.29	0.28
SummaCConv	0.36	0.39
AlignScore	0.54	0.37

Table 9: Performance of faithfulness evaluators at the summary level. "Pearson" and "Spearman" represent Pearson correlation and Spearman rank correlation coefficients with human judgments.

support evidence is in parenthesis. We obtained the support evidence from the generated text by extracting the content in parentheses.

One thing worth mentioning is that the content might include ellipsis, i.e., "...", indicating that there is an omission in the generated text. We expanded such omission back to its original content by explicitly matching the beginning and the end of the given response.

H Analysis of Annotated Supporting Sentences

Table 11 shows the examples of the annotation of supporting sentences where multiple sentences can solely support the summary sentence and annotators overlook the others.

Llama-3	<p>Yes, the summary is supported by the source. The following sentences from the Background section support the summary:</p> <p>“Bochdalek hernia is a diaphragmatic hernia usually diagnosed during the neonatal period. It typically occurs in the left hemi-diaphragm and presents with severe respiratory and circulatory compromise.”</p> <p>“Adult Bochdalek hernia is rare, and most are also found on the left side of the diaphragm because the right pleuroperitoneal canal closes earlier and the liver buttresses the right diaphragm, minimizing the opportunity for herniation into the right thoracic cavity [1][2][3].</p> <hr/> <p>Yes, the summary is supported by the source. The following sentences from the Introduction section support the claim that the Sumida watercress farm provides educational opportunities, aesthetic beauty, and a sense of heritage:</p> <p>“Previous research has demonstrated that, in some areas, small farms are rarely economically viable by crop production alone and that many successful small farms rely on grant and other revenue streams based on diverse benefits (e.g. aesthetic value; farm experience) provided by these systems [14][15][16].”</p> <p>“Specifically, we utilized mixed methods (Table 1) including trend analysis, field water quality and microbial sampling, and semi-structured interviews to address the following research questions:</p> <p>... 3. What additional socio-cultural benefits are provided by the farm to the Sumida family and surrounding community?”</p>
GPT-4	<p>Yes</p> <p>Supporting sentences from the Source:</p> <ul style="list-style-type: none"> - “Bochdalek hernias are one of the most common types of diaphragmatic hernia, with most cases diagnosed during the neonatal period.” - “In contrast, diagnosis of a Bochdalek hernia in an adult is rare and is typically observed on the left side of the diaphragm.” <hr/> <p>Yes</p> <p>Supporting sentences from the Source:</p> <ol style="list-style-type: none"> 1. “We describe a case of an 89-year-old female who presented with abdominal pain...which led to the diagnosis of a strangulated diaphragmatic hernia.” 2. “The patient underwent emergent laparotomy and required a partial resection of the necrotic ileum and a hernia repair with direct closure.” 3. “Considering the severity of this condition, accurate diagnosis and proper treatment are needed.” 4. “A tailored operative approach is required on an individual case basis.”
Gemini-1.5	<p>Yes</p> <p>Supporting sentences: “Even more rare is the diagnosis of a right-sided Bochdalek hernia in an adult, where there is concurrent visceral malformation in most cases... However, as a corollary to this rationale, organ malformation consisting primarily of hepatic atrophy can increase the risk of a right-sided diaphragmatic hernia. Few of these rare cases have been reported.”</p> <hr/> <p>Yes</p> <p>Supporting sentences:</p> <p>* **Adult Bochdalek hernia is rare... **</p> <p>* **Considering the severity of this condition, accurate diagnosis and proper treatment are needed. **</p>
Llama-3	<p>Answer: Yes</p> <p>Supporting sentences:</p> <p>“An adult Bochdalek hernia is usually precipitated by a state of increased intra-abdominal pressure, such as that induced by pregnancy or from prolonged operations under pneumoperitoneum [4]. Thus, as the number of laparoscopic-assisted operations increases, it is important to consider a Bochdalek hernia as a potential intraoperative or postoperative complication. Iatrogenic weakness of the diaphragm that occurs as a result of abdominal surgery, such as hepatectomy, esophagogastrectomy, or transthoracic hiatal hernia repairs, may also cause an adult Bochdalek hernia [5,6].”</p> <hr/> <p>These sentences support the summary by stating that adult Bochdalek hernias can be caused by increased abdominal pressure, previous surgeries, or a pre-existing weakness in the diaphragm.</p> <hr/> <p>Answer: Yes</p> <p>Supporting sentences: 11Finally, the core/shell NPs were loaded with a chemotherapeutic agent doxorubicin (DOX)... the cytotoxic effects of the different core-shell Ag/polymeric NPs-based combinatorial therapeutics were tested in-vitro against breast cancer (MCF-7) and human fibroblast (1BR hTERT) cell lines.”</p> <p>Note: The Summary is supported by the Source because it accurately reflects the content of the Source, specifically the loading of the NPs with DOX and the testing of their cytotoxic effects against breast cancer and human fibroblast cell lines.</p>

Table 10: Different types of responses for extracting support evidence.

Source Article #1	Title: Mechanisms regulating resistance to inhibitors of topoisomerase II
	Authors: Ram Ganapathi, [...]
	Abstract: Inhibitors of topoisomerase II (topo II) are clinically effective in the management of hematological malignancies and solid tumors. Resistance to the various classes of anti-tumor agents (Curt et al., 1984) has been suggested to involve reduced drug accumulation and/or retention, conformational changes and/or over production of the target enzyme, and reduced activation and/or increased catabolism of drug. Doxorubicin (DOX) is a clinically effective anti-tumor agent against a spectrum of neoplastic diseases (Carter, 1975; Myers and Chabner, 1990). Although DOX is an inhibitor of topoisomerase II (topo II), multifactorial mechanisms are involved in the cytotoxic response (Siegfried et al., 1985; Louie et al., 1986; Bhushan et al., 1989; Doroshow et al., 1990). [...]
	Introduction: The emergence of drug-resistant tumor cells continues to be a major problem confronting advances in cancer chemotherapy. Resistance to the various classes of anti-tumor agents (Curt et al., 1984) has been suggested to involve reduced drug accumulation and/or retention, conformational changes and/or over production of the target enzyme, and reduced activation and/or increased catabolism of drug. Doxorubicin (DOX) is a clinically effective anti-tumor agent against a spectrum of neoplastic diseases (Carter, 1975; Myers and Chabner, 1990). Although DOX is an inhibitor of topoisomerase II (topo II), multifactorial mechanisms are involved in the cytotoxic response (Siegfried et al., 1985; Louie et al., 1986; Bhushan et al., 1989; Doroshow et al., 1990). [...]
Summary Sentence #1	Cancer cells can develop resistance to certain chemotherapy drugs, such as topoisomerase II inhibitors, which are used to treat various types of cancer.
Source Article #2	Title: miR-135 family members mediate podocyte injury through the activation of Wnt/ β -catenin signaling
	Authors: Xianggui Yang, [...]
	Abstract: [...] The ectopic expression of miR-135a and miR-135b led to severe podocyte injury and the disorder of the podocyte cytoskeleton. Our findings demonstrated that miR-135a and miR-135b activated Wnt/ β -catenin signaling and induced the nuclear translocation of β -catenin. Using luciferase reporter assays, reverse transcription-quantitative polymerase chain reaction (RT-qPCR) and western blot analysis, glycogen synthase kinase 3 β (GSK3 β) was identified as a target gene of miR-135a and miR-135b. To the best of our knowledge, this is the first study to demonstrate that members of the miR-135 family (specifically miR-135a and miR-135b) regulate the expression of GSK3 β , thus playing a role in the development of podocyte injury and the disorder of the podocyte cytoskeleton. This is an important finding as it may contribute to the development of novel therapeutics for podocyte injury-associated glomerulopathies.
	Introduction: [...] In the present study, we aimed to determine the roles and mechanisms of action of miR-135a and miR-135b in podocyte injury, and to elucidate the mechanisms underlying podocyte injury. We found that miR-135a and miR-135b were overexpressed in patients with FSGS and in models of podocyte injury, and that the ectopic expression of these miRNAs promoted podocyte injury by activating Wnt/ β -catenin signaling through the suppression of glycogen synthase kinase 3 β (GSK3 β) expression. Our findings demonstrate that miR-135a and miR-135b play an important role in podocyte injury. Our findings may provide new insight into the understanding of the molecular mechanisms underlying podocyte injury, which may be crucial for the development of novel therapeutic agents for the treatment of podocytopathy.
Summary Sentence #2	Overall, the study suggests that miR-135a and miR-135b play a role in podocyte injury and may be potential targets for developing new treatments for kidney diseases.

Table 11: Examples of supporting sentences annotated by the two annotators. Different colors represent different annotations. Multiple sentences can solely support the summary sentence and the annotators annotated different supporting sentences.