

Understanding LLMs' summarization capabilities: an analysis of biomedical abstract and lay summary generation


Batuhan Nursal Cassie S. Mitchell

Georgia Institute of Technology

bnursal3@gatech.edu,
cassie.mitchell@bme.gatech.edu

Abstract

Scientific abstracts and lay summaries serve distinct but critical roles in research communication. Abstracts use technical language for academic audiences, while lay summaries aim to make findings accessible to non-specialists. With the rise of large language models (LLMs), there is increasing interest in automating the generation of both types of summaries—especially in the biomedical domain, where clarity and factual accuracy are essential. This study evaluates the performance of lightweight LLMs (under 8B parameters) in generating biomedical abstracts and lay summaries in a zero-shot setting. We assess outputs across three key dimensions: relevance, readability, and factuality. Additionally, we introduce a novel analysis of the sectional origin and desirability of information—where desirability reflects the utility of content from the reader's perspective. We further compare human and LLM preferences using an objective ranking task. Our results show that LLM-generated summaries often contain comparable levels of desirable information to gold-standard human references. In several cases, LLM outputs are preferred by human evaluators and occasionally mistaken for human-authored text. These findings demonstrate the potential of lightweight LLMs for scalable, high-quality summarization and suggest their practical use in domains requiring both technical and accessible communication.

The codebase for this study is publicly available on  GitHub: <https://github.com/batuinmetz/Understanding-LLMs-summarization-capabilities>.

1 Introduction

The exponential growth of biomedical literature (González-Márquez et al., 2024) has intensified the need for effective summarization techniques to distill complex research into accessible formats. Two primary forms of summarization address distinct

Same Finding, Different Language

Technical vs. Lay Summary (Single Example)

Technical Summary (Abstractive)
Scientific & Domain-Specific

“Long-term exposure to high glucose levels leads to endothelial dysfunction, contributing to the development of diabetic complications.”

Source: Abstract

Lay Summary
Clear & Accessible

“Having high blood sugar for a long time can damage blood vessels, which helps explain why diabetes can cause other health problems.”

Source: Lay Summary

Measurement / Condition Biological Mechanism Health Consequence

Takeaway: Lay summaries replace technical terms with everyday language while preserving the core scientific meaning.

Figure 1: Language level difference between a sentence taken from an abstract and a sentence taken from a lay summary.

audiences: **scientific abstracts**, which offer concise overviews of a study's objectives, methods, and findings for researchers, and **lay summaries**, which translate technical content into plain language for broader public understanding (see Figure 1). While abstracts are central to information retrieval in academic databases, lay summaries support science communication in contexts such as patient education, public health, and policy-making.

Recent advances in natural language processing (NLP), especially large language models (LLMs), have enabled new approaches to biomedical summarization. Models like BioBART (Yuan et al., 2022) and LED (Beltagy et al., 2020; Luo et al., 2022) have demonstrated strong performance in generating both technical and lay summaries. In parallel, biomedical information extraction tasks—such as named entity recognition and

entity linking (Kartchner et al., 2023; Al-Hussaini et al., 2022; Kartchner et al., 2025; Bathala et al., 2025)—have contributed to structuring and grounding content. Still, summarization remains critical for distilling essential information for interdisciplinary use and public communication.

More recently, studies have emphasized lay summarization. For example, BioLaySumm (Goldsack et al., 2024) shows that LLMs can produce lay summaries approaching human quality. Domain enrichment strategies, such as integrating biomedical entity linking or external knowledge graphs like SemNet 2.0 (Kirkpatrick et al., 2022), have improved summary factuality and relevance (Goldsack et al., 2023). However, comparisons across models remain difficult due to varying fine-tuning approaches, datasets, and proprietary architectures. Even recent work on zero-shot evaluation (Goldsack et al., 2025) is limited in scale and scope. Notably, no prior study has systematically analyzed the sentence-level content or information desirability in LLM-generated biomedical summaries.

With growing interest in deployable AI systems, it is essential to assess lightweight, resource-efficient LLMs—especially given the prohibitive cost of long-document inference with larger models in biomedical settings. To this end, our paper makes the following contributions:

- We perform an unprecedented unified comparative analysis of scientific and lay summaries using the PLOS and eLife datasets (Goldsack et al., 2022).
- We present an original baseline evaluation of sub-8B parameter LLMs in a zero-shot setting using standard summarization metrics.
- We perform a novel sentence-level lexical analysis of LLM outputs across two input lengths to quantify divergence from human references.
- We perform a complete hybrid evaluation that combines LLM-based and human-based judgments to assess summary quality.

2 Related work

There is growing interest in using NLP—particularly large language models (LLMs) (Raiaan et al., 2024)—to generate summaries of scientific documents, with an increasing focus on lay summarization. Goldsack

et al. (Goldsack et al., 2025) explored zero-shot lay summarization using a two-stage prompting strategy, achieving strong results with 70B+ parameter models. Their study demonstrated the potential of prompt engineering to improve clarity and even proposed using LLM-generated preference panels as proxies for human evaluation. However, their work was limited to the eLife dataset, relied heavily on automated metrics, and did not systematically compare lightweight LLMs. Other studies have investigated methods to improve the relevance and factual grounding of lay summaries. Ming et al. (2025) showed that prompting with MeSH terms enhances the informativeness of biomedical lay summaries, although their evaluation focused on expert alignment rather than lay reader comprehension or content desirability. Similarly, Pakull et al. (2024) fine-tuned LLMs and applied dynamic expert selection to optimize readability and factuality, achieving strong results in the BioLaySumm shared task. These shared task reports (Goldsack et al., 2024) have advanced the field by emphasizing human-centric criteria—readability, accuracy, and clarity—while documenting the growing role of LLM-based systems. Recent work has also focused on aligning scientific summarization with diverse communicative goals and reader personas. Fonseca and Cohen (Fonseca and Cohen, 2024) investigated whether LLM summarizers can adapt to different scientific communication objectives—such as writing for experts versus lay audiences—and found that current models often fail to consistently align content with the intended communicative goal. Similarly, Mullick et al. (Mullick et al., 2024) proposed persona-based summarization for domain-specific documents, showing that conditioning on reader profiles (e.g., “expert,” “novice,” or “policy-maker”) can significantly alter the level of technicality and narrative framing in generated summaries. Complementary to these efforts, Ragazzi et al. (2024) introduced a differentiable perturbed Top- k token selection mechanism that improves factual grounding and coherence in scientific document summarization, highlighting the value of controllable token-level selection for abstractive generation. Despite these advances, most prior work has emphasized fine-tuned models and leaderboard performance across automated metrics. Few studies have conducted zero-shot evaluations,

and even fewer have analyzed the sentence-level content or assessed the performance of smaller LLMs.

3 Methodology

3.1 Datasets

We utilize two biomedical datasets previously employed in lay summarization tasks to facilitate comparison across the literature.: **eLife** and **PLOS**. In our study, we specifically explore two summarization levels: the abstract and the lay summary, reflecting different levels of technical complexity. We adopt the original training, validation, and test splits provided by the authors, and do not incorporate the mixed split configuration used in Bio-LaySumm 2024. The summary length statistics are given in Appendix (Table 9). Below, we provide a brief overview of the two datasets:

- **eLife**: Contains 4,828 biomedical research articles published in the eLife journal. Each article includes a scientific abstract and a corresponding lay summary written by professional science communicators. The dataset offers a clear separation between technical and non-technical summaries, making it well-suited for evaluating lay-level generation.
- **PLOS**: Comprises 27,525 articles from PLOS journals, each paired with an abstract and an “Author Summary” section. The author-written lay summaries aim to make scientific content more accessible to a broader audience, providing another perspective for evaluating lay summarization models.

3.2 Metrics

We adopt a comprehensive evaluation approach to assess the performance of the selected small-scale LLMs on the eLife and PLOS datasets. Following common practices, we assess model outputs based on three main criteria: relevance, readability, and factuality. Each criterion is measured using established automatic evaluation metrics:

- **Relevance**
 - **ROUGE** Lin (2004): A set of metrics that measure n-gram overlap between the generated and reference texts. ROUGE-1, ROUGE-2, and ROUGE-L are commonly used variants that assess content similarity and coverage.

- **BERTScore** Zhang* et al. (2020): Calculates semantic similarity between candidate and reference texts using contextual embeddings from a pre-trained BERT model. Unlike ROUGE, it captures deeper semantic alignment beyond surface-level word overlap.

- **Readability**

- **Flesch Reading Ease (FRE)**: A classic readability measure that evaluates how easy a text is to read based on sentence length and syllable count. Higher scores indicate simpler, more accessible language, making it particularly useful for assessing lay summaries intended for non-expert audiences.
- **Coleman–Liau Index (CLI)**: A readability metric that estimates the U.S. grade level needed to understand a piece of text, based on the average number of letters per word and the number of sentences per 100 words.
- **Dale-Chall Readability Score (DCRS)**: Another readability formula that assesses text difficulty based on a list of 3,000 familiar words. The score incorporates both sentence length and the proportion of difficult words, making it suitable for evaluating clarity in generated content.

- **Factuality**

- **BARTScore** Yuan et al. (2021): Uses a pre-trained BART model to score the likelihood of the reference article and the generated summary, capturing fluency and factual alignment.
- **AlignScore** Zha et al. (2023): Measures factual consistency through natural language inference (NLI) by aligning each sentence of the generated summary with supporting evidence from the source text. It leverages a RoBERTa-based model fine-tuned on factuality datasets, offering state-of-the-art performance and strong correlation with human factuality judgments.

3.3 Models

The families of LLMs utilized include:

- **Llama 3 family** (Grattafiori et al., 2024): Llama 3.1-8B-Instruct and Llama 3.2-3B-Instruct.

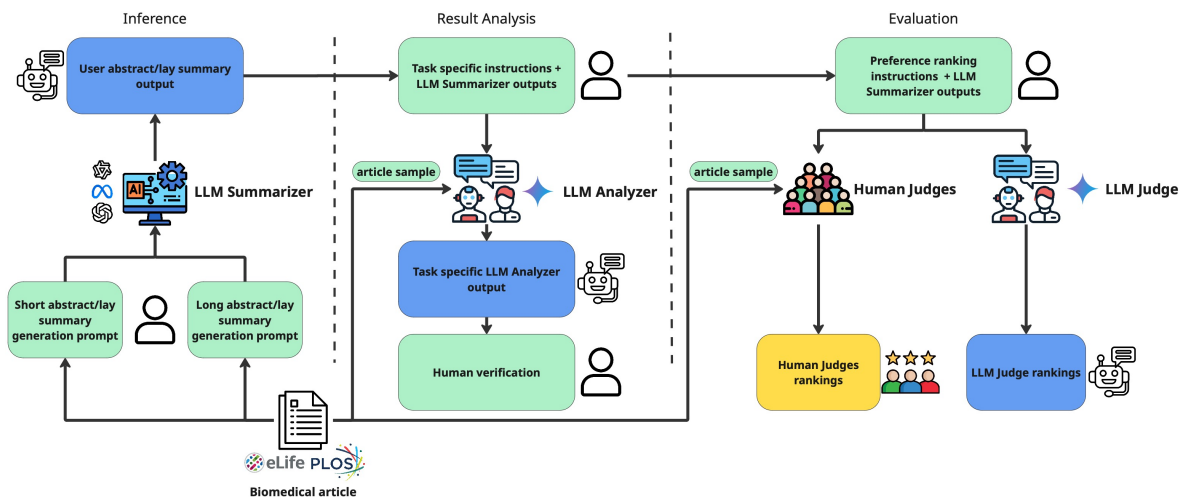


Figure 2: Framework for LLM-based inference and evaluation: The user provides a prompt containing the full article (without section names) for abstract or lay summary generation at a specified length. The generated summary is analyzed by the LLM Analyzer, which either classifies each sentence or generates questions from the source article/target summary and answers from the summary. Outputs from different models are then ranked by the LLM Judge or human evaluators according to explicit instructions.

- **Qwen 2 family** (Qwen et al., 2025): Qwen 2.5-3B-Instruct and Qwen 3.2.5-7B-Instruct.
- **GPT4 family** (OpenAI et al., 2024): We include GPT4o to compare how well lightweighted LLMs compare with significantly bigger models. GPT4o will only be used for inference on the eLife test set due to cost constraints.

4 Results

We adopt the inference setup illustrated in Figure 2, evaluating various small-scale LLMs under zero-shot prompting on the eLife and PLOS test sets. Two summarization tasks are examined: generating long abstracts and short abstracts. Corresponding results are shown in Table 1 and Table 2, additionally summary length statistics are given in Appendix (Table 8). During lay summary generation, the full article (excluding section headers) is provided as input along with an instruction prompt (see Appendix A.2), whereas the abstract is excluded during abstract generation. This choice forces the model to make a choice when picking information from a single block corpora rather than picking a certain number of information from each section and it enables a better investigation of how summary length affects relevance, readability, and factuality.

4.1 Abstract generation

Relevance and Semantic Similarity- Across both datasets, Llama and Qwen models typically yield higher ROUGE scores, indicating stronger lexical overlap with the gold abstracts. In contrast, GPT-4o consistently attains higher BERTScore, suggesting better semantic alignment despite less overlap in exact phrasing. This pattern highlights a trade-off between lexical similarity and conceptual relevance.

Readability- The average FRE, CLI and DCRS for generated abstracts are comparable to those observed in the reference abstracts of the test sets. Notably, abstract length appears to significantly influence readability metrics: when constrained to generate shorter abstracts, models tend to use more abstract or condensed vocabulary, which naturally increases readability scores. This effect is more pronounced in PLOS, where reference abstracts are often longer than 150 words. In contrast, eLife reference abstracts are already concise, leading to closer length alignment and less deviation in readability.

Factuality- Factuality, measured using AlignScore, is slightly better for Llama models and GPT-4o compared to Qwen models in the eLife dataset. This aligns with expectations given their stronger semantic understanding. Surprisingly, all generated summaries reach better factuality scores than the reference summaries on both datasets. Addi-

| | Rouge Metrics \uparrow | | | BERTscore \uparrow | FRE \uparrow | CLI \downarrow | DCRS \downarrow | BARTscore \downarrow | Alignscore \uparrow |
|-----------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------|------------------------------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 41.29 \pm 0.88 | 12.91 \pm 0.27 | 21.66 \pm 0.24 | 85.57 \pm 0.35 | 25.60 \pm 0.45 | 16.72 \pm 0.06 | 10.86 \pm 0.27 | -2.48 \pm 0.015 | 78.77 \pm 0.77 |
| Llama-3.1-8B-Instruct | 42.03 \pm 0.59 | 13.49 \pm 0.05 | 22.14 \pm 0.17 | 85.99 \pm 0.06 | 25.89 \pm 0.12 | 16.61 \pm 0.04 | 10.87 \pm 0.05 | -2.51 \pm 0.005 | 79.62 \pm 0.56 |
| Qwen-2.5-3B-Instruct | 40.96 \pm 0.69 | 11.85 \pm 0.09 | 20.64 \pm 0.28 | 85.76 \pm 0.27 | 18.54 \pm 0.18 | 18.77 \pm 0.15 | 12.17 \pm 0.16 | -2.83 \pm 0.012 | 76.78 \pm 0.43 |
| Qwen-2.5-7B-Instruct | 42.27 \pm 0.39 | 12.72 \pm 0.20 | 21.36 \pm 0.41 | 85.97 \pm 0.04 | 19.67 \pm 0.19 | 18.69 \pm 0.12 | 12.13 \pm 0.10 | -2.83 \pm 0.009 | 76.71 \pm 0.33 |
| GPT4o | 41.54 | 12.65 | 21.03 | 86.27 | 18.23 | 18.86 | 11.73 | -2.74 | 77.59 |
| eLife abstracts | - | - | - | - | 23.72 | 17.53 | 11.87 | -3.50 | 63.42 |
| PLOS Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 45.69 \pm 0.03 | 15.09 \pm 0.05 | 22.75 \pm 0.02 | 85.61 \pm 0.01 | 26.61 \pm 0.08 | 16.64 \pm 0.05 | 11.06 \pm 0.04 | -2.39 \pm 0.004 | 80.91 \pm 0.11 |
| Llama-3.1-8B-Instruct | 45.10 \pm 0.74 | 15.17 \pm 0.48 | 23.36 \pm 0.04 | 86.11 \pm 0.14 | 25.11 \pm 0.02 | 16.89 \pm 0.17 | 11.20 \pm 0.04 | -1.82 \pm 0.009 | 86.80 \pm 0.06 |
| Qwen-2.5-3B-Instruct | 42.79 \pm 1.17 | 13.41 \pm 0.38 | 21.25 \pm 0.36 | 85.56 \pm 0.06 | 19.49 \pm 0.09 | 18.58 \pm 0.10 | 12.15 \pm 0.12 | -2.72 \pm 0.005 | 79.92 \pm 0.02 |
| Qwen-2.5-7B-Instruct | 43.82 \pm 0.70 | 14.04 \pm 0.39 | 21.93 \pm 0.52 | 85.77 \pm 0.06 | 20.33 \pm 0.18 | 18.57 \pm 0.10 | 12.16 \pm 0.04 | -2.66 \pm 0.004 | 79.65 \pm 0.07 |
| PLOS abstracts | - | - | - | - | 27.55 | 16.39 | 11.06 | -3.33 | 63.40 |

Table 1: Results for long abstracts generation on eLife/PLOS test sets (\uparrow :higher is better / \downarrow :lower is better).

| | Rouge Metrics \uparrow | | | BERTscore \uparrow | FRE \uparrow | CLI \downarrow | DCRS \downarrow | BARTscore \downarrow | Alignscore \uparrow |
|-----------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------|------------------------------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 42.94 \pm 0.08 | 13.14 \pm 0.08 | 22.23 \pm 0.06 | 85.92 \pm 0.05 | 25.17 \pm 0.55 | 16.97 \pm 0.09 | 11.38 \pm 0.06 | -2.47 \pm 0.017 | 78.70 \pm 0.37 |
| Llama-3.1-8B-Instruct | 42.91 \pm 0.64 | 13.74 \pm 0.13 | 22.57 \pm 0.33 | 86.12 \pm 0.09 | 26.03 \pm 0.24 | 16.62 \pm 0.04 | 11.20 \pm 0.04 | -2.53 \pm 0.027 | 79.69 \pm 0.40 |
| Qwen-2.5-3B-Instruct | 40.20 \pm 0.20 | 11.35 \pm 0.09 | 20.68 \pm 0.13 | 85.99 \pm 0.04 | 19.54 \pm 0.65 | 19.08 \pm 0.10 | 12.66 \pm 0.02 | -2.86 \pm 0.008 | 78.83 \pm 0.10 |
| Qwen-2.5-7B-Instruct | 41.35 \pm 0.27 | 12.22 \pm 0.14 | 21.48 \pm 0.13 | 86.24 \pm 0.05 | 20.56 \pm 0.32 | 18.90 \pm 0.06 | 12.63 \pm 0.02 | -2.75 \pm 0.016 | 78.45 \pm 0.59 |
| GPT4o | 42.26 | 12.03 | 21.89 | 86.58 | 17.31 | 19.53 | 12.80 | -2.64 | 79.72 |
| eLife abstracts | - | - | - | - | 23.72 | 17.53 | 11.87 | -3.50 | 63.42 |
| PLOS Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 43.97 \pm 0.09 | 14.48 \pm 0.10 | 22.36 \pm 0.03 | 85.55 \pm 0.02 | 26.22 \pm 0.11 | 16.80 \pm 0.02 | 11.46 \pm 0.01 | -2.39 \pm 0.005 | 81.84 \pm 0.07 |
| Llama-3.1-8B-Instruct | 44.29 \pm 0.40 | 14.62 \pm 0.41 | 23.04 \pm 0.03 | 86.05 \pm 0.13 | 25.74 \pm 0.11 | 16.82 \pm 0.12 | 11.32 \pm 0.26 | -1.88 \pm 0.012 | 86.85 \pm 0.12 |
| Qwen-2.5-3B-Instruct | 37.00 \pm 0.09 | 11.69 \pm 0.08 | 19.57 \pm 0.08 | 85.40 \pm 0.02 | 20.43 \pm 0.05 | 18.82 \pm 0.01 | 12.62 \pm 0.01 | -2.71 \pm 0.002 | 82.96 \pm 0.01 |
| Qwen-2.5-7B-Instruct | 38.34 \pm 0.08 | 12.61 \pm 0.08 | 20.46 \pm 0.06 | 85.65 \pm 0.02 | 21.27 \pm 0.09 | 18.68 \pm 0.03 | 12.60 \pm 0.01 | -2.60 \pm 0.005 | 82.64 \pm 0.21 |
| PLOS abstracts | - | - | - | - | 27.55 | 16.39 | 11.06 | -3.33 | 63.40 |

Table 2: Results for short abstracts generation on eLife/PLOS test sets (\uparrow :higher is better / \downarrow :lower is better).

tionally, we observe a comparable range of values across models between long and short abstracts. This suggests that models can maintain a high factuality while shortening their text generation.

Effect of Length Constraints- Imposing stricter length limits affects multiple metrics. Shorter abstracts often achieve higher recall-oriented ROUGE scores, possibly because models condense content into essential points. However, this brevity may occasionally compromise completeness or fluency as some information are discarded or simplified depending on the model’s summarization strategy.

4.2 Lay summary generation

We use a similar zero-shot setup for lay summarization, generating long lay summaries and short versions. Prompt details are provided in the Appendix A.2. Results are presented in Table 3 and Table 4. Summary length statistics are given in Table 7.

Relevance and Semantic Similarity- Lay summary generation proves more challenging, as reflected by lower ROUGE and BERTScore compared to abstract generation. This is expected given the need to both identify salient information

and simplify technical content. On eLife, GPT-4o achieves the highest BERTScore, with Llama and Qwen models closely trailing. Notably, Llama 3.1-8B surpasses GPT-4o in ROUGE for short lay summaries. In PLOS, absolute ROUGE scores are substantially lower than those for eLife, but performance improves with shorter summaries, suggesting a better alignment with reference length.

Readability- All models generate lay summaries with improved readability relative to abstracts, confirming their ability to adapt to the simpler target style. CLI and DCRS scores generally fall within expected ranges whereas FRE scores are higher than the references. eLife remains more challenging, with models struggling to reduce CLI scores and increase FRE scores significantly. By contrast, the PLOS dataset shows clearer improvements, suggesting more effective simplification of technical content.

Factuality- Factual consistency remains relatively stable across models and summary lengths. Similarly to abstracts, generated lay summaries tend to be more factual than the reference lay summaries, this tendency is more pronounced on the eLife dataset. Interestingly, BARTScore and Align-

| | Rouge Metrics \uparrow | | | BERTscore \uparrow | FRE \uparrow | CLI \downarrow | DCRS \downarrow | BARTscore \downarrow | Alignscore \uparrow |
|-----------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-------------------------------------|------------------------------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 42.01 \pm 0.26 | 09.88 \pm 0.08 | 18.98 \pm 0.04 | 83.69 \pm 0.07 | 45.66 \pm 0.27 | 13.38 \pm 0.10 | 08.27 \pm 0.14 | -2.52 \pm 0.007 | 58.11 \pm 0.01 |
| Llama-3.1-8B-Instruct | 39.54 \pm 0.29 | 09.05 \pm 0.05 | 18.29 \pm 0.12 | 83.22 \pm 0.03 | 40.10 \pm 0.28 | 14.55 \pm 0.03 | 08.19 \pm 0.02 | -2.70 \pm 0.005 | 66.19 \pm 0.01 |
| Qwen-2.5-3B-Instruct | 39.77 \pm 0.71 | 08.57 \pm 0.12 | 17.07 \pm 0.18 | 83.72 \pm 0.06 | 55.52 \pm 0.62 | 12.74 \pm 0.12 | 09.03 \pm 0.10 | -3.16 \pm 0.012 | 59.26 \pm 0.02 |
| Qwen-2.5-7B-Instruct | 42.13 \pm 0.14 | 09.65 \pm 0.10 | 17.88 \pm 0.09 | 83.68 \pm 0.16 | 46.79 \pm 0.06 | 14.13 \pm 0.17 | 09.50 \pm 0.05 | -3.18 \pm 0.007 | 62.82 \pm 0.01 |
| GPT4o | 46.08 | 11.76 | 19.18 | 84.86 | 40.57 | 14.80 | 09.48 | -2.58 | 68.39 |
| eLife Lay summaries | - | - | - | - | 51.83 | 12.50 | 08.93 | -3.28 | 43.42 |
| PLOS Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 32.47 \pm 0.36 | 08.43 \pm 0.06 | 17.41 \pm 0.16 | 83.84 \pm 0.08 | 46.86 \pm 0.07 | 13.20 \pm 0.06 | 08.17 \pm 0.04 | -2.46 \pm 0.003 | 61.65 \pm 0.05 |
| Llama-3.1-8B-Instruct | 32.15 \pm 0.27 | 09.08 \pm 0.05 | 17.52 \pm 0.09 | 83.77 \pm 0.23 | 41.72 \pm 0.14 | 14.32 \pm 0.04 | 08.29 \pm 0.02 | -2.66 \pm 0.007 | 68.11 \pm 0.35 |
| Qwen-2.5-3B-Instruct | 32.35 \pm 1.04 | 06.57 \pm 0.32 | 15.88 \pm 0.40 | 83.68 \pm 0.26 | 56.03 \pm 0.49 | 12.49 \pm 0.33 | 08.84 \pm 0.22 | -3.07 \pm 0.004 | 62.21 \pm 0.23 |
| Qwen-2.5-7B-Instruct | 36.00 \pm 1.30 | 08.66 \pm 0.09 | 17.69 \pm 0.54 | 84.22 \pm 0.43 | 45.09 \pm 0.08 | 14.35 \pm 0.04 | 09.75 \pm 0.16 | -2.97 \pm 0.004 | 67.44 \pm 0.09 |
| PLOS lay summaries | - | - | - | - | 30.03 | 15.90 | 10.91 | -2.94 | 65.48 |

Table 3: Results for long lay summaries generation on eLife/PLOS test sets (\uparrow :higher is better / \downarrow :lower is better).

| | Rouge Metrics \uparrow | | | BERTscore \uparrow | FRE \uparrow | CLI \downarrow | DCRS \downarrow | BARTscore \downarrow | Alignscore \uparrow |
|-----------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-------------------------------------|------------------------------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 40.65 \pm 0.55 | 09.61 \pm 0.20 | 18.79 \pm 0.20 | 84.02 \pm 0.04 | 47.77 \pm 0.58 | 13.12 \pm 0.12 | 09.02 \pm 0.17 | -2.55 \pm 0.010 | 59.38 \pm 0.01 |
| Llama-3.1-8B-Instruct | 41.30 \pm 0.14 | 09.80 \pm 0.05 | 19.03 \pm 0.05 | 83.87 \pm 0.04 | 40.10 \pm 0.28 | 13.27 \pm 0.14 | 08.59 \pm 0.04 | -2.59 \pm 0.017 | 62.26 \pm 0.02 |
| Qwen-2.5-3B-Instruct | 33.84 \pm 0.28 | 07.31 \pm 0.08 | 15.76 \pm 0.05 | 83.78 \pm 0.05 | 48.20 \pm 0.57 | 14.31 \pm 0.26 | 10.12 \pm 0.16 | -3.01 \pm 0.007 | 65.14 \pm 0.01 |
| Qwen-2.5-7B-Instruct | 35.28 \pm 0.33 | 08.11 \pm 0.12 | 16.43 \pm 0.10 | 83.97 \pm 0.03 | 43.59 \pm 0.25 | 14.85 \pm 0.06 | 10.47 \pm 0.04 | -2.90 \pm 0.008 | 68.95 \pm 0.02 |
| GPT4o | 36.89 | 08.56 | 17.33 | 84.78 | 41.43 | 15.10 | 10.51 | -2.60 | 68.81 |
| eLife Lay summaries | - | - | - | - | 51.83 | 12.50 | 08.93 | -3.28 | 43.42 |
| PLOS Dataset | | | | | | | | | |
| Llama-3.2-3B-Instruct | 37.47 \pm 0.23 | 08.96 \pm 0.08 | 19.13 \pm 0.09 | 84.72 \pm 0.04 | 48.68 \pm 0.04 | 12.85 \pm 0.05 | 09.06 \pm 0.08 | -2.51 \pm 0.012 | 64.33 \pm 0.52 |
| Llama-3.1-8B-Instruct | 35.92 \pm 0.04 | 09.03 \pm 0.04 | 18.73 \pm 0.03 | 84.26 \pm 0.01 | 46.88 \pm 0.19 | 13.38 \pm 0.01 | 08.64 \pm 0.02 | -2.58 \pm 0.003 | 66.64 \pm 0.02 |
| Qwen-2.5-3B-Instruct | 35.52 \pm 0.06 | 07.79 \pm 0.04 | 17.43 \pm 0.02 | 84.85 \pm 0.02 | 49.23 \pm 0.09 | 14.04 \pm 0.02 | 10.02 \pm 0.01 | -2.85 \pm 0.003 | 70.06 \pm 0.23 |
| Qwen-2.5-7B-Instruct | 37.66 \pm 0.05 | 08.73 \pm 0.02 | 18.51 \pm 0.04 | 85.07 \pm 0.01 | 44.94 \pm 0.36 | 15.03 \pm 0.14 | 10.43 \pm 0.03 | -2.80 \pm 0.004 | 71.00 \pm 0.04 |
| PLOS lay summaries | - | - | - | - | 30.03 | 15.90 | 10.91 | -2.94 | 65.48 |

Table 4: Results for short lay summaries generation on eLife/PLOS test sets (\uparrow :higher is better / \downarrow :lower is better).

Score for lay summaries generated on the PLOS dataset are markedly respectively lower and higher than those for eLife, and a bit higher than the corresponding abstract generation scores. This may suggest that, despite the simplification required, the models are still able to preserve factual alignment in this setting.

Effect of Length Constraints- As with abstract generation, length constraints influence performance across all metrics. Shorter lay summaries often yield better ROUGE scores for PLOS and for Llama models on eLife, possibly due to more concise and focused content that aligns more closely with the reference text. These findings underscore the importance of prompt calibration and length control in producing high-quality, user-aligned lay summaries.

4.3 Lexical and Content Origin Analysis

While earlier sections focused on comparing generated summaries to gold references using automated metrics, such evaluations often fail to capture why certain information is included and how it aligns with human preferences. To address this, we conduct two complementary analyses. First, we per-

form sentence origin classification, tracing each summary sentence back to its corresponding section in the source article (e.g., Introduction, Methods, Results, Discussion). Second, we implement a QA-based content utility assessment, evaluating whether the summary contains sufficient information to answer key questions automatically generated from the original article.

Sentence type classification. Following the result analysis framework illustrated in Figure 2, we employ Gemini 2.0 Flash as the LLM to perform sentence-level classification. For this task, we sample 20 articles each from the eLife and PLOS datasets. For each article, we provide the LLM with the full article (including section headers), followed by the generated abstract and lay summary. The LLM is prompted to assign each sentence from the summary to the most appropriate section of the article. To maintain strictness and interpretability, we use hard labels. Each sentence is assigned to only one section, without modeling potential overlaps. To ensure labeling quality, we restrict the analysis to a small, manually verifiable set. For this study, we analyze summaries generated by GPT-4o for the eLife dataset and Llama 3.1 for the PLOS

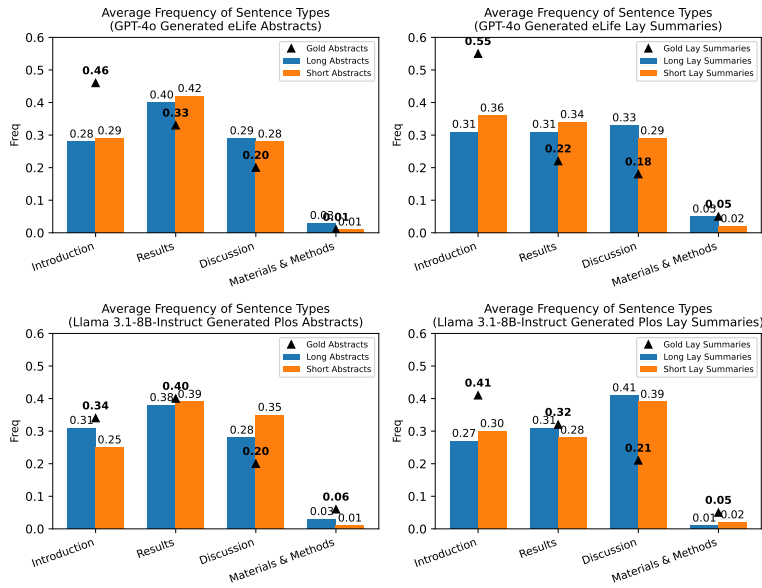


Figure 3: Sentence type distribution across sections for generated abstracts and lay summaries with GPT4o and Llama 3.1-8B on the samples from eLife and PLOS.

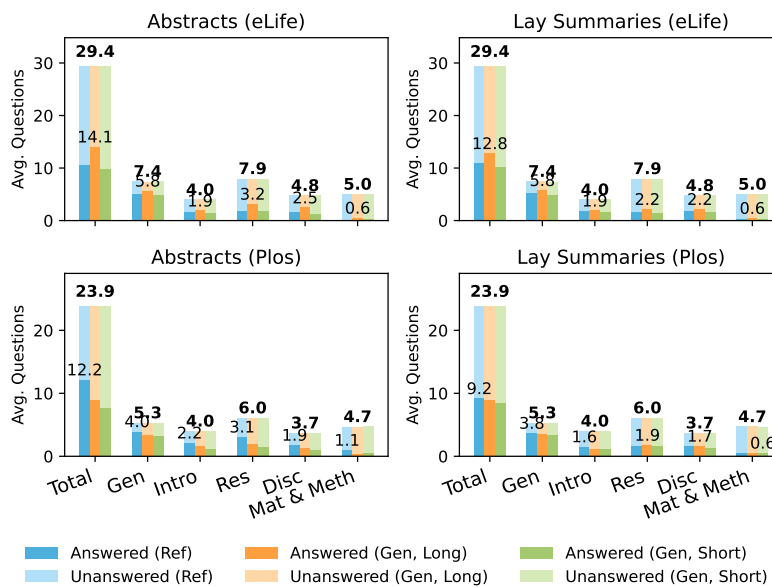


Figure 4: Answerable and unanswerable question frequencies for eLife and PLOS samples.

dataset, as these models consistently achieve higher performance according to automated metrics across both summary types. For comparison, we perform the same analysis on the gold reference summaries. The resulting sentence-type distributions are presented in Figure 3. The details of the prompt are given in Appendix A.3.

Quantifying Content Desirability Beyond classifying sentence types, we quantified “desirable information”—content providing useful insights for readers. Using Gemini 2.0 Flash, we generated

question–answer pairs from full articles (with section annotations) and prompted the LLM to answer each question using only the generated summary (see Appendix A.3). The proportion of answerable questions assessed how well summaries captured essential content (Figure 4). Additionally, we generated questions from gold reference summaries to see how many earlier summaries (GPT-4o, Llama 3.1-8B-Instruct) could answer, providing a finer measure of content desirability (Appendix Table 10).

Results Discussion Sentence-type classification reveals that both long and short summaries exhibit similar distributions of content derived from the Introduction, Results, and Discussion sections. However, longer summaries are more likely to include material from the Materials & Methods section, suggesting that length facilitates greater inclusion of methodological details. Compared to gold references, LLM-generated summaries tend to contain approximately twice as many Discussion-related sentences and half as many from the Introduction, while proportions for Results and Methods remain relatively consistent. As shown in Figure 4, GPT-4o’s long summaries on the eLife dataset consistently achieve a higher proportion of answerable questions across all content types compared to gold references, whereas short summaries perform comparably. In contrast, for the PLOS dataset, gold summaries outperform LLM-generated ones in terms of answerability. These discrepancies may stem from differences in article length and writing style between the two sources. Notably, PLOS summaries generated by LLMs frequently fail to adequately answer questions related to Results & Methods. This underscores both the overlapping nature of key content across sections and the broader finding that, when considering both datasets together, LLM-generated summaries can convey a comparable amount of desirable information to that of human-written references.

5 Human vs LLM as judges

To complement automatic metrics and content analysis, we perform a preference-based evaluation of summary quality using both LLM and human judges in a restrained sample set containing 40 articles (20 articles from eLife and 20 articles from PLOS). This dual approach allows us to assess how well the summaries meet key criteria of an ideal abstract or lay summary.

5.1 LLM judge

We follow the evaluation protocol outlined in Figure 2. For each article in our sampled test set, we compile a set of five abstracts generated by the different LLMs included in our study. These candidate summaries are presented to Gemini 2.0 Flash, which serves as judge. The judge is prompted to rank the summaries from most to least preferred according to objective criteria expected from an ideal abstract or ideal lay summary (see Appendix A.4 for the detailed prompt with criteria). We first ex-

amine the absolute top-ranked summaries chosen by the LLM for each article. Then, we compute summary-level statistics to determine which models consistently produce higher-ranked summaries. Table 5 shows the number of times a summary generated by each LLM was selected as the top choice, compared to the gold target summaries. Additionally, we asked the LLM judge to classify the summaries for each article as human-generated or LLM-generated. We show the result obtained for reference summaries and generated summaries in Table 6.

| Summary Type | Judge | LLM Top Choice | Gold Top Choice |
|--------------|--------------|----------------|-----------------|
| Abstract | LLM Judge | 0.65 | 0.35 |
| | Human Judges | 0.85 | 0.15 |
| Lay Summary | LLM Judge | 0.73 | 0.27 |
| | Human Judges | 0.83 | 0.17 |

Table 5: Top 1 ranked summaries by LLM and human judges across both summary types.

| Summary Type | Judge | Summary Source | LLM-Generated | Human-Written |
|--------------|-------------|----------------|---------------|---------------|
| Abstract | LLM Judge | Reference | 0.10 | 0.90 |
| | LLM Judge | LLM-Generated | 0.57 | 0.43 |
| | Human Judge | Reference | 0.05 | 0.95 |
| | Human Judge | LLM-Generated | 0.19 | 0.81 |
| Lay Summary | LLM Judge | Reference | 0.20 | 0.80 |
| | LLM Judge | LLM-Generated | 0.57 | 0.43 |
| | Human Judge | Reference | 0.05 | 0.95 |
| | Human Judge | LLM-Generated | 0.44 | 0.56 |

Table 6: Classification of summaries as LLM-generated or human-written by LLM and human judges across both summary types and sources.

5.2 Human judges

We used the same sampled test set for both human and LLM-based evaluations to enable a direct comparison. A cohort of 80 undergraduate volunteers ranked both the abstracts and lay summaries using the same instructions and criteria as the LLM Judge. Rankings were aggregated at the document level to compute summary-level statistics across the full sample. Table 5 presents results for the top-1 ranked summaries, while Table 6 shows results for distinguishing human- and LLM-generated summaries. More explicit ranking frequencies per model generated summaries and rating settings are given in Appendix A.6, A.1.

5.3 Judges results

Overall, human and LLM judges tended to converge on the same top-ranked summary, regardless of summary type. A similar trend was observed in the classification task: the LLM judge correctly identified many human-written reference

summaries but showed limited accuracy in distinguishing between LLM- and human-generated content for the remaining examples. This mirrors the difficulty human evaluators faced in reliably identifying LLM-generated lay summaries, suggesting a growing overlap in quality and style.

6 Conclusion

Lightweight LLMs demonstrates significant capabilities for zero-shot biomedical summary generation with automated evaluation metrics. Our studies further indicates that summaries from GPT4o and Llama 3.1-8B contain by default more information extracted from Results, Discussion and Introduction with a lesser extend Materials & Methods related content. The ability to answer specific questions requiring more granular information from the article as frequently as gold reference summaries highlights that the core objective of both types of summaries is reached. Further LLM and human based rankings indicates that LLM generated summaries can often be objectively preferred and be confused with human written summaries.

Limitations

This study has several limitations. First, we used a fixed base prompt for summary generation, which may have constrained the diversity and quality of outputs. More tailored or adaptive prompting strategies could potentially improve results and warrant future investigation. Second, the inherent variability of LLM outputs poses challenges for consistent evaluation, particularly in the absence of sampling or decoding controls. Additionally, we did not account for potential data contamination in the models. Our analysis is further limited by sample size: due to resource and time constraints, only a subset of summaries was selected for detailed lexical analysis and evaluation by both human and LLM judges. Finally, we focused on a small set of lightweight pretrained LLMs in a zero-shot setting. While this setup reflects realistic deployment scenarios, such models may be less capable of producing highly abstracted or lay-accessible summaries due to limited domain-specific knowledge. In-context learning approaches or biomedical fine-tuned models, which may improve domain-specific summarization performance, were not explored and remain avenues for future work.

Acknowledgements

Funding support provided by National Science Foundation CAREER grant 1944247, National Institute of Health grant R35GM152245, and Chan Zuckerberg Initiative grant 253558 to C.S.M.

References

- Irfan Al-Hussaini, Davi Nakajima An, Albert J. Lee, Sarah Bi, and Cassie S. Mitchell. 2022. [Ccs explorer: Relevance prediction, extractive summarization, and named entity recognition from clinical cohort studies](#). In *2022 IEEE International Conference on Big Data (Big Data)*, page 5173–5181. IEEE.
- Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner, and Cassie S. Mitchell. 2025. [BioEL: A comprehensive python package for biomedical entity linking](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1709–1721, Albuquerque, New Mexico. Association for Computational Linguistics.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#).
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. [Huatuogpt-01, towards medical complex reasoning with llms](#).
- Marcio Fonseca and Shay Cohen. 2024. [Can large language model summarizers adapt to diverse scientific communication goals?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8599–8618, Bangkok, Thailand. Association for Computational Linguistics.
- Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2025. [Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond](#).
- Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. [Overview of the BioLay-Summ 2024 shared task on the lay summarization of biomedical research articles](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. 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.
- Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. [Enhancing biomedical lay summarisation with external knowledge graphs](#).

Rita González-Márquez, Luca Schmidt, Benjamin M. Schmidt, Philipp Berens, and Dmitry Kobak. 2024. [The landscape of biomedical research](#). *Patterns*, 5(6):100968.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alonsoius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Van-

denhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle

- Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).
- David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. [A comprehensive evaluation of biomedical entity linking models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478, Singapore. Association for Computational Linguistics.
- David Kartchner, Haydn Turner, Christophe Ye, Irfan Al-Hussaini, Batuhan Nursal, Albert J. B. Lee, Jennifer Deng, Courtney Curtis, Hannah Cho, Eva L. Duvaris, Coral Jackson, Catherine E. Shanks, Sarah Y. Tan, Selvi Ramalingam, and Cassie S. Mitchell. 2025. [Trialsieve: A comprehensive biomedical information extraction framework for pico, meta-analysis, and drug repurposing](#). *Bioengineering*, 12(5).
- Anna Kirkpatrick, Chidozie Onyeze, David Kartchner, Stephen Allegri, Davi Nakajima An, Kevin McCoy, Evie Davalbhakta, and Cassie S. Mitchell. 2022. [Optimizations for computing relatedness in biomedical heterogeneous information networks: Semnet 2.0](#). *Big Data and Cognitive Computing*, 6(1).
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#).
- 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.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. [Readability controllable biomedical document summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Shufan Ming, Yue Guo, and Halil Kilicoglu. 2025. [Towards knowledge-guided biomedical lay summarization using large language models](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. [On the persona-based summarization of domain-specific documents](#).
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew,

Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakob Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lillian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu,

Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shiron Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghamman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yuri Malkov. 2024. [Gpt-4o system card](#).

Tabea Margareta Grace Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A. Horn, and Christoph M. Friedrich. 2024. [WisPerMed at BioLaySumm: Adapting autoregressive large language models for lay summarization of scientific articles](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 769–779, Bangkok, Thailand. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang

- Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Luca Ragazzi, Paolo Italiani, Gianluca Moro, and Mattia Panni. 2024. [What are you token about? differentiable perturbed top- \$k\$ token selection for scientific document summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9427–9440, Bangkok, Thailand. Association for Computational Linguistics.
- Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Saddam Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. [A review on large language models: Architectures, applications, taxonomies, open issues and challenges](#). *IEEE Access*, 12:26839–26874.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2026. [Medgemma technical report](#).
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs](#).
- Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaying Zhang, Yutao Xie, and Sheng Yu. 2022. [Biobart: Pretraining and evaluation of a biomedical generative language model](#).
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BARTScore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems*.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [Alignscore: Evaluating factual consistency with a unified alignment function](#).
- 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*.

A Appendix

A.1 Experimental setup

To ensure the reproducibility of our results, we describe in this section the experimental conditions adopted at all stages of our study.

- **Inference:** All inferences for abstract and lay summary generation with lightweight models were conducted on a single Nvidia A40 GPU. We used the open-source Hugging Face Transformers library to facilitate inference. Each model was run five times on the eLife test set and four times on the PLOS test set, with variances calculated for all metrics. ROUGE scores and BERTScore were computed using the Hugging Face Evaluation library to ensure consistency. For FRE, CLI and DCRS, we used the `textstat` library, and for BARTScore and AlignScore, we relied on the authors' original implementation available on GitHub with 'facebook/bart-large-cnn' pretrained weights for BARTScore and 'AlignScore-large' for AlignScore. For abstracts, we consider readability scores closest to the target values as top scores, since the goal is not necessarily to maximize technicality. In contrast, for lay summaries, we define top scores as those achieving the highest FRE and the lowest CLI and DCRS, reflecting the aim to produce more accessible, lay-oriented text. Inference with GPT-4o was performed via the OpenAI Python API on a single run, which requires an API key. All models were run with the following parameters: temperature = 1, top p = 1.
- **Result analysis and ranking evaluation:** We selected Google's Gemini 2.0 Flash as both the LLM Analyzer and LLM Judge. Due to the limited size of our sampled test set, we used the web interface directly.
- **Human evaluators:** We recruited a cohort of 80 undergraduate volunteers with a background in biomedical sciences to rank both abstracts and lay summaries, using the same instructions and criteria provided to the LLM Judge (Figure 12 and Figure 13). Each human judge received the same sample of 40 documents (20 from eLife and 20 from PLOS) and was tasked with ranking five abstracts and five lay summaries for each document according

to the specified criteria. Judges had a time window of three days to complete the task and were free to work at their own pace.

A.2 Prompt setup for LLM Generator

We used different prompts depending on the task. Figure 7 shows an example prompt provided to the LLM Generator for long abstract generation. Similarly, Figure 8 shows the prompt used for generating a short lay summary. The prompts are structurally similar across short and long formats, differing only in the required output length.

A.3 Prompt setup for LLM Analyzer

We employed a multi-turn prompting strategy: first, we presented the article with section headers, followed by the task instructions and the summary to analyze. We used two prompts with the LLM Analyzer—one for sentence type classification and another for the question answering setup. Figure 9 shows an example of the sentence classification task, while Figure 10 illustrates the QA task.

A.4 Prompt setup for LLM Judge

We used a similar multi-turn prompting strategy for the LLM Judge. In this case, the model was presented with five summary candidates and two evaluation instructions. The specific criteria used to rank abstracts and lay summaries are shown in Figure 12 and Figure 13, respectively.

A.5 Additional tables for generated summaries and targets

During inference, we generate two summary lengths for both lay summaries and abstracts. A long lay summary is limited to approximately 500 words, while a short lay summary is limited to 200 words. Similarly, a long abstract is capped at 300 words, and a short abstract at 150 words. These length choices are motivated by the average summary lengths observed in the target test sets (eLife and PLOS), as detailed in Table 9.

| Model | eLife | | PLOS | |
|-----------------------|--------|--------|--------|--------|
| | Short | Long | Short | Long |
| Llama-3.2-3B-Instruct | 265.45 | 399.93 | 228.89 | 415.53 |
| Llama-3.1-8B-Instruct | 317.97 | 465.16 | 299.63 | 416.38 |
| Qwen 25-3B-Instruct | 179.70 | 278.48 | 176.78 | 280.52 |
| Qwen 25-7B-Instruct | 187.28 | 322.27 | 189.16 | 278.83 |

Table 7: Average number of words in generated lay summaries.

| Model | eLife | PLOS |
|-----------------------|---------------|---------------|
| | Short / Long | Short / Long |
| Llama-3.2-3B-Instruct | 176.14/210.21 | 162.39/205.99 |
| Llama-3.1-8B-Instruct | 179.55/224.57 | 156.37/199.08 |
| Qwen 25-3B-Instruct | 119.94/173.22 | 116.23/165.90 |
| Qwen 25-7B-Instruct | 117.38/199.08 | 117.92/180.68 |

Table 8: Average number of words in generated abstracts.

| Dataset | Lay Summary (avg \pm std) | Abstract (avg \pm std) |
|---------|-----------------------------|--------------------------|
| eLife | 383.09 \pm 66.04 | 166.28 \pm 21.03 |
| PLOS | 194.70 \pm 38.48 | 269.32 \pm 67.91 |

Table 9: Average and standard deviation of reference summary lengths (in words) across eLife and PLOS test sets for lay summaries and abstracts.

| Dataset | Avg. questions | Summary type | Correct frequency |
|---------|----------------|-------------------|-------------------|
| eLife | 9.8 | Long lay summary | 0.80 |
| | | Short lay summary | 0.68 |
| | 8.9 | Long abstract | 0.75 |
| | | Short abstract | 0.60 |
| Plos | 8.4 | Long lay summary | 0.70 |
| | | Short lay summary | 0.58 |
| | 10.4 | Long abstract | 0.49 |
| | | Short abstract | 0.46 |

Table 10: Question answering statistics for questions generated from the reference summaries in each subset of PLOS and eLife.

A.6 Additional figure for human evaluation

This section provides supplemental visualizations from the human evaluation phase. Figure 5 illustrates the explicit ranking distribution for summaries generated by the base models, offering a granular view of how often each model was preferred relative to others. Additionally, Figure 6 details the rater classification distribution. These distributions further elucidate the comparative performance and perceived quality of the summaries beyond aggregate mean scores.

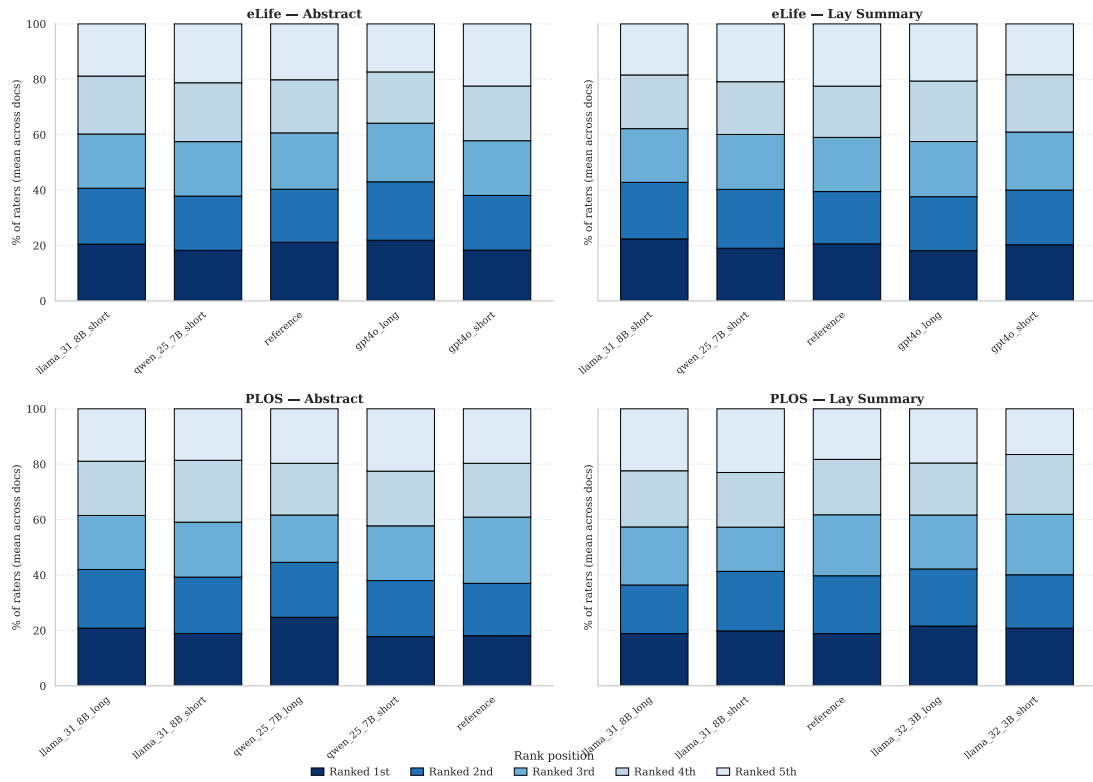


Figure 5: Explicit ranking distribution for each given summaries during human evaluation.

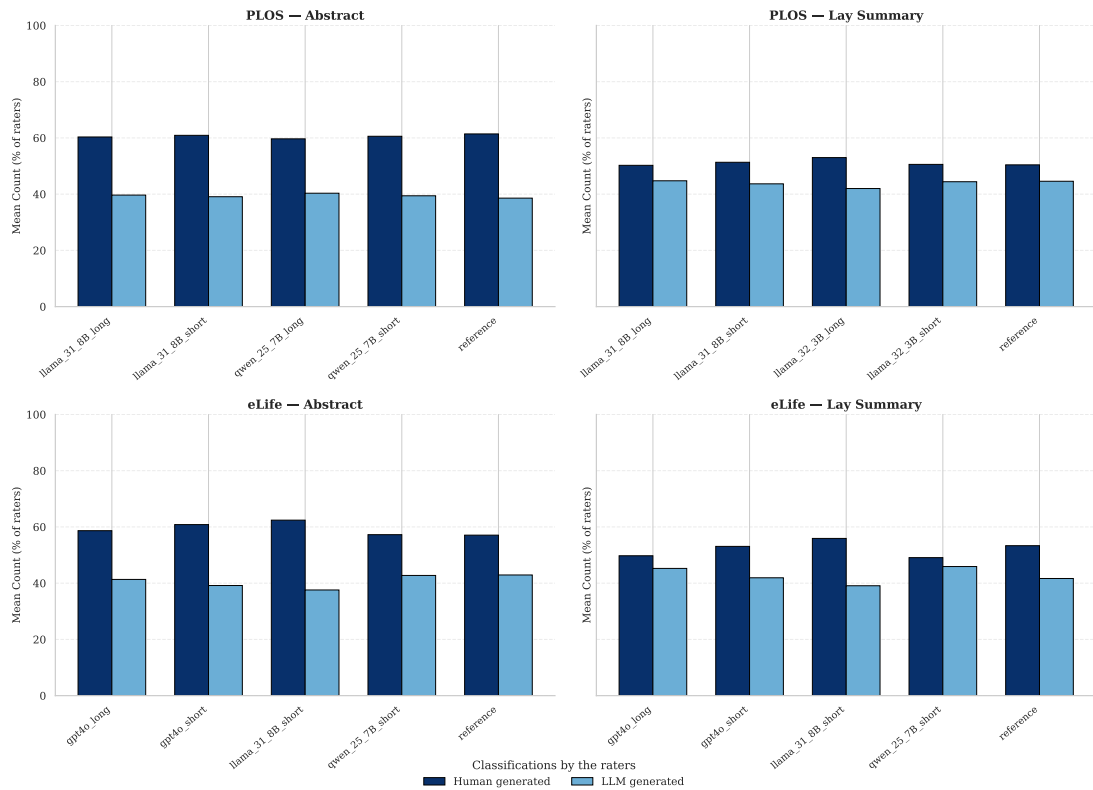


Figure 6: Detailed distribution of rater classifications for each model's generated summaries during the human evaluation.

A.7 Pilot study on domain-specific lightweight models

To evaluate the performance of general-purpose LLMs against domain-specific counterparts, we extended our evaluation of lay and abstractive summary generation to three specialized biomedical models: MedGemma-4B-it (Sellergren et al., 2026), MedReason-8B (Wu et al., 2025), and HuatuoGPT-o1-7B (Chen et al., 2024).

Notably, MedReason-8B and HuatuoGPT-o1-7B are specialized medical reasoning models that generate an internal reasoning trace prior to producing a final response. For these models, we extracted only the final generated answer for evaluation, discarding the preceding reasoning steps to ensure a fair comparison with the non-reasoning base models. We employed the same experimental configurations as the primary study, specifically targeting long lay summaries (approx. 500 words) and long abstracts (approx. 300 words). Inference was conducted using the vLLM library (Kwon et al., 2023) with a temperature of 0 and a maximum generation length of 2048 tokens. All experiments were performed in a single pass on an NVIDIA A40 GPU. Results are detailed in Tables 11 and 12.

Key Findings Overall, domain-specific models did not yield higher relevance scores than the base models used in our main study. However, they surprisingly achieved readability scores (CLI, DCRS, and FRE) closer to the reference summaries, whereas base models exhibited a broader variance in these metrics. Additionally, specialized models demonstrated high factual alignment—as measured by BARTScore and AlignScore—comparable to the average performance of general-domain models. These results suggest that while domain tuning may not significantly boost raw relevance, it may offer better calibration toward human-like readability. Further qualitative content analysis or human evaluation is warranted to fully discern the nuances between specialized and general-domain summaries and to determine user preference.

| | Rouge Metrics ↑ | | | BERTscore ↑ | FRE ↑ | CLI ↓ | DCRS ↓ | BARTscore ↓ | Alignscore ↑ |
|----------------------------|-----------------|-------|-------|-------------|-------|-------|--------|-------------|--------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| MedGemma-4B-it | 40.90 | 09.18 | 18.07 | 83.45 | 44.18 | 14.19 | 09.16 | -2.89 | 67.26 |
| HuatuoGPT-o1-7B | 37.71 | 08.49 | 17.17 | 84.24 | 35.03 | 15.84 | 10.62 | -2.59 | 68.90 |
| MedReason-8B | 31.09 | 07.04 | 15.62 | 83.51 | 33.42 | 15.92 | 10.59 | -2.27 | 75.84 |
| eLife Lay summaries | - | - | - | - | 51.83 | 12.50 | 08.93 | -3.28 | 43.42 |
| PLOS Dataset | | | | | | | | | |
| MedGemma-4B-it | 37.18 | 10.01 | 18.76 | 84.19 | 41.87 | 14.66 | 09.44 | -2.75 | 73.08 |
| HuatuoGPT-o1-7B | 40.60 | 10.68 | 20.04 | 85.72 | 35.51 | 15.65 | 10.61 | -2.47 | 73.57 |
| MedReason-8B | 36.53 | 09.93 | 19.16 | 84.87 | 34.90 | 15.67 | 10.01 | -2.38 | 76.26 |
| PLOS lay summaries | - | - | - | - | 30.03 | 15.90 | 10.91 | -2.94 | 65.48 |

Table 11: Results for long lay summaries generation on eLife/PLOS test sets with domain specific models (↑:higher is better / ↓:lower is better).

| | Rouge Metrics ↑ | | | BERTscore ↑ | FRE ↑ | CLI ↓ | DCRS ↓ | BARTscore ↓ | Alignscore ↑ |
|------------------------|-----------------|-------|-------|-------------|-------|-------|--------|-------------|--------------|
| | R-1 | R-2 | R-L | | | | | | |
| eLife Dataset | | | | | | | | | |
| MedGemma-4B-it | 37.75 | 11.54 | 19.69 | 85.36 | 28.68 | 16.92 | 11.13 | -2.38 | 81.60 |
| HuatuoGPT-o1-7B | 40.26 | 10.75 | 20.25 | 85.87 | 26.03 | 17.23 | 11.65 | -2.59 | 70.93 |
| MedReason-8B | 34.05 | 10.07 | 18.43 | 85.04 | 29.90 | 16.38 | 11.53 | -2.37 | 81.86 |
| eLife abstracts | - | - | - | - | 23.72 | 17.53 | 11.87 | -3.50 | 63.42 |
| PLOS Dataset | | | | | | | | | |
| MedGemma-4B-it | 40.53 | 13.34 | 20.94 | 85.37 | 27.92 | 17.03 | 11.32 | -2.38 | 83.25 |
| HuatuoGPT-o1-7B | 40.45 | 11.37 | 19.98 | 85.32 | 26.01 | 17.20 | 11.62 | -2.51 | 75.55 |
| MedReason-8B | 35.00 | 11.30 | 18.23 | 84.45 | 28.60 | 16.32 | 11.29 | -2.25 | 83.93 |
| PLOS abstracts | - | - | - | - | 27.55 | 16.39 | 11.06 | -3.33 | 63.40 |

Table 12: Results for long abstracts generation on eLife/PLOS test sets with domain specific models (↑:higher is better / ↓:lower is better).

User prompt :

"You are a skilled physician able to understand biomedical documents and write their Abstract. An abstract provides an overview of the study's key aspects. It typically includes: Background, Objective, Methods, Results, and Conclusions. Don't mention those sections in your abstract. The abstract should be less than 300 words."

"Provide an Abstract for the following article: \nMost movements are comprised of sequences . From the complex routines that gymnasts perform to intricate dance numbers to simply reaching for an object , our movements are comprised of sequences of movements that are learned through practice . The cerebellum has been long implicated in learning and execution of accurate movements (Doyon et al . , 1997; Shin and Ivry , 2003; Lehericy et al . , 2005; Krupa et al . , 1993; Lisberger , 1994; Diener and Dichgans , 1992) . Movement sequences as well as multi-joint movements are particularly sensitive to cerebellar dysfunction (Shin and Ivry , 2003; Diener and Dichgans , 1992; Doyon et al . , 2002)"

LLM Generator output:

"The cerebellum plays a crucial role in the learning and execution of movement sequences. While previous studies have focused on single-component movements, this study investigates the cerebellar mechanisms underlying the learning of movement sequences. The researchers used a simple elaboration of cerebellar mechanisms that mediate single-component movements to explain cerebellar learning and implementation of movement sequences. They tested the hypothesis that a feedback signal (FS) from one movement component can serve as a cue for the next component in the sequence. Using electrical stimulation of mossy fibers as a cue..."

Figure 7: Prompt example for an eLife article in an abstract generation case followed by the LLM Generator output.

User prompt :

"You are a skilled physician able to understand biomedical documents and write their Lay summary. A Lay summary is a brief summary of a research article that is written in a way that a non-expert will understand. It should be written in plain English and be easy to read. The lay summary should be less than 200 words."

"Provide a Lay summary of the article:

Seasonal epidemics of influenza virus result in ≈ 36000 deaths annually in the United States . Current vaccines against influenza virus elicit an antibody response specific for the envelope glycoproteins. However , high mutation rates result in the emergence of new viral serotypes , which elude neutralization by preexisting antibodies"

LLM Generator output:

"Every year, the flu virus causes about 36,000 deaths in the United States. Current flu vaccines work by creating antibodies that specifically target the flu virus's outer proteins. However, the flu virus is constantly changing, making it difficult for vaccines to keep up. Researchers have been searching for a "universal" flu vaccine that can protect against multiple flu strains.\n\nIn this study, scientists tested a new approach to creating a universal flu vaccine. They found that a combination of two immune responses - one from a specific type of white blood cell called a CD8+ T cell and another from non-neutralizing antibodies - can provide complete protection against the flu virus.\n\nThe CD8+ T cells recognize internal proteins of the flu virus, while the non-neutralizing antibodies target the virus's outer protein...."

Figure 8: Prompt example for an eLife article in a lay summary generation case followed by the LLM Generator output.

User prompt :

"Here is an article from PLOS dataset:

Introduction

The gastrointestinal tract hosts a huge number of bacteria species . In mice , these bacteria play a major role in shaping local and systemic immune responses , notably by the induction of different effector and regulatory T cell subsets , whose adequate balance is required for the maintenance of gut homeostasis [1] . It may be postulated that this balance relies on the microbiota composition , as suggested by the observation that dysbiosis are frequent in chronic immune disorders , especially Inflammatory Bowel Disease (IBD) [2]-[4]"

User prompt :

"Here is an abstract generated by Llama 3.1-8B-Instruct. For each sentence in the abstract, I want you to tell me from which section of the article it is taken from, only give the section that fits the best:

A novel subset of T cells, referred to as double-positive CD4CD8 $\alpha\alpha$ (DP8 α) T cells, was identified in the human colonic lamina propria and blood. These cells lack the transcription factor Foxp3 but share many regulatory functions with Foxp3 Treg, including the secretion of IL-10 and the inhibition of CD4 T cell proliferation and dendritic cell maturation. Importantly, a significant fraction of DP8 α T cells specifically recognize F..."

LLM Generator output:

"Here is the breakdown of each sentence in the generated abstract, indicating the section of the article it most likely originates from:

"A novel subset of T cells, referred to as double-positive CD4CD8 $\alpha\alpha$ (DP8 α) T cells, was identified in the human colonic lamina propria and blood." - Introduction (This sentence introduces the key finding of the study and the location of these cells, which is a primary focus of the introduction.)

"These cells lack the transcription factor Foxp3 but share many regulatory functions with Foxp3 Treg, including the secretion of IL-10 and the inhibition of CD4 T cell proliferation and dendritic cell maturation." - Results(These functional characteristics and the lack of Foxp3 are key findings presented in the "Results" section, particularly in figures 2 and 3.)"

Figure 9: Prompt example for a PLOS article in a sentence type classification task.

User prompt :

"Here is an article from PLOS dataset:

Introduction

The gastrointestinal tract hosts a huge number of bacteria species . In mice , these bacteria play a major role in shaping local and systemic immune responses , notably by the induction of different effector and regulatory T cell subsets , whose adequate balance is required for the maintenance of gut homeostasis [1] . It may be postulated that this balance relies on the microbiota composition , as suggested by the observation that dysbiosis are frequent in chronic immune disorders , especially Inflammatory Bowel Disease (IBD) [2]-[4]"

User prompt :

"Now that you have the full Plos article with section Headings, generate relevant questions from the original article. Ask general questions and section specific questions."

LLM Generator output:

" General Questions:

What is the primary focus of this research article?

What were the main findings of this study regarding DP8α T cells?

...

Section-Specific Questions:

Introduction:

According to the introduction, what is the established role of gut bacteria in mice regarding immune responses?

What knowledge gap regarding microbiota-induced regulatory T cells in humans does the introduction highlight?..."

User prompt:

"Take the questions you generated. Now, based on the informations provided ONLY AND ONLY by the following long abstract generated by Llama 3.1-8B-Instruct, try to answer as much previous questions as possible. If you cannot simply say that you cannot."

LLM Generator output:

"What is the primary focus of this research article? The primary focus is on a newly identified subset of T cells, DP8α T cells, and their potential role in the context of gut microbiota, inflammation, and Inflammatory Bowel Disease (IBD).

What were the main findings of this study regarding DP8α T cells? The main findings are the identification of DP8α T cells that lack Foxp3 but share regulatory features with Foxp3 Treg (IL-10 secretion, inhibition of CD4 T cell proliferation and dendritic cell maturation), their specific recognition of *F. prausnitzii*, and their reduced frequency in IBD patients...."

Figure 10: Prompt example for a PLOS article in a question answering task.

User prompt :
 "Here is an article from PLOS dataset:
 Introduction
 The gastrointestinal tract hosts a huge number of bacteria species . In mice , these bacteria play a major role in shaping local and systemic immune responses , notably by the induction of different effector and regulatory T cell subsets , whose adequate balance is required for the maintenance of gut homeostasis [1] . It may be postulated that this balance relies on the microbiota composition , as suggested by the observation that dysbiosis are frequent in chronic immune disorders , especially Inflammatory Bowel Disease (IBD) [2]-[4]"

User prompt :
 "Objective:
 You will be presented with 5 scientific abstracts. Your task is to rank them from 1st (best) to 5th (worst) based on how well they communicate the core elements of a scientific study.
 Additionally, try to guess which ones were generated by an LLM.
 We consider the following definition for an abstract:
 "A biomedical abstract is a concise, self-contained summary of a scientific study that clearly communicates the study's background, objective, methods, main results, and conclusions, with a focus on clinical, biological, or translational significance. It should allow readers (clinicians, researchers, or policymakers) to quickly understand the study's purpose, how it was conducted, and why it matters.

 "

User prompt :
 "Abstract 1

 Abstract 2
 ...
 Abstract 5"

LLM Generator output:
 "Ranking:
 Abstract 1
 Abstract 4
 Abstract 3
 Abstract 5
 Abstract 2
 Reasoning and LLM Guesses:
 Abstract 1: This abstract is very well-structured. It clearly lays out the background (gut microbiota, IBD), the problem (unknown mechanisms, human Treg), the method (identification of DP8α T cells), key results (lack Foxp3, regulatory functions, *F. prausnitzii* induction, reduction in IBD), and the conclusion/contribution (novel Treg, potential for IBD control, diagnostic/therapeutic strategies). It's clear, concise, and informative. I suspect this might be human-written.
 Abstract 1...."

Figure 11: Prompt example for a PLOS article presented with 5 abstracts in a LLM Judge setup.

Ideal abstract instructions :

Objective:

You will be presented with 5 scientific abstracts. Your task is to rank them from 1st (best) to 5th (worst) based on how well they communicate the core elements of a scientific study.

Additionally, try to guess which ones were generated by an LLM.

We consider the following definition for an abstract:

"A biomedical abstract is a concise, self-contained summary of a scientific study that clearly communicates the study's background, objective, methods, main results, and conclusions, with a focus on clinical, biological, or translational significance. It should allow readers (clinicians, researchers, or policymakers) to quickly understand the study's purpose, how it was conducted, and why it matters."

Please base your ranking on the following five criteria:

Clarity and Readability

Is the abstract written clearly and concisely ?

Is the language understandable with some domain knowledge (specific jargon) ?

Motivation and Problem Definition

Does the abstract clearly state what problem is being addressed and why it matters ?

Methods Description

Is the approach or method briefly but adequately described ?

Do you understand what was done to solve the problem ?

Key Results and Findings

Are the results presented in a way that is informative and meaningful ?

Are quantitative outcomes (if applicable) mentioned ?

Conclusion and Contribution

Does the abstract clearly state the conclusion and/or impact of the study ?

Is the contribution to the field evident ?

Figure 12: Criteria used to rank abstracts with an LLM judge and human judges.

Ideal lay summary instructions :

Objective:

You will be presented with 5 scientific lay summaries. Your task is to rank them from 1st (best) to 5th (worst) based on how well they communicate the core elements of a scientific study.

Additionally, try to guess which ones were generated by an LLM.

We consider the following definition for a lay summary:

“A biomedical lay summary is a brief, accessible explanation of a scientific study written for a non-specialist audience. It translates complex biomedical research findings into plain language that can be understood by patients, caregivers, policymakers, or the general public — without sacrificing the core message or significance of the work.”

example:

Scientific Abstract:

“We performed a double-blind, placebo-controlled trial of a novel IL-6 inhibitor in 178 RA patients, measuring changes in DAS28 scores over 12 weeks...”

Lay Summary:

“We tested a new drug to reduce joint pain and swelling in people with rheumatoid arthritis. The study showed that the drug worked better than a placebo and helped patients feel better after three months...”

Please base your ranking on the following five criteria:

Clarity and Simplicity of Language

Is the summary written in plain, understandable language ?

Are technical terms avoided or clearly explained ?

Explanation of the Problem / Context

Does the summary clearly state the health or scientific issue being addressed ?

Is it easy to understand why this research matters ?

Description of What Was Done

Is it clear what kind of research was conducted ?

Is the study design or approach described in simple terms ?

Communication of Key Findings

Are the main results explained in a way that a layperson can understand ?

Are the findings accurate, without exaggeration ?

Relevance and Impact

Does the summary explain the potential real-world implications (e.g., for patients, policy, or future research) ?

Does it feel relevant or important to everyday life ?

Figure 13: Criteria used to rank lay summaries with an LLM judge and human judges.