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

**Biaoyan Fang** and **Xiang Dai** and **Sarvnaz Karimi**
CSIRO Data61
Sydney, Australia
{byron.fang;dai.dai;sarvnaz.karimi}@csiro.au

## Abstract

Generating plain biomedical summaries 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 **Fa**ithfulness and **Re**asoning on plain **Bio**medical Summaries. This dataset consists of 175 plain summaries (1,445 sentences) generated by seven different LLMs, paired with source articles. Using our dataset, we identify the performance gap of LLMs in generating faithful plain biomedical summaries and observe a negative correlation between abstractiveness and faithfulness. We also show that current faithfulness evaluation metrics do not work well in the biomedical domain and confirm the over-confident tendency of LLMs as faithfulness evaluators. To better understand the faithfulness judgements, we further benchmark LLMs in retrieving supporting evidence and show the gap of LLMs in reasoning faithfulness evaluation at different abstractiveness levels. 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]

## 1 Introduction

Generating plain text summaries—summarising 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 performance achieved by LLMs (Jahan et al., 2024; Guo et al., 2024; Sim et al., 2023), the *faithfulness*, which is to what extent the generated text is consistent with the source articles, and *factuality*, that is to what extent the generated text is consistent with the external knowledge, have been known problems of LLMs (Pagnoni et al., 2021; Ji et al., 2023) and have not been well-studied in the biomedical domain (Joseph et al., 2024).

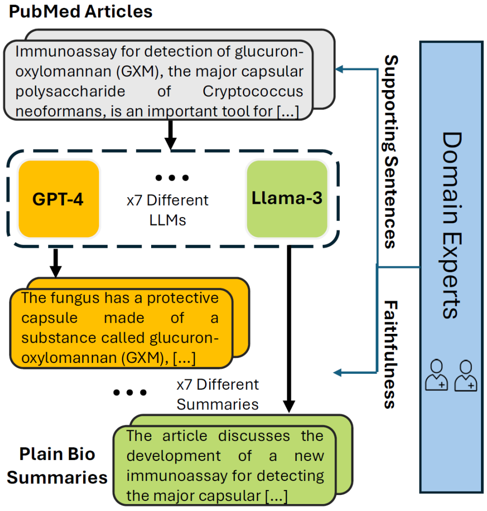

PubMed Articles
Immunoassay for detection of glucuron-oxylomannan (GXM), the major capsular polysaccharide of Cryptococcus neoformans, is an important tool for [...]
GPT-4
x7 Different LLMs
Llama-3
The fungus has a protective capsule made of a substance called glucuron-oxylomannan (GXM), [...]
x7 Different Summaries
Plain Bio Summaries
The article discusses the development of a new immunoassay for detecting the major capsular [...]
Supporting Sentences
Faithfulness
Domain Experts

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

Apart from labor-intensive and costly manual examination, prior work (Scialom et al., 2021; Laban et al., 2022; Zha et al., 2023) has proposed various metrics to automatically evaluate the faithfulness of the generated text. However, these metrics are mainly designed to measure faithfulness in the general domain, such as news. To what extent it can be used in domain-specific areas, such as 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 alignments with human judgment, they do not always provide correct reasoning for their decisions, e.g., correct rationales or supporting sentences from text. Examining to what extent LLMs can provide correct reasoning for their decisions could help better understand the reasoning behind LLMs, especially in the biomedi-

[1] Dataset available at https://data.csiro.au/.

*Findings of the Association for Computational Linguistics: EACL 2024*, pages 9890–9911
November 12-16, 2024 ©2024 Association for Computational Linguistics

cal domain where accurate evidence is paramount.

To address these problems, we propose a benchmark dataset, FAREBIO, on evaluating **Fa**ithfulness and **Re**asoning of LLMs on plain **Bio**medical 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 LLM-based summarisation systems and highlight the corresponding supporting sentences from source articles.

In Section 4, we answer the four major research questions on the faithfulness evaluation of 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 correlate with faithfulness? (3) How do current automatic faithfulness evaluators align with human judgment? (4) Do LLMs consider their own generation more faithful than other LLMs in zero-shot evaluation?

We further evaluate capability of LLMs in providing correct reasoning for their judgment in Section 5. Specifically, we focus on the reasoning as extracting the supporting sentences and address three major research questions: (5) Can LLMs identify the supporting sentences from the source article when they are instructed to evaluate the faithfulness of generate summaries? (6) How 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 with the identification of supporting sentences for plain biomedical summaries. We find that generated summaries demonstrate a high degree of non-faithful hallucination and the level of abstractiveness shows a negative correlation with the faithfulness of plain summaries. Current faithfulness evaluators tailored for the general domain do not directly transfer well to the biomedical domain. We also observe a tendency where LLMs, as evaluators, favour their generation when evaluating faithfulness. However, the construction of the prompt could also impact such a tendency. Additionally, we identify the performance gap of LLMs in retrieving supporting sentences from source articles, either for high or low abstractiveness summaries. We hope our dataset would enable further studies on understanding the faithfulness evaluation of LLMs as well as its reasoning.

## 2 Related Work

Faithfulness, where the generated summary is consistent with the source (Maynez et al., 2020), has been 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 evaluated faithfulness and factuality—factual consistency with the source and external knowledge, respectively—in the biomedical domain. For instance, Ramprasad et al. (2024) measured the factuality of zero-shot summaries from GPT-3.5 (Brown et al., 2020) and Flan-T5-XL (Chung et al., 2024). 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). Guo et al. (2023) proposed an evaluation framework for plain language summarisation and measured the faithfulness of GPT-based text simplification on biomedical abstracts.

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), utilising 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) Learning-based faithfulness evaluation (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 faithfulness evaluators.

Apart from solely evaluating the binary faithfulness label of the generated summary, a natural extension 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 claims based on Wikipedia articles and provided extracted facts from 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. Similarly, FACTPICO (Joseph et al., 2024) contains expert-written rationales for factuality annotations. Ghosal et al. (2024) proposed a shared task in identifying all grounding context from the scholarly paper discussing methodological details in the claim.[2]

## 3 Dataset Creation

### 3.1 Model Selection

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

**GPT-4** (Achiam et al., 2023), a large closed-source model developed by Open AI. The GPT family is adopted in various NLP tasks including summarisation (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 closed-source model developed by Anthropic. It has been seen to outperform the GPT family in certain tasks, e.g., open-domain conversations (Lin and Chen, 2023) and reading comprehension tests (Kuo et al., 2024). We use `claude-3-sonnet` in our experiments.

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

**Llama-3** (Meta, 2024), an 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., 2024; 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 summarisation tasks (Sim et al., 2023; Alqahtani et al., 2023). we investigate `Flan-T5-XL` for our plain biomedical summarisation task.

**Finetuned-Llama-3** To investigate the impact of customising the plain summary for a specific type, we further finetune `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 finetune `Flan-T5-XL` on the PLOS dataset and investigate the faithfulness of the generated summaries from this finetuned model.

For the selected models, we use them to generate the plain summary based on the source article, detailed in Section 3.2. Specifically, we follow the previous work (Goldsack et al., 2022; Cardenas et al., 2023) on generating plain summaries for scientific literature, considering the title, authors, abstract, and the first section of the content as the input.[4]

### 3.2 Annotation Data

To generate plain biomedical summaries, we obtained English 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[6] that (1) were published in PubMed; (2) were published no later than 2010, ensuring more recent PubMed articles are included; and (3) contained metadata of title, authors, abstract, and full content. For each selected article, as shown in Figure 1, we then generate 7 different plain summaries using various types of LLMs (Section 3.1), resulting in 175 summaries.

To provide a more fine-grained level of faithfulness analysis, we segment the generated summaries into sentences and ask annotators to annotate the faithfulness of generated summaries at the sentence level (Section 3.3). Note that due to the imperfection of the off-the-shelf segmentation tool,[7] segmented text could result in only a partial segment

[2] https://github.com/oasisresearchlab/context24

[3] Details of finetuning `Llama-3-8B-Instruction` and `Flan-T5-XL` on the PLOS dataset are in Appendix A. The choice of PLOS over the eLife dataset is provided in Appendix B.

[4] Detailed prompt constructions are in Appendix C.

[5] https://api.semanticscholar.org/api-docs/graph

[6] The MeSH classifications of the selected articles are provided in Appendix K.

[7] We used Spacy sentencizer with "en_core_web_sm". https://spacy.io/api/sentencizer

|  | **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 |  |
| 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.

of a sentence. To address this, we filter out the segmented 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 process 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 faithful, highlight the part that is not consistent in the summary sentence.[8]

Additionally, in line with the previous literature (Maynez et al., 2020; Ramprasad et al., 2024), we also ask annotators to flag the sentences that are factually hallucinated. That is, the generated sentence is supported by external knowledge but not by the source article. This helps better understand to what extent external knowledge is utilised to generate plain summaries as the task requires plain explanations of technical terms.

We recruit two medical doctors via Upwork.[9] Specifically, before the annotators started the annotation separately, we started with the annotation training by giving the two annotators 6 summaries (34 sentences) generated from different LLMs based on different source articles. We calculate the inter-annotator agreement (IAA) at the sentence level, i.e., binary faithfulness labels, and at the summary level, considering the summary as faithful if all sentences are annotated as faithful. We achieve a percentage agreement of 0.94 and 0.83 and Cohen's Kappa (McHugh, 2012) of 0.48 and 0.57 at the sentence and summary level, respectively. Similar to the observations from previous work (Ramprasad et al., 2024; Joseph et al., 2024), faithfulness annotation is imbalanced, e.g., on average 2 out of 34 sentences in those 6 summaries are hallucinated, resulting in an expected higher percentage agreement and lower Cohen's kappa score.

We further calculate the IAA on annotated supporting sentences based on the subset that both annotators consider the generated sentence to be faithful. Specifically, we consider the agreement where both annotators highlight the same supporting sentences, resulting in Cohen's Kappa of 0.49 and Precision, Recall, and F1 of 0.47, 0.56, and 0.51, respectively. Despite the challenge of understanding and finding supporting sentences in biomedical literature (Van Auken et al., 2014), one possible reason for such agreement is that we did not ask annotators to highlight all related supporting sentences. Multiple sentences from the source article could solely support the summary sentence. Annotators might overlook other supporting sentences once they find one.[10]

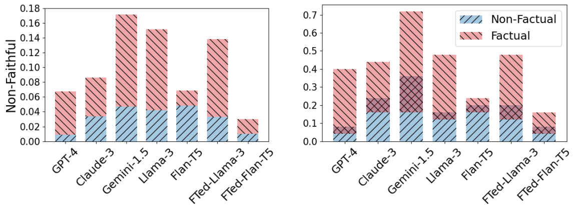
## 4 Faithfulness Evaluation and Analysis on Plain Biomedical Summaries

**RQ1. How faithful are generated summaries across different LLMs?** Figure 2 shows the non-faithful hallucination across selected LLMs. As we use summarisers to generate plain summaries, external domain knowledge might be required to explain technical concepts in simple terms. We label the information where it is correct but cannot be attributed to the source article as factual hallucination (Cao et al., 2022; Li et al., 2024). Specifically, we break down the non-faithful hallucination into two categories: (1) non-factual and (2) factual hallucinations.

Overall, across different LLMs, we observe the non-faithful hallucination rate (blue and forward-slashed bars + red and backward-slashed bars) ranging from 3% to 17% and from 16% to 72% at the sentence and summary level, respectively, indicating the performance variances of LLMs in

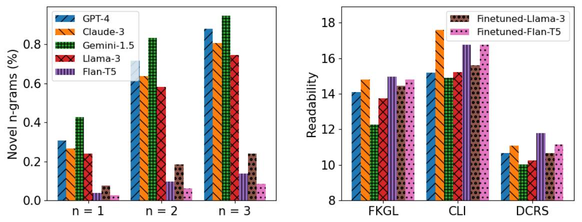
[8] Annotation details are provided in Appendix G.
[9] https://www.upwork.com/

[10] More detailed analysis of the annotation of supporting sentences are provided in Appendix I.



(a) Sentence Level (b) Summary level

Figure 2: Non-faithful hallucination across models at sentence and summary levels. We break down the non-faithful hallucination into Non-Factual (blue and forward-slashed bars) and Factual (red and backward-slashed bars) hallucinations. The overlap of non-factual and factual at the summary level indicates the percentage of summaries that have both non-factual and factual hallucinated sentences.



(a) Abstractiveness ↑ (b) Readability ↓

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

generating faithful plain biomedical summaries.

Comparing the non-factual and factual hallucinations (blue and forward-slashed bars vs., red and backward-slashed bars), most LLMs show a high rate of non-faithful but factual hallucination at both sentence and summary levels. Flan-T5 and its finetuned version, on the contrary, show a low factual hallucination rate. One possible reason is due to the different levels of abstractiveness in generated summaries. That is, the generated summary is inherently more factual if only extracting sentences from the source article instead of providing plain summaries (Ladhak et al., 2022). We further investigate this in the following research question.

For non-faithful and non-factual hallucinations (blue and forward-slashed bars), i.e., the generated text is neither faithful nor factual, we observe a small hallucination rate across all models at the sentence level, with less than 5% of sentences that are hallucinated. However, we observe a higher rate at the summary level, with at least 8% of summaries containing hallucination (i.e., at least 2 out of 25 summaries), indicating the performance gap in generating faithful and factual plain biomedical summaries.

**RQ2. How abstractive and readable are the plain biomedical summaries, and how do they relate to faithfulness?** As discussed in Section 1, plain summaries from biomedical articles might incorporate external knowledge, e.g., explaining jargon, to make it more readable for general audiences (Goldsack et al., 2022). This could affect the summary's abstractiveness and potentially introduce more hallucinations. To measure the abstractiveness of the summary, we compare the *n*-gram novelty (See et al., 2017; 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 Gemeini-1.5, and the open-source model Llama-3 demonstrate high abstractiveness. Interestingly, although plain 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 further calculate the Spearman *r* correlation between *n*-grams novelty and the ratio of non-faithful hallucination in summary, i.e., the percentage of non-faithful sentences in summary. We separate the correlation based on two groups: (1) non-factual and (2) factual hallucinations.[11] 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 of the factors impacting non-factual hallucination. Also, weak to moderate 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

[11] Detailed correlations between abstractiveness and readability with faithfulness are shown in Appendix E.

|  | Agreement |  | Prediction Performance |  |
|---|---|---|---|---|
|  | Cohen Kappa ↑ | P. Agreement (%) ↑ | Recall ↑ | 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 automatic faithfulness evaluators at the sentence level. We report the annotation correlation between human experts and automatic evaluators. “P. Agreement (%)" represents the percentage agreement. “Pred. Non-faithful (%)" represents the percentage of non-faithful instances predicted by automatic evaluators.

given text, by employing experimental formulas on the number of characters, words, and sentences.

As shown in Figure 3b, We observe that Gemini-1.5 and Llama-3 show a lower readability score, that is, generate more readable summaries across the three metrics. We investigate the correlation between abstractiveness and readability and observe a Spearman *r* correlation of at least -0.31 ($p<0.05$) (Appendix E) 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.

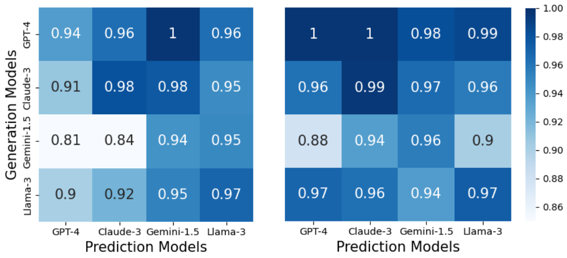
**RQ3. To what extent do current automatic faithfulness evaluators align with human judgment in plain biomedical summarisation?** We compare the human annotations with different types of automatic faithfulness evaluators. We consider two QA-based faithfulness metrics, Questeval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2022) which utilise 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 zero-shot 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*, prompting LLMs to provide only faithfulness labels; and (2) *label & sentences*, prompting LLMs to provide faithfulness labels and supporting sentences from the source (Section 5).[12] We exclude using Flan-T5 as an evaluator because it cannot produce meaningful results from our prompt.

In Table 2, we measure the faithfulness evaluation agreement between automatic evaluation metrics and human judgment at the sentence level, considering non-faithful hallucination from factual and non-factual hallucinations (Figure 2).

We observe a performance gap in improving faithfulness evaluation agreements with human annotation across all automatic metrics. Specifically, current metrics trained on the general domain (i.e., Questeval, QAFactEval, Summac, and AlignScore), do not achieve strong agreement with human annotations, indicating the difficulty of directly transferring those metrics to the biomedical domain. One possible reason is the lack of domain knowledge. For instance, QA-based metrics rely on the extracted named entities to generate ques-

[12] Prompt constructions are provided in Appendix D.



(a) *only label*

(b) *label&sentences*

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

tions and answers, where the extraction framework mostly focuses on general named entities (e.g., person and location), instead of domain-specific categories (e.g., disease and species) (Lee et al., 2020). 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 non-faithfulness sentences annotated by the annotators (Adams et al., 2023b), i.e., Recall. Although QAFactEval, Questeval, and Summac achieved high recall (Table 2, Column 4), they predicted more than 60% of sentences as non-faithful, making it impractical considering overall only 11% of sentences are labeled as non-faithful. Among other metrics, we observe a low recall in identifying non-faithful sentences, indicating the gap in this direction.

We also aggregate annotations at the sentence level to the summary level and evaluate the Pearson and Spearman $r$ correlations (Appendix F). We observe a similar performance gap in the alignment between current automatic faithfulness evaluators and human judgment.

**RQ4. Do LLMs consider their generation more faithful than other LLMs in zero-shot evaluation?** Inspired by previous work (Tam et al., 2023; Panickssery et al., 2024), which shows LLMs tend to be over-confident in evaluating their own generated text in the general domain, we investigate where this holds in our faithfulness evaluation of plain biomedical summaries. Specifically, we consider a subset of the annotations, only focusing on the summaries from GPT-4, Claude-3, Gemini-1.5, and Llama-3.

| | **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 sentence identification.

Figure 4 shows the heatmap of predicted faithfulness percentage across selected LLMs. We find that although LLMs rate most of their generations as faithful, there is no strong evidence showing they tend to favour their generation over those generated by other models. In addition, we find that the different construction of prompts would also impact the model tendency regarding their faithfulness evaluation. For example, when evaluating summaries generated by Gemini-1.5 and Llama-3, using the prompts of returning *label&sentences* tends to rate higher than returning *only label*.

## 5 Supporting Sentence 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 a subset of our dataset where sentences are labeled as faithful and supporting evidence is provided. The statistic of this subset is shown in Table 3.

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.[13]

It is worth mentioning that, as discussed in Section 3.3, annotators might overlook the supporting evidence. Although we provide Precision, Recall, and F1 on extracted sentence matching, we focus

[13] Additional experiment on Okapi BM25 Top $k$, where $k \in \{1, 2, 3\}$, is provided in Appendix J.

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.[14] 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 source articles. Additionally, we observe that models have higher Recall compared to Precision. This might be due to the incomprehensive annotation of the supporting evidence from annotators (Section 3.3).

We further perform an error analysis on the extracted supporting evidence from LLMs. 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. These errors could contribute to the low precision in the supporting evidence extraction from LLMs.

**RQ6. Does abstractiveness impact the identification of supporting evidence?** As shown in Figure 3a, summaries from different LLMs demonstrate different levels of abstractiveness. 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: (1) High abstractiveness, i.e., GPT-4, Claude-3, Gemini-1.5, and Llama-3; and (2) Low abstractiveness, i.e., Flan-T5, Finetuned-LLama-3, and Finetuned-Flan-T5.

Table 4 shows the performance on different levels of abstractiveness. Compared to low abstractiveness, we observe a consistent performance drop in the high abstractiveness subset across all models, indicating the impact of abstractiveness and the difficulty in identifying supporting evidence from high abstractive summaries. Okapi BM25 achieves the best performance in low abstractiveness summaries but it suffers when the abstractiveness of summaries increases. LLMs achieve high recall in identifying supporting sentences, with either high or low abstractiveness. Specifically, GPT-4 achieves a Recall of 0.72 and 0.85 in the high and low abstractiveness, respectively, indicating its potential in identifying supporting evidence.

| | **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 Abstractiveness | | |
| 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 Abstractiveness | | |
| 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: Performance on supporting sentence identification across different models. Overall is separated into two subsets of high and low abstractiveness.

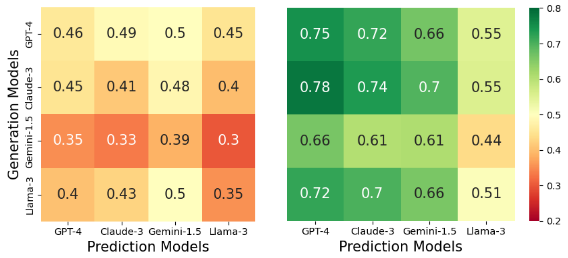
**RQ7. Do LLMs perform better when extracting evidence for their generated summaries?** That is, assuming LLMs generate summaries based on their reasoning, would LLMs be able to retrieve evidence for their generation, as they might follow a similar reasoning process?

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.

[14] The post-processing of the identified sentences from LLMs are shown in Appendix H.

| | |
|---|---|
| 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.ConclusionThe 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:<br>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<br>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<br>It has potential to enhance optometry training by offering flexible, accessible international learning experiences. |
| Extraction #4 | Error: Irrelevant Sentences<br>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.



(a) Precision

(b) Recall

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 observe a positive correlation between abstractiveness and non-faithful hallucination and find that the construction of prompts could also affect faithfulness prediction preferences. We further benchmark the capability of LLMs in retrieving supporting sentences for the plain summaries at different levels of abstractiveness.

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. We expect our research would enable further studies on understanding the faithfulness evaluation of LLMs in its reasoning.

## 7 Limitations

Our dataset provides the faithfulness annotation at the sentence level, with 1445 sentences from 175 generated summaries (25 documents × 7 different LLMs). Having a larger size of annotation would further enhance our analysis. However, one of the main challenges in benchmarking the faithfulness of plain biomedical summaries is the cost involved in hiring highly skilled domain experts. For our annotation, we hire two medical doctors at $50 USD/hr. Although we facilitated the annotation by grouping summaries that are generated based on the same source article, minimising 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 the scalability of annotations challenging.

Another challenge of the faithfulness annotation in the biomedical domain is to understand the generated hallucination. Our IAA of faithfulness evaluation aligns with the previous work (Ramprasad et al., 2024). We also ask the annotators to highlight the inconsistent part from the summary and provide a brief rationale for their judgement. These could be used for further categorising and analysing the errors of LLMs in generating plain biomedical summaries. Combining the analysis on faithful and non-faithful sentences could further provide a complementary understanding of summary generation in the biomedical domain.

We follow the previous work (Goldsack et al., 2022; Cardenas et al., 2023), considering the title, authors, abstract, and the first section of the literature as the input to generate plain summaries for scientific literature. To cover comprehensive information from the literature and improve the generalization ability of the faithfulness study, utilising the full article as the input would be an important direction to explore.

For our supporting sentence 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 overlook from 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 sentence identification. Additionally, IR-based metrics (Manning, 2008) such as precision at $k$ or mean average precision can be employed to further understand the reasoning capability of LLMs.

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 supporting 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 it is 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 studies. 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 supporting evidence is customising different prompts for different LLMs and employing more advanced prompt engineering methods, e.g., automatic prompt generation (Ha et al., 2023; Zhou et al., 2023; Li and Liang, 2021) and Chain-of-Thought (CoT) prompts (Kojima et al., 2022; Yu et al., 2023).

Based on our annotated dataset, we show the poor transformability of current faithfulness metrics to the biomedical domain. Developing a reliable faithfulness evaluator is critical in generating faithful summaries for the biomedical domain. Approaches of how to transfer general faithfulness evaluators to biomedical domains, such as finetuning the off-the-shelf evaluators on our annotations, would be worth exploring.

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

Our study was reviewed and approved by the Health and Medical Human Research Ethics Committee at CSIRO, under the category of low risk (2024_029_LR).

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 in their residence.

For copyright, we obtained the PubMed article from S2ORC (Lo et al., 2020), which is under the licence ODC-By 1.0.[15] OpenAI provides Terms of Use[16] for the usage of GPT-4. Anthropic provides the Consumer Terms of Service for Claude-3.[17] Gemini-1.5 follows the Google Generative AI terms [18] Llama-3 is under licence “META LLAMA 3 COMMUNITY LICENSE AGREEMENT".[19] Flan-T5 is under licence “Apache License 2.0".[20]

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 rea-

[15] https://opendatacommons.org/licenses/by/1-0/

[16] https://openai.com/policies/terms-of-use/

[17] https://www.anthropic.com/legal/consumer-terms

[18] https://policies.google.com/terms/generative-ai/use-policy

[19] https://llama.meta.com/llama3/license/

[20] https://choosealicense.com/licenses/apache-2.0/

soning 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 and not commercial usage. Additionally, The PubMed articles might contain authors' information and associated affiliations. We discourage the use of this information to target individuals.

## References

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

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023a. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Griffin Adams, Jason Zuckerg, and Noémie Elhadad. 2023b. A Meta-Evaluation of Faithfulness Metrics for Long-Form Hospital-Course Summarization. In *Machine Learning for Healthcare Conference*, pages 2–30.

Amal Alqahtani, Rana Salama, Mona Diab, and Abdou Youssef. 2023. Care4Lang at MEDIQA-chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 524–528, Toronto, Canada. Association for Computational Linguistics.

Anthropic. 2024. Meet Claude. Accessed on 31 May 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1202, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25:1–53.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the*

*North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Tirthankar Ghosal, Philipp Mayr, Anita de Waard, Aakanksha Naik, Shannon Shen, Amanpreet Singh, Orion Weller, Yanxia Qin, and Yoonjoo Lee, editors. 2024. *The 4th Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Bangkok, Thailand.

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 *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2023. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. *arXiv preprint arXiv:2305.14341*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2433–2445, Singapore. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2311.05232*.

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. *Computers in Biology and Medicine*, page 108189.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).

Sebastian Antony Joseph, Lily Chen, Jan Trienes, Hannah Louisa Göke, Monika Coers, Wei Xu, Byron C Wallace, and Junyi Jessy Li. 2024. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. In *The 62nd Annual Meeting of the Association for Computational Linguistics*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Bor-Chen Kuo, Pei-Chen Wu, and Chen-Huei Liao. 2024. GPT-3.5, GPT-4, Bard, and Claude's Performance on the Chinese Reading Comprehension Test. In *Joint Proceedings of LAK 2024 Workshops, co-located with 14th International Conference on Learning Analytics and Knowledge*, Kyoto, Japan.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Barrett Lattimer, Patrick CHen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1703, Singapore. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10879–10899.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *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)*, pages 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Christopher D Manning. 2008. Introduction to information retrieval.

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

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

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed on 05 31, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv preprint arXiv:2404.13076*.

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. *Journal of machine learning research*, 21(140):1–67.

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

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *TREC*, Gaithersburg, MD, US.

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. *arXiv preprint arXiv:2404.18416*.

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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and Sarvnaz Karimi. 2023. CSIRO Data61 team at BioLaySumm task 1: Lay summarisation of biomedical research articles using generative models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 629–635, Toronto, Canada. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

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

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium, ADCS 2014, Melbourne, VIC, Australia, November 27-28, 2014*, page 58. ACM.

Kimberly Van Auken, Mary L Schaeffer, Peter McQuilton, Stanley JF Laulederkind, Donghui Li, Shur-Jen Wang, G Thomas Hayman, Susan Tweedie, Cecilia N Arighi, James Done, et al. 2014. BC4GO: a full-text corpus for the BioCreative IV GO task. *Database*, 2014:bau074.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association 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 evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, 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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via ChatGPT for faithful summary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

## A Finetuing on the PLOS Dataset

We follow the instructions on Appendix C to fintune Llama-3 and Flan-T5 on the PLOS training dataset (Goldsack et al., 2022).

For Llama-3, we follow the hyper-parameters from *llama-recipes*[21] and finetune `Llama-3-8B-Instruction` on the PLOS training set for 10 epochs. The model is finetuned on 3 H100 GPUs for 5 hours. We select the checkpoint that has the best perplexity performance on the PLOS dev set.

For Flan-T5, we follow the approach in Sim et al. (2023) to finetune the model. We finetune the `Flan-T5-xl` (3B) model on the PLOS training set for 5 epochs and use a beam search decoder during inference—a beam width of four—to generate up to 386 tokens.

Table 6 shows the effectiveness of finetuned models on the PLOS test set.

## B Comparison of the PLOS and eLife Datasets

Goldsack et al. (2022) collected plain summaries from two sources, i.e., PLOS and eLife. For our model finetuning (Appendix A), we consider PLOS over eLife due to the following reasons: (1) Dataset size. Compared to the eLife dataset with 4,828 documents, the PLOS dataset contains a larger training set of 27,525 documents; (2) Diversity of plain summaries. The plain summaries from eLife and PLOS were written by editors and authors, respectively. Being written by various authors might potentially lead to more diverse styles of plain summaries; (3) Domain. eLife has a specific focus on biomedical and life sciences, whereas PLOS includes other areas of science and medicine, covering a wider range of topics; and, (4) Abstractiveness. Although plain summaries from eLife are more abstractive compared to PLOS, summaries from PLOS are also considered to be abstractive, with nearly 80% novelty in 3-grams.

## C Prompt Construction for Plain Summary Generation

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

[21] https://github.com/meta-llama/llama-recipes/tree/main/recipes/finetuning

| **Model** | 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: Effectiveness of finetuned models on the PLOS test set (R = average ROUGE F1-score).

*Summarize this article for non-experts:*

*Article:*

*Title: [Title]*

*Authors: [Authors]*

*Abstract: [Abstract]*

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

*Summary:*

where *[Title]*, *[Authors]*, and *[Abstract]* represent the content of the title, authors, and abstract, respectively. *[First section Name]* and *[First section context]* denote the name of the first section of the source article (e.g., *Introduction*) and the corresponding content, respectively.

## D Prompt Construction for Faithfulness Evaluation

To utilise LLMs as faithfulness evaluators, we adopt the evaluation prompt from prior work (Gekhman et al., 2023). Specifically, we use the following template for GPT-4, Claude-3, Gemini-1.5, and Llama-3:

*Source:*

*Title: [Title]*

*Authors: [Authors]*

*Abstract: [Abstract]*

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

*Summary: [Summary sentence]*

*[Evaluation prompt]*

where *[Title]*, *[Authors]*, *[Abstract]*, *[First section Name]*, and *[First section context]* are denoted as in Appendix C. *[Summary sentence]* represents the sentence from the generated summary. For cases where only prompting LLMs to return the faithfulness label, *[Evaluation prompt]* represent *Is the Summary supported by the Source? Answer using "Yes" or "No" only.*; For cases where prompting LLMs to return the faithfulness label and supporting sentences from the source, *[Evaluation prompt]* represent *Is the Summary supported by the Source? Answer using "Yes" or "No" and extract the supporting sentences from the Source..*

| | Non-factual | Factual |
|---|---|---|
| | Abstractiveness | |
| $n = 1$ | 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 (novel *n*-grams, i.e., "$n = k$" where $k \in \{1, 2, 3\}$) and Readability (i.e., FKGL, CLI, and DCRS) with the ratio of hallucination (i.e., non-factual and factual hallucinations) in summary.

| | 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 abstractiveness (novel *n*-grams, i.e, "$n = k$" where $k \in \{1, 2, 3\}$) and readability.

## E Correlations between Abstractiveness and Readability with Faithfulness

The Spearman *r* correlations between *n*-gram novelty and readability scores with faithfulness in shown in Table 7.

Table 8 shows the Spearman *r* correlations between *n*-gram novelty and readability.

## F Performance of Faithfulness Evaluators at the Summary Level

Table 9 shows the performance of faithfulness evaluators aligning with human judgments at the summary level.

## G Annotation Guidelines

### G.1 Annotators

We recruited two experienced medical doctors by advertising our task on the Upwork platform, a global platform connecting various experts for different tasks. Specifically, we hired the annotators

| Evaluator | 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: Effectiveness of faithfulness evaluators at the summary level. “Pearson" and “Spearman" represent Pearson correlation and Spearman *r* correlation with human judgments.

based on their profiles, where both had indicated that they are native English speakers and have many years of experience in writing medical articles. We only considered expertise as a factor and do not explicitly limit the recruitment to location or other factors. One annotator was based in the United States and the other was based in Germany. One was a board-certified physician and surgeon and the other was a medical doctor with a background in Epidemiology and Biostatistics.

We recorded the hours spent and managed the payment via the Upwork platform.

### G.2 Annotation Interface

We use Label Studio as our annotation platform.[22] We include screenshots of the annotation interface of our task in Figure 6, Figure 7, and Figure 8.

## H Post-processing of Extracted Sentences

We use the prompt from Appendix D to evaluate the capability of LLMs in extracting supporting sentences. As shown in the examples in Table 11, based on our observation, we find that almost all supporting evidence is in parenthesis. We obtained the supporting 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.

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Overall | | | |
| Okapi BM25 Top 1 | 0.73 | 0.56 | 0.63 |
| Okapi BM25 Top 2 | 0.44 | 0.67 | 0.53 |
| Okapi BM25 Top 3 | 0.32 | 0.74 | 0.45 |
| 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 Abstractiveness | | | |
| Okapi BM25 Top 1 | 0.61 | 0.41 | 0.49 |
| Okapi BM25 Top 2 | 0.41 | 0.56 | 0.47 |
| Okapi BM25 Top 3 | 0.32 | 0.65 | 0.43 |
| 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 Abstractiveness | | | |
| Okapi BM25 Top 1 | 0.94 | 0.90 | 0.92 |
| Okapi BM25 Top 2 | 0.49 | 0.94 | 0.65 |
| Okapi BM25 Top 3 | 0.33 | 0.97 | 0.50 |
| 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 10: Supporting sentence identification using information retrieval and LLM-based methods.

## I Analysis of Annotated Supporting Sentences

Table 12 shows the examples of the annotation of supporting sentences where multiple sentences can solely support the summary sentence and annotators overlook the others.

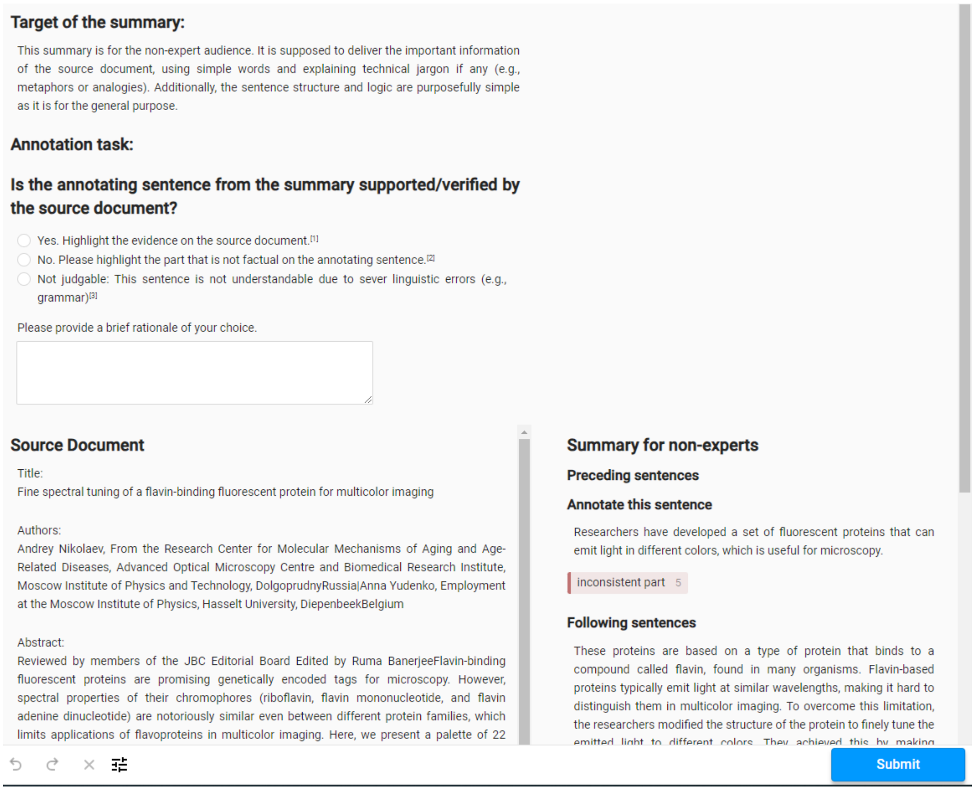
## J Additional Experiment on Support Sentence Identification

Table 10 shows the additional experiment results on the support sentence identification task. We consider the maximum number of *k* to be 3 based on the statistic that the average support sentences per instance are 1.31 (Table 3).

## K MeSH Classifications of the Dataset

Table 13 shows the MeSH classifications of the dataset from PubMed.

[22] https://labelstud.io/



**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 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

**Summary for non-experts**

**Preceding sentences**

**Annotate this sentence**

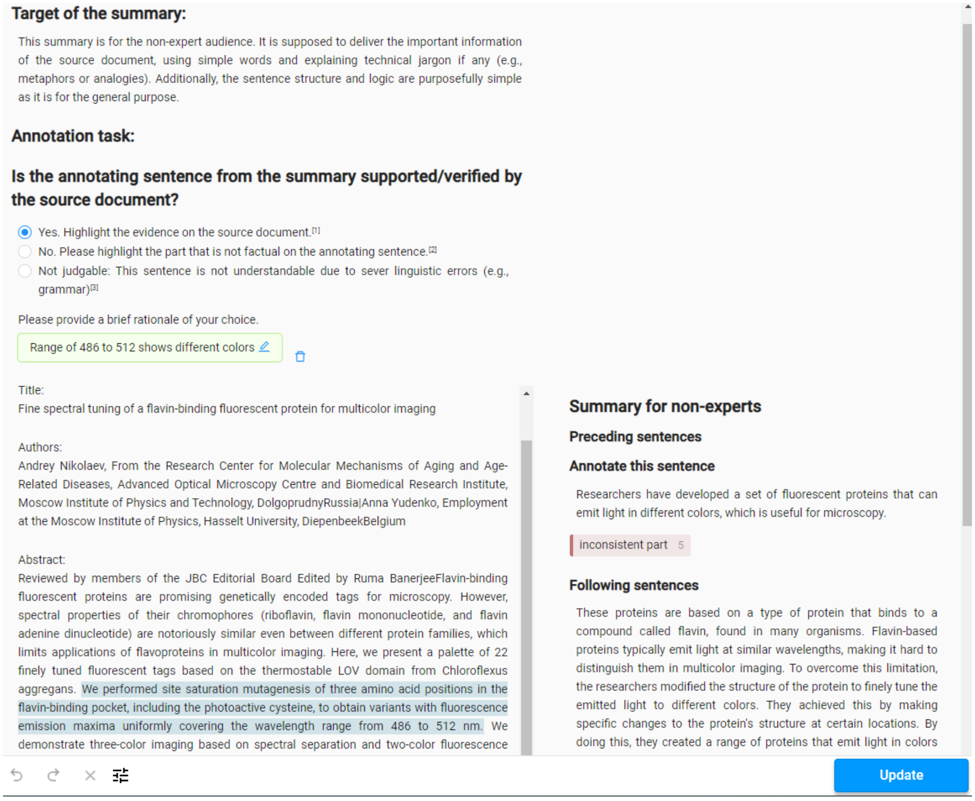
Researchers have developed a set of fluorescent proteins that can emit light in different colors, which is useful for microscopy.

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

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.

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 specific changes to the protein's structure at certain locations. By doing this, they created a range of proteins that emit light in colors

Figure 7: An example of the faithful annotation.

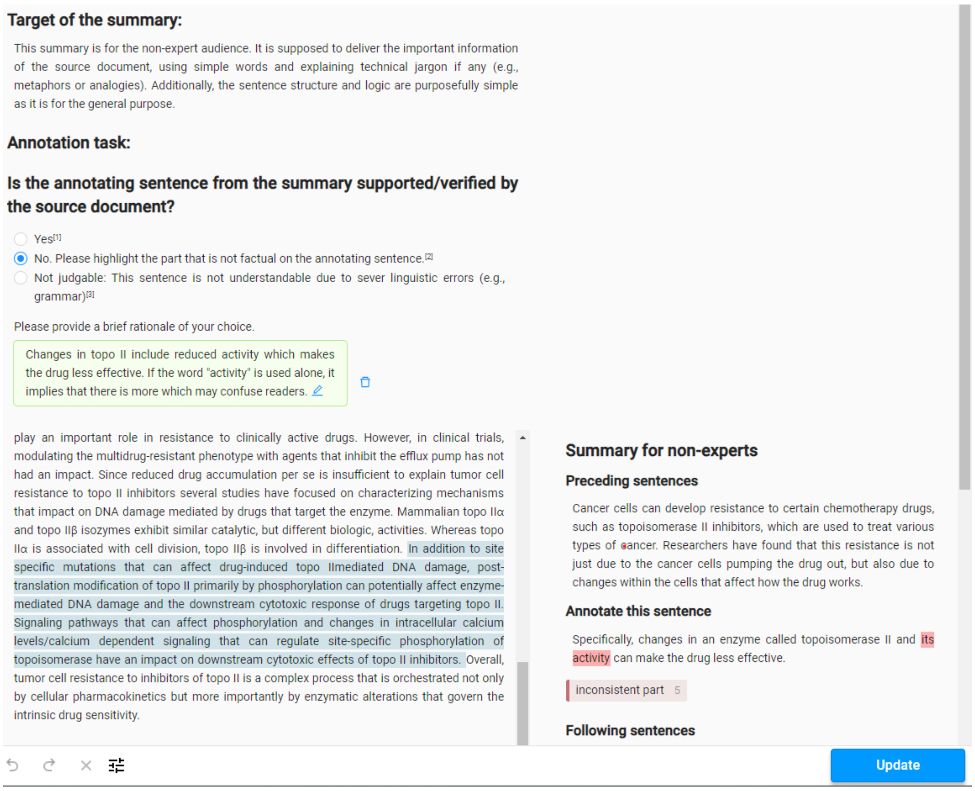
| | |
|---|---|
| GPT-4 | Yes<br><br>Supporting sentences from the Source:<br>- “Bochdalek hernias are one of the most common types of diaphragmatic hernia, with most cases diagnosed during the neonatal period."<br>- “In contrast, diagnosis of a Bochdalek hernia in an adult is rare and is typically observed on the left side of the diaphragm." |
| | Yes<br><br>Supporting sentences from the Source:<br><br>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."<br>2. “The patient underwent emergent laparotomy and required a partial resection of the necrotic ileum and a hernia repair with direct closure."<br>3. “Considering the severity of this condition, accurate diagnosis and proper treatment are needed."<br>4. “A tailored operative approach is required on an individual case basis." |
| Claude-3 | Yes, the summary is supported by the source. The following sentences from the Background section support the summary:<br><br>“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."<br><br>“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, minimising the opportunity for herniation into the right thoracic cavity [1][2][3]. |
| | 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:<br><br>“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]."<br><br>“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:<br>...<br>3. What additional socio-cultural benefits are provided by the farm to the Sumida family and surrounding community?" |
| Gemini-1.5 | Yes<br><br>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." |
| | Yes<br><br>Supporting sentences:<br><br>* **“Adult Bochdalek hernia is rare..."**<br>* **“Considering the severity of this condition, accurate diagnosis and proper treatment are needed."** |
| Llama-3 | Answer: Yes<br><br>Supporting sentences:<br>“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]."<br><br>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. |
| | Answer: Yes<br><br>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."<br><br>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. |

Table 11: Different types of responses for extracting supporting evidence.

| | |
|---|---|
| Source Article #1 | Title:<br>Mechanisms regulating resistance to inhibitors of topoisomerase II<br><br>Authors:<br>Ram Ganapathi, [...]<br><br>Abstract:<br>Inhibitors of topoisomerase II (topo II) are clinically effective in the management of hematological malignancies and solid tumors. The efficacy of anti-tumor drugs targeting topo II is often limited by resistance and studies with in vitro cell culture models have provided several insights on potential mechanisms. Multidrug transporters that are involved in the efflux and consequently reduced cytotoxicity of diverse anti-tumor agents suggest that they play an important role in resistance to clinically active drugs. [...]<br><br>Introduction:<br>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:<br>miR-135 family members mediate podocyte injury through the activation of Wnt/$\beta$-catenin signaling<br><br>Authors:<br>Xianggui Yang, [...]<br><br>Abstract:<br>[...] 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/$\beta$-catenin signaling and induced the nuclear translocation of $\beta$-catenin. Using luciferase reporter assays, reverse transcription-quantitative polymerase chain reaction (RT-qPCR) and western blot analysis, glycogen synthase kinase 3$\beta$ (GSK3$\beta$) 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$\beta$, 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.<br><br>Introduction:<br>[...] 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/$\beta$-catenin signaling through the suppression of glycogen synthase kinase 3$\beta$ (GSK3$\beta$) 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 12: Examples of supporting sentences annotated by the two annotators. Different colours represent different annotations. Multiple sentences can solely support the summary sentence and the annotators annotated different supporting sentences.

| Title | Lnk | MeSH Terms |
|---|---|---|
| Monoclonal antibodies specific for immunorecessive epitopes of glucuronoxylomannan, the major capsular polysaccharide of Cryptococcus neoformans, reduce serotype bias in an immunoassay for cryptococcal antigen | https://pubmed.ncbi.nlm.nih.gov/21697342/ | Antibodies, Monoclonal*; Antigens, Fungal / immunology*; Cryptococcosis / diagnosis*; Cryptococcosis / microbiology; Cryptococcus neoformans / classification*; Cryptococcus neoformans / isolation & purification; Enzyme-Linked Immunosorbent Assay / methods; Epitopes / immunology; Humans; Mycological Typing Techniques / methods*; Polysaccharides / immunology*; Sensitivity and Specificity; Serotyping / methods |
| Phacomatosis pigmentovascularis of cesioflammea type | https://pubmed.ncbi.nlm.nih.gov/28300894/ | Female; Humans; Melanosis / pathology*; Middle Aged; Neurocutaneous Syndromes / pathology*; Nevus, Pigmented / pathology*; Port-Wine Stain / pathology; Rare Diseases / pathology; Skin / pathology; Skin Neoplasms / pathology* |
| Right-sided Bochdalek hernia in an elderly adult: a case report with a review of surgical management | https://pubmed.ncbi.nlm.nih.gov/29030793/ | N/A |
| Keratoameloblastoma of the mandible | https://pubmed.ncbi.nlm.nih.gov/21731268/ | N/A |
| The impact of smoking and alcohol consumption on rosacea: a multivariable Mendelian randomization study | https://pubmed.ncbi.nlm.nih.gov/38439759/ | Alcohol; Drinking / epidemiology; Genome-Wide Association Study*; Humans; Mendelian Randomization Analysis; Rosacea* / epidemiology; Smoking / adverse effects; Smoking / epidemiology |
| The total solar irradiance during the recent solar minimum period measured by SOHO/VIRGO | https://pubmed.ncbi.nlm.nih.gov/37513380/ | N/A |
| Collaborative research to support urban agriculture in the face of change: The case of the Sumida watercress farm on O’ahu | https://pubmed.ncbi.nlm.nih.gov/32702038/ | Agriculture*; Bacteria / genetics; Bacteria / isolation & purification; Brassicaceae / growth & development*; Crop Production; Ecosystem; Farms; Hawaii; Nitrogen Cycle; RNA, Ribosomal, 16S / chemistry; RNA, Ribosomal, 16S / genetics; RNA, Ribosomal, 16S / metabolism; Soil Microbiology; Urbanization; Water Quality; |
| Virtual simulated international placements as an innovation for internationalisation in undergraduate programs: a mixed methods study | https://pubmed.ncbi.nlm.nih.gov/37072745/ | Australia; Cross-Sectional Studies; Curriculum*; Humans; Learning; Students* |
| Prediction of pandemic risk for animal-origin coronavirus using a deep learning method | https://pubmed.ncbi.nlm.nih.gov/34689829/ | Animals; Coronavirus Infections* / epidemiology; Coronavirus Infections* / veterinary; Coronavirus* / isolation & purification; Deep Learning; Humans; Models, Statistical; Pandemics*; Risk Assessment / methods |
| A multidisciplinary review of triphalangeal thumb | https://pubmed.ncbi.nlm.nih.gov/30318985/ | Abnormalities, Multiple / epidemiology; Fingers / embryology; Gene Duplication / genetics; Hand Deformities, Congenital / epidemiology; Hand Deformities, Congenital / genetics*; Hedgehog Proteins / physiology; Humans; Nerve Tissue Proteins / physiology; Phenotype; Point Mutation. Thumb / abnormalities*; Zinc Finger Protein Gli3 / physiology |
| The Influence of Oxidation and Nitrogenation on the Physicochemical Properties and Sorption Capacity of Activated Biocarbons Prepared from the Elderberry Inflorescence | https://pubmed.ncbi.nlm.nih.gov/37513380/ | N/A |
| Reducing the global burden of musculoskeletal conditions | https://pubmed.ncbi.nlm.nih.gov/29875522/ | Cost of Illness*; Global Health*; Humans; Musculoskeletal Diseases* / diagnosis; Musculoskeletal Diseases* / therapy; Sickness Impact Profile |
| Detection and Mitigation of IoT-Based Attacks Using SNMP and Moving Target Defense Techniques | https://pubmed.ncbi.nlm.nih.gov/36772751/ | N/A |
| A Custom Made Electrode Construct and Reliable Implantation Method That Allows for Long-Term Bilateral Deep Brain Stimulation in Mice | https://pubmed.ncbi.nlm.nih.gov/32385967/ | Animals; Deep Brain Stimulation* ; Electrodes, Implanted; Mice; Parkinson Disease* / therapy; Reproducibility of Results; Subthalamic Nucleus* |
| Diagnostic accuracy of automated occlusion detection in CT angiography using e-CTA | https://pubmed.ncbi.nlm.nih.gov/33527886/ | Cerebral Angiography; Computed Tomography Angiography*; Humans; Predictive Value of Tests; Sensitivity and Specificity; Stroke* / diagnostic imaging |
| Efficacy of life skills training on general health in students | https://pubmed.ncbi.nlm.nih.gov/23922605/ | N/A |
| Core-Shell Silver/Polymeric Nanoparticles-Based Combinatorial Therapy against Breast Cancer In-vitro | https://pubmed.ncbi.nlm.nih.gov/27491622/ | Antineoplastic Agents / pharmacology*; Breast Neoplasms / drug therapy*; Cell Line; Cell Proliferation / drug effects; Cell Survival / drug effects; Doxorubicin / pharmacology*; Drug Therapy, Combination; Female; Humans; In Vitro Techniques; MCF-7 Cells; Nanoshells / chemistry; Polyethylene Glycols / chemistry; Polymers / chemistry*; Polyvinyl Alcohol / chemistry; Povidone / chemistry; Silver / chemistry* |



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 IIα and topo IIβ isozymes exhibit similar catalytic, but different biologic, activities. Whereas topo IIα is associated with cell division, topo IIβ is involved in differentiation. In addition to site specific mutations that can affect drug-induced topo IImediated 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.

inconsistent part 5

Following sentences

Figure 8: An example of the hallucinated annotation.

| Title | Link | MeSH Terms |
|---|---|---|
| Evaluation of Natural and Factitious Food Sources for Pronematus ubiquitus on Tomato Plants | https://pubmed.ncbi.nlm.nih.gov/34940199/ | N/A |
| Progression on Citrullination of Proteins in Gastrointestinal Cancers | https://pubmed.ncbi.nlm.nih.gov/30740359/ | N/A |
| Liver Abnormalities in Turner Syndrome: The Importance of Estrogen Replacement | https://pubmed.ncbi.nlm.nih.gov/36111277/ | N/A |
| Analysis of rare variants in the C3 gene in patients with age-related macular degeneration | https://pubmed.ncbi.nlm.nih.gov/24736606/ | Aged; Aged, 80 and over; Alleles; Amino Acid Substitution; Case-Control Studies; Complement C3 / genetics*; Female; Genetic Predisposition to Disease; Genetic Variation*; Genome-Wide Association Study; Genotype; Humans; Macular Degeneration / diagnosis; Macular Degeneration / genetics*; Male; Middle Aged; Polymorphism, Single Nucleotide; Sequence Analysis, DNA; Severity of Illness Index |
| Intracholecystic papillary neoplasm acquiring malignant characteristics and leading to multiple liver metastases: A case report | https://pubmed.ncbi.nlm.nih.gov/38162850/ | N/A |
| Flexible Krylov Methods for Edge Enhancement in Imaging | https://pubmed.ncbi.nlm.nih.gov/34677302/ | N/A |
| Osteosarcopenia and Long-COVID: a dangerous combination | https://pubmed.ncbi.nlm.nih.gov/36317068/ | N/A |
| Gastric ischaemia as an unusual presentation of median arcuate ligament compression syndrome | https://pubmed.ncbi.nlm.nih.gov/30363266/ | N/A |

Table 13: The corresponding MeSH terms for the selected 25 PubMed articles. “N/A” denotes that the MeSH terms are unavailable from PubMed.