

Retrieval-Augmented Simplification of Sectioned Biomedical Abstracts

Pascal Mathas and Jan Bakker and Jaap Kamps

Institute for Logic, Language and Computation (ILLC)

University of Amsterdam

Amsterdam, The Netherlands

pascal.mathas@student.uva.nl, {j.bakker, j.kamps}@uva.nl

Abstract

Cochrane systematic reviews pair each technical abstract with a plain-language summary (PLS), but most biomedical papers lack a PLS. We propose a retrieval-augmented generation (RAG) pipeline that generates Cochrane PLS from biomedical abstracts. The pipeline exploits the section-level alignment of the COCHRANE-SECTIONS dataset in two ways. First, it provides the LLM with abstract-PLS section pairs from topically related reviews, giving more granular in-context examples than whole abstract-PLS pairs. Second, each retrieved PLS section is split into its rephrased and inserted sentences, showing the LLM both the rephrasing and the insertion behavior it should reproduce. Our pipeline outperforms a non-RAG LLM and three RAG ablations on the COCHRANE-SECTIONS test set, with the same gain trend across three different LLM families.

1 Introduction

In order to answer a specific (health-related) question, systematic reviews aim to find, evaluate, and combine all available empirical evidence from related studies (Bakker and Kamps, 2025). The Cochrane Database of Systematic Reviews (CDSR) pairs every technical abstract with a plain-language summary (PLS), intended to be readable without medical training, so that the review becomes accessible to a broader audience. This can be tedious work, making an automated approach appealing. Generating a PLS from an abstract is, however, non-trivial, as large parts of the abstract have to be rephrased into simpler language, and additional background content that is not in the abstract has to be inserted into the PLS (Guo et al., 2024).

Earlier work has stayed at the document- or sentence-level, whether on whole abstract-PLS pairs (Guo et al., 2021), on the results and conclusions only (Devaraj et al., 2021; Bakker and Kamps,

2024), or on aligned sentence pairs (Joseph et al., 2023). Bakker and Kamps (2025) address this by releasing COCHRANE-SECTIONS, the first dataset in which the PLS of 7.7K Cochrane reviews are divided into the same seven sections as their parallel abstracts.

In this paper, we use COCHRANE-SECTIONS to build a retrieval-augmented generation (RAG) pipeline for the section-level simplification of Cochrane abstracts. The main contributions of our paper are as follows:

- We propose a RAG pipeline for the generation of Cochrane PLS that uses abstract-PLS section pairs from topically related reviews as in-context examples.
- We split each retrieved PLS section into its rephrased and inserted sentences, and present both parts separately in the prompt, showing the model which source content to keep and what background to add.
- We show that our pipeline outperforms four strong baselines on the COCHRANE-SECTIONS test set, with the same gain trend across three different LLM families.

We make all code, prompts, and settings publicly available at github.com/pascalmathas/cochrane_sections_rag.

2 Related Work

Retrieval-augmented text simplification. A closely related work on RAG pipelines for lay language generation is Guo et al. (2024). The authors focus on what they call background explanation, content that is absent from the complex abstract and thus poses a challenge for simplification, using CELLS, a corpus of abstract-lay summary pairs from 12 biomedical journals. They tackle this by

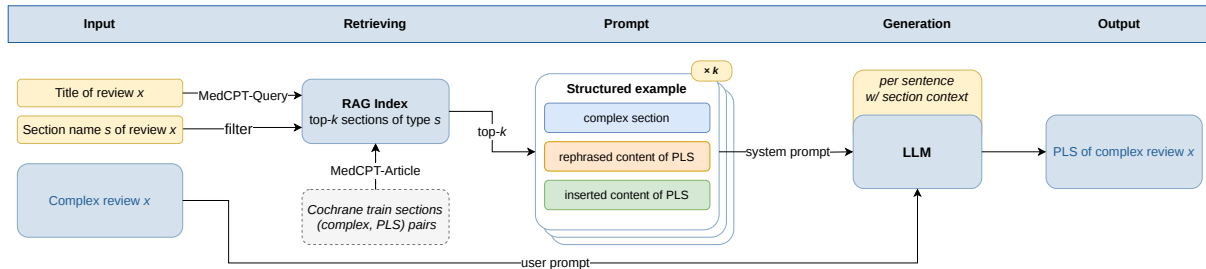


Figure 1: Overview of the RAG pipeline for the simplification of sectioned biomedical abstracts. For each section of a complex abstract that the LLM is asked to simplify, we retrieve the top- k most similar reviews from the RAG index and take from each the matching section. Every retrieved example is then put in the prompt as the complex abstract section together with its PLS section, which is split into rephrased and inserted sentences.

retrieving external knowledge such as term definitions from UMLS and Wikipedia. You et al. (2024) take a similar approach, incorporating Wikipedia knowledge into their biomedical lay summarization pipeline through an extract-then-summarize method. The addition of content when simplifying text is also investigated by Srikanth and Li (2021), who introduce elaborative simplification, a simplification method that inserts clarifications, explanations, or definitions. We address the same insertion problem, but instead of external knowledge, we retrieve closely related abstract-PLS section examples into the prompt, split into rephrased and inserted sentences.

Dataset. Our RAG pipeline explicitly relies on the two features of the COCHRANE-SECTIONS dataset introduced by Bakker and Kamps (2025). First, the section-level alignment between abstracts and PLS lets us use section-level in-context examples rather than full abstract-PLS pairs. Second, the underlying sentence-level alignments let us split each retrieved PLS section into rephrased content (the sentences aligned to the abstract) and inserted content (the sentences added only in the PLS). Both are used as inputs to our pipeline.

3 Method

3.1 Dataset

The RAG pipeline of this paper is built and evaluated on the COCHRANE-SECTIONS dataset (Bakker and Kamps, 2025). The data consists of 7.7K abstract-PLS pairs aligned at the section level in English, derived from 7,703 Cochrane systematic reviews in the CDSR. Each entry consists of a technical abstract, intended for healthcare decision-makers, and a PLS, written to be understandable to readers without a university education.

The technical abstracts are divided into seven sections: *Background*, *Objectives*, *Search methods*, *Selection criteria*, *Data collection and analysis*, *Main results*, and *Authors' conclusions*. The corresponding PLS of these reviews are, however, written more freely and do not adhere to the same structure. Bakker and Kamps (2025) address this misalignment with a two-step procedure that first aligns each PLS sentence to a sentence in the parallel abstract, based on the neural CRF aligner of Jiang et al. (2020), and then classifies the remaining unaligned sentences into one of the seven section labels.

The resulting dataset provides, for each review, a PLS divided into the same seven sections as its corresponding technical abstract, thereby creating parallel abstract-PLS pairs at the section level. We use the same train/validation/test split as Bakker and Kamps (2025), who adopted it from MULTICOCHRANE (Joseph et al., 2023). This gives us 7,676 training pairs, 5 validation pairs, and 22 test pairs. Our RAG index combines the manually aligned part (74 train pairs) with the automatically aligned part (7,602 train pairs). We additionally evaluate on the CLEF 2026 SimpleText test data (Ermakova et al., 2026). This data contains 176 Cochrane reviews covering all seven sections, published after COCHRANE-SECTIONS was constructed.

3.2 Pipeline

Figure 1 provides an overview of our COCHRANE-SECTIONS RAG pipeline. The COCHRANE-SECTIONS dataset allows us to generate PLS from technical abstracts in two novel ways. First, instead of using whole abstract-PLS pairs as in-context examples, we use individual abstract-PLS section pairs, which gives the LLM more granular and

topically aligned simplification examples. Second, within each retrieved PLS section, we can split the sentences into two groups: the sentences aligned to the abstract, which we treat as rephrasings of the technical content, and the sentences that required a classification pass into the seven sections, which we treat as content that was added only in the PLS. A PLS sentence is thus rephrased if it has a counterpart in the abstract, and inserted if it does not. Rephrased sentences show which abstract content to preserve, and inserted sentences show what background to add. These behaviors correspond directly to the keep and add components of SARI.

During inference, given a sentence from a section of a new complex abstract, we retrieve the same section from the top- k most similar Cochrane reviews in the training split. For each retrieved example, we include in the prompt the complex abstract section, the PLS sentences that were aligned to it (the rephrased content), and the PLS sentences that were assigned to it through the classification step (the inserted content). This way, the LLM is shown section-level examples of both the rephrasing and the insertion behavior it should reproduce when writing the PLS for the target section. The LLM is asked to simplify sentence by sentence, using the section as context. The full prompts can be found in Appendix H, or on [GitHub](#).

4 Experiments

4.1 Setup

RAG. For embedding the training pool and the test reviews at inference time, we use the MedCPT article and query encoders ([Jin et al., 2023](#)). For both, we embed the title of the review, as we found that this gave the best performance (see Table 2 in Appendix A). At inference, we retrieve the top- k reviews based on cosine similarity, take from each the section we are currently simplifying, and split that section into rephrased and inserted sentences. We skip reviews where this section is empty. We choose $k = 4$ as this gives the best results (see Table 3 in Appendix B).

Generation. For generation we use Qwen2.5-72B-Instruct ([Qwen Team et al., 2025](#)), run with the vLLM library ([Kwon et al., 2023](#)) for faster inference and lower memory use. We run the model in bfloat16 and set the temperature to 0.2, top- p to 0.9, repetition penalty to 1.1, and max new tokens to 2048. In order to investigate whether our findings generalize to different LLM families,

we also run the same setup with Gemma-3-27b-it ([Gemma Team et al., 2025](#)) and Llama-3.1-70B-Instruct ([Grattafiori et al., 2024](#)) and present their results in Appendix G.

Metrics. We evaluate our simplifications by regrouping the generated sentences back into their documents and calculating several document-level metrics. We use two reference-based metrics: (a) SMART, which compares the output to the reference at the sentence level, matching sentences against each other ([Amplayo et al., 2023](#)), and (b) SARI, which assesses simplification quality by comparing n-gram operations (add, keep, delete) between the output, source, and reference ([Xu et al., 2016](#); [Alva-Manchego et al., 2019](#)). Additionally, we use the reference-free Flesch-Kincaid Grade Level (FKGL), which measures text readability ([Kincaid et al., 1975](#)). Due to the small test set of COCHRANE-SECTIONS, we evaluate across three runs, with three different seeds.

Human evaluation. We additionally conduct a small human evaluation, in which we show participants the complex section of a review together with the simplifications of both the baseline LLM and the RAG retrieved-split ($k = 4$) variant, and ask them to choose which one is the better simplification or to indicate no preference, with an optional explanation field. The two simplifications are shown side by side as *A* and *B*. For each participant, we shuffle the order of the items and randomly swap the ordering of *A* and *B*. We also provide the option to expand the reference simplification. We filter the test set in several ways. The *Search methods* and *Data collection and analysis* sections are dropped because both models leave them empty equally often, and we additionally remove sections where one model has no output or produces only a single sentence, where the two outputs are nearly identical, or that cover sensitive health topics we preferred not to show participants. We sample up to 3 items per section across the 5 remaining sections. Because some sections contain only 2 items after filtering, this yields 13 items in total rather than 15. Each of our 6 participants judges all 13 items. Participants are unpaid volunteers from the authors’ academic environment, not authors of this paper, who were told their judgments would compare two simplification systems. No personal data was collected. We report the percentage of times that the RAG retrieved-split variant was preferred over the baseline, together with a 95% confidence interval

(Wilson, 1927). We also report the inter-annotator agreement with Fleiss’ κ (Fleiss, 1971), computed over the preferred system.

Computational requirements. We run Qwen2.5-32B-Instruct and Gemma-3-27b-it with a single H100 80GB GPU (32B and 27B parameters), and Llama-3.1-70B-Instruct with two H100 80GB GPUs (70B parameters). With each model, we generate 15 runs (5 systems, 3 seeds each), for a total of 45 runs. Generating simplifications for a single system across 3 seeds takes ~ 10 minutes.

Reproducibility. All experiments in this paper are run with fixed random seeds and deterministic Torch and CUDA behavior to ensure reproducibility. For a detailed overview of all parameters for every system and library versions, refer to the [README](#), [shell scripts](#), and [requirements](#) in our GitHub repository. For library versions and the licenses of any artifacts used in this paper, see Appendices C and D.

4.2 Baselines

We run four baselines to assess the performance of our RAG pipeline, each one removing a part of the pipeline. First, we run a non-RAG LLM simplification baseline, where the LLM is asked to simplify the abstract sentence by sentence, given the section context, without any retrieved examples. Second, we keep the same prompt but replace the retrieved examples with k random examples from the training pool to check whether semantically similar examples are actually beneficial (denoted as -rand). Third, we keep the retrieval but give the LLM the whole PLS section of each example, to check whether splitting it into rephrased and inserted content makes a difference (denoted as -full). Last, we provide both k random examples and show the whole PLS section.

4.3 Results & Discussion

Table 1 presents the document-level results of our pipelines. We report SARI together with its add, keep, and delete components, as these indicate which of the three n-gram operations drive the differences between systems.

First, we can see that the SMART metric alone is not the best indicator of simplification quality, as the complex input actually achieves a higher SMART F1 (35.16) than any of the generated simplifications (31.01 to 33.21). This is expected, as

SMART is made to measure semantic similarity, and not simplification quality. We include it because it is still useful for comparing the simplifications to each other.

Second, FKGL is roughly constant across systems (10.19 to 10.51), well below the 13.00 of the input and close to the 11.39 of the reference, so the LLM lowers the grade level regardless of whether retrieved examples are added.

Providing random abstract-PLS pairs (RAG rand-) in the prompt, whether they are split into rephrased and inserted or not, has little effect on the simplification quality, besides making the outputs slightly longer. SARI moves only marginally over the non-RAG LLM baseline (43.69 to 43.78 and 43.74), and the underlying add and keep components are similarly flat. This means that the in-context examples have to be topically related to the target abstract to be useful, as dissimilar PLS sections do not provide the model with simplification patterns it can transfer, whether shown whole or split into their rephrased and inserted parts.

If we retrieve closely related section examples, we can see that simplification quality clearly improves, with SARI rising from 43.78 (rand-full) to 44.26 (retrieved-full). This gain is driven by keep (+1.41) and add (+0.62), while delete remains roughly stable. Splitting these retrieved sections into rephrased and inserted content boosts simplification quality further, resulting in the best overall scores across SMART F1 (33.21), SARI (45.17), add (7.83), and keep (37.45). The improvement again comes from the keep and add components, which rise by another 1.81 and 0.61 points, respectively, over retrieved-full, while delete remains essentially flat. This behavior matches what the split was designed to do, as the rephrased sentences show the model which abstract content should be preserved in the PLS (driving keep), and the inserted sentences make it clear what kind of background content should be added in the PLS (driving add). The per-section SARI breakdown in Appendix E further shows that the gains are largest for the sections the LLM baseline handles worst without examples.

Comparing the outputs of the base LLM against the RAG retrieved-split pipeline, we find that the base LLM paraphrases the abstract sentences more freely than the RAG pipeline, which stays closer to the structural format of the PLS. The LLM often drops framing sentences that introduce a section or a specific measurement methodology, as well

System	SMART F1 $\uparrow$	FKGL $\downarrow$	SARI $\uparrow$				Length	
			add	keep	del	all	Tok.	Sent.
Input	35.16	13.00	0.00	23.16	0.00	7.72	864.6	25.64
Reference	100.00	11.39	100.00	100.00	100.00	100.00	363.8	14.68
LLM	31.01	10.25	6.46	33.65	90.96	43.69	396.0	18.00
RAG rand-full ($k = 4$)	31.52	10.30	6.60	34.23	90.50	43.78	425.0	19.14
RAG rand-split ($k = 4$)	31.58	10.51	6.51	34.33	90.38	43.74	428.5	18.91
RAG retrieved-full ($k = 4$)	32.53	10.19	7.22	35.64	89.92	44.26	446.5	20.44
RAG retrieved-split ($k = 4$)	33.21	10.29	7.83	37.45	90.24	45.17[†]	448.9	20.26

Table 1: Simplification results on the COCHRANE-SECTIONS test set, averaged over three runs with three different seeds. Best results are in **bold**. [†] marks results that are significantly better than every other system on per-review SARI (paired bootstrap (Koehn, 2004), $N = 1000$, $p < 0.01$).

as side-effect descriptions that should be kept in the PLS. The RAG retrieved-split pipeline instead keeps these elements and rewrites them in plain language, and occasionally replaces medical jargon with lay terms, for example “jaw surgery” instead of “orthognathic surgery”. Some examples of simplifications can be found in Table 10 in Appendix I.

Finally, the RAG gains generalize beyond our main setup. The same gain trend holds with both Gemma-3-27B-it and Llama-3.1-70B-Instruct, with all three model families gaining 1.3 to 1.6 SARI points over their non-RAG LLM baseline (Appendix G), and the pipeline also outperforms the non-RAG baseline on the larger, held-out CLEF 2026 SimpleText test set (Appendix F).

After analyzing the results of our human evaluation, we find that participants preferred the RAG retrieved-split pipeline over the LLM baseline 52.6% of the time (95% CI [41.6, 63.3]). This preference is not statistically significant, and inter-annotator agreement is negligible (Fleiss’ $\kappa = 0.08$). Preference varies by section and is strongest for the *Background* section, where the RAG pipeline is preferred 83.3% of the time (95% CI [55.2, 95.3]). The *Background* section relies most on inserted background content, so it makes sense that the RAG system is preferred, although its confidence interval is wide.

Even though statistically one system is not preferred over the other, the optional explanations that annotators left for their choice give qualitative insight into a trade-off between the two systems. Participants who preferred the LLM baseline described its output as more concise and easier to read, valuing its brevity. Participants who preferred the RAG pipeline instead valued its completeness, comment-

ing that the pipeline “explains terms in the text” and that the baseline LLM can leave out information, in one case introducing ambiguity by dropping the fact that an absence of evidence was due to a lack of studies rather than to studies that found no effect. This trade-off between the two systems can also be seen in their SARI scores, with RAG having high add and keep, reflecting that it inserts more explanatory background information and retains more of the source content than the LLM baseline.

5 Conclusion

In this paper, we have presented a retrieval-augmented generation pipeline for plain-language summaries of Cochrane systematic reviews. The pipeline retrieves topically related reviews and shows the LLM their abstract-PLS section pairs, with each retrieved PLS section split into its rephrased and inserted sentences. Our setup outperforms a non-RAG LLM and three RAG ablations on the COCHRANE-SECTIONS test set, with the same gain trend holding across three different LLM families. The gains also hold on the larger, held-out CLEF 2026 SimpleText test set, while a small human evaluation finds the two systems close in perceived quality, with annotators trading off the baseline’s brevity against our pipeline’s completeness. Future work could extend the pipeline to multilingual settings, supplement the retrieval pool with external encyclopedic knowledge, or move from sentence-level to document-level generation.

Limitations

Pipeline choices. We make two choices in our pipeline based on which configuration produced the best metric scores. The first is that we retrieve

based on the review title rather than the section or the whole review. Even though this yielded the best results, it is plausible that reviews are sometimes retrieved that are only loosely topically aligned. The second is that we simplify sentence by sentence, which may sacrifice consistency between sentences and sections, as the model has no knowledge of how it simplified previous sentences.

Dataset. Using the COCHRANE-SECTIONS dataset has advantages, but it also has some limitations. First, the test set contains only 22 reviews, which means that the performance of individual reviews can have a noticeable impact on the overall score, although we have attempted to mitigate this by averaging across three seeds, reporting paired-bootstrap significance, conducting a small human evaluation, and evaluating on the larger CLEF 2026 SimpleText test set. Second, our pipeline relies on two alignments: (a) the classification of the PLS into sections, and (b) the alignment of the PLS sentences to their complex counterparts. Both of these only operate at the sentence level, so we can only split the retrieved sections into rephrased and inserted content per sentence. These alignments are applied only to the training pool, not to the abstract being simplified at inference, so any noise affects only the in-context examples. That retrieved-split outperforms retrieved-full suggests the labels carry enough signal despite this. Finally, the COCHRANE-SECTIONS dataset is currently available only in English, and therefore our pipeline has been evaluated only in English.

Risk & ethics. With our work, we aim to make biomedical research more accessible to readers without medical training by lowering the cost of producing PLS. However, automatically generating PLS from complex medical reviews carries potential risks. First, LLMs can hallucinate content that is not in the source abstract. Second, the generated PLS are not medical guidance and should not be interpreted as such. Therefore, our pipeline is intended as research only and should not be used without additional human review.

Acknowledgments

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker is funded by the Netherlands Organization for Scientific Research (NWO

NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Statement on the use of Generative AI

During the preparation of this paper, we used generative AI for spelling, grammar, and polishing. All content, experiments, and claims are the authors' own.

References

- Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. [EASSE: Easier automatic sentence simplification evaluation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.
- Reinold Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2023. [SMART: sentences as basic units for text evaluation](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- M Aouad-Maroun, CK Raphael, SK Sayyid, F Farah, and EA Akl. 2016. [Ultrasound-guided arterial cannulation for paediatrics](#). *Cochrane Database of Systematic Reviews*, (9).
- Jan Bakker and Jaap Kamps. 2024. [Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.
- Jan Bakker and Jaap Kamps. 2025. [Section-level simplification of biomedical abstracts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13819–13833, Suzhou, China. Association for Computational Linguistics.
- R Brignardello-Petersen, A Carrasco-Labra, I Araya, N Yanine, L Cordova Jara, and J Villanueva. 2015. [Antibiotic prophylaxis for preventing infectious complications in orthognathic surgery](#). *Cochrane Database of Systematic Reviews*, (1).
- Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. [Paragraph-level simplification of medical texts](#). In *Proceedings of the 2021*

- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.
- Liana Ermakova, Hosein Azarbonyad, Jan Bakker, Gautam Kishore Shahi, Benjamin Vendeville, and Jaap Kamps. 2026. [CLEF 2026 SimpleText track: Simplify scientific text \(and nothing more\)](#). In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026*, volume 16486 of *Lecture Notes in Computer Science*, pages 215–224, Cham. Springer.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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.
- Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.
- Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. [Multilingual simplification of medical texts](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. 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. Technical report, Naval Technical Training Command, Millington, TN.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.
- 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. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- VM Musini, P Rezapour, JM Wright, K Bassett, and CD Jauca. 2015. [Blood pressure-lowering efficacy of loop diuretics for primary hypertension](#). *Cochrane Database of Systematic Reviews*, (5).
- M Nagendran, CD Toon, BR Davidson, and KS Gurusamy. 2014. [Laparoscopic surgical box model training for surgical trainees with no prior laparoscopic experience](#). *Cochrane Database of Systematic Reviews*, (1).
- H Prabhakar, GP Singh, C Mahajan, I Kapoor, M Kalaivani, and V Anand. 2016. [Intravenous versus inhalational techniques for rapid emergence from anaesthesia in patients undergoing brain tumour surgery](#). *Cochrane Database of Systematic Reviews*, (9).
- Qwen Team, 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, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- R Rajaratnam, J Halpern, A Salim, and C Emmett. 2010. [Interventions for melasma](#). *Cochrane Database of Systematic Reviews*, (7).
- A Saarto, JH Verbeek, MC Lavoie, and M Pahwa. 2011. [Blunt versus sharp suture needles for preventing percutaneous exposure incidents in surgical staff](#). *Cochrane Database of Systematic Reviews*, (11).
- Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137.
- Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLay-Summ: an extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143.

A Retrieval Embedding Source

Embed source	SMART F1 ↑	FKGL ↓	SARI ↑				Length	
			add	keep	del	all	Tok.	Sent.
Title	33.21	10.29	7.83	37.45	90.24	45.17	448.9	20.26
Section	32.94	10.45	7.41	36.85	90.03	44.76	454.8	20.18
Abstract	32.87	10.56	7.44	35.87	89.86	44.39	461.5	20.33

Table 2: Effect of the retrieval embedding source on the RAG retrieved-split ($k=4$) pipeline, averaged over three runs with three different seeds. The Title row is the RAG retrieved-split ($k=4$) system from Table 1. Best results are in **bold**.

B Different k Values

k	SMART F1 ↑	FKGL ↓	SARI ↑				Length	
			add	keep	del	all	Tok.	Sent.
$k=2$	32.90	10.44	7.23	36.60	90.98	44.94	404.8	18.06
$k=3$	32.55	10.31	7.35	36.69	90.51	44.85	437.4	19.50
$k=4$	33.21	10.29	7.83	37.45	90.24	45.17	448.9	20.26
$k=5$	32.97	10.35	7.72	37.05	89.82	44.86	447.3	20.20

Table 3: Effect of the number of retrieved section pairs k on the RAG retrieved-split pipeline, averaged over three runs with three different seeds. Best results are in **bold**.

C Library Versions

Package	Version
torch	2.9.1
vLLM	0.14.0
transformers	4.57.6
easse	git@6a4352e
sacrebleu	2.5.1
nlTK	3.9.2
Levenshtein	0.27.3
scikit-learn	1.7.2

Table 4: Versions of the packages that affect the results.

D Artifact Licenses

Artifact	License
Cochrane-sections	GPL-3.0
Qwen2.5-32B-Instruct	apache-2.0
Gemma-3-27b-it	gemma
Llama-3.1-70B-Instruct	llama3.1
EASSE	GPL-3.0
MedCPT	public-domain

Table 5: Artifact licenses. All artifacts are used consistent with their intended use.

E Per-Section SARI

Section	LLM	RAG retr.-split	Gain
Background	38.95	42.29	+3.34
Objectives	37.49	38.99	+1.50
Search methods	31.62	38.03	+6.41
Selection criteria	27.87	38.88	+11.01
Data coll. & analysis	32.14	40.77	+8.63
Main results	43.18	43.09	−0.09
Authors’ conclusions	37.24	38.92	+1.68

Table 6: Per-section SARI for the LLM baseline and RAG retrieved-split ($k=4$), averaged over three runs. The gains are largest for the sections the LLM baseline handles worst without retrieved examples. Best results are in **bold**.

F Generalization to CLEF 2026 SimpleText

System	SARI ↑
LLM	46.70
RAG retrieved-split ($k=4$)	47.98

Table 7: SARI on the CLEF 2026 SimpleText simple_full test set. Best results are in **bold**.

Section	LLM	RAG retr.-split
Background	43.46	43.96
Objectives	37.36	40.64
Search methods	5.17	33.53
Selection criteria	23.80	39.08
Data coll. & analysis	32.85	33.71
Main results	43.78	43.79
Authors’ conclusions	43.82	43.63
Mean	32.89	39.76

Table 8: Per-section SARI on the CLEF 2026 SimpleText simple_full test set, for the LLM baseline and RAG retrieved-split ($k=4$). Best results are in **bold**.

G Generalization Across Model Families

System	Qwen-32B	Gemma-27B	Llama-70B
LLM	43.69	42.79	43.33
RAG rand-full ($k=4$)	43.78	43.04	43.47
RAG rand-split ($k=4$)	43.74	42.98	43.59
RAG retrieved-full ($k=4$)	44.26	44.08	44.31
RAG retrieved-split ($k=4$)	45.17	44.10	44.96
Δ (LLM to retrieved-split)	+1.48	+1.31	+1.63

Table 9: Document-level SARI on the COCHRANE-SECTIONS test set for all five systems across three LLM families, averaged over three runs with three different seeds. The split’s contribution over retrieved-full varies by family (+0.91 Qwen, +0.65 Llama, +0.02 Gemma). For Gemma, the gain comes almost entirely from retrieval. Best results are in **bold**.

H Prompts

The prompts can also be found on [GitHub](#).

H.1 Baseline Text Simplification

All RAG pipelines use the <<HEADER>> block:

<<HEADER>>

You are rewriting sentences from a Cochrane biomedical review abstract as they would appear in a
→ Cochrane plain-language summary (PLS). A PLS is written for a general reader with no medical
→ training.

You work one sentence at a time. You receive the section name, the section text for context, and the
→ current sentence. Decide whether the sentence belongs in the PLS and rewrite it, or drop it. Use
→ the section name to judge how much is likely to stay: Search methods, Selection criteria, and Data
→ collection and analysis are almost always dropped; Main results and Authors' conclusions are
→ mostly rewritten; Background and Objectives sit in between.

Plain-language style

- Short, direct sentences.
- Replace clinical jargon with everyday words ("randomised controlled trial" -> "study";
→ "participants" -> "people"; "adverse events" -> "side effects").
- Remove effect sizes, confidence intervals, p-values, statistical tests, and methodology. Report
→ findings qualitatively ("more effective than", "no clear difference").
- Drop procedural detail (database names, search dates, risk-of-bias tools, inclusion criteria
→ wording).
- Use active voice where natural. Refer to the review as "this review" or "we".

When to drop a sentence

If the sentence is purely procedural (search strategy, statistical method, risk-of-bias assessment,
→ inclusion criteria wording, trial registration detail) or reports only numeric results without a
→ clinically meaningful finding, output exactly: [DELETE]

Output

- Output only the rewritten sentence or [DELETE].
- No explanations, no prefixes, no markdown.

H.2 Retrieval-Augmented Text Simplification

The -split variants extend the baseline prompt with the <<EXAMPLE INTRO SPLIT>> block:

<<EXAMPLE INTRO SPLIT>>

Below are {k} examples of how the {section_name} section of a Cochrane PLS is typically written for
→ reviews similar in topic. For each example we show three things: the original abstract section,
→ the PLS sentences that are rewrites of kept abstract sentences, and the PLS sentences that the
→ reviewer added without a counterpart in the abstract. Use them to guide STYLE, PHRASING, LENGTH,
→ the KIND OF CONTENT that belongs in this section, and the KIND OF EXPLANATORY ADDITIONS PLS
→ authors typically introduce. Do NOT copy specific findings, drug names, or numbers from them;
→ those belong to other reviews.

The -full variants use the <<EXAMPLE INTRO FULL>> block instead:

<<EXAMPLE INTRO FULL>>

Below are {k} examples of how the {section_name} section of a Cochrane PLS is typically written for
→ reviews similar in topic. For each example we show two things: the original abstract section and
→ the full plain-language summary of that section as written by the reviewer. Use them to guide
→ STYLE, PHRASING, LENGTH, and the KIND OF CONTENT that belongs in this section. Do NOT copy
→ specific findings, drug names, or numbers from them; those belong to other reviews.

I Qualitative Review Examples

Complex	LLM	RAG retrieved-split ($k=4$)
CD010479 (Nagendran et al., 2014) – Background – Laparoscopic surgical box model training Surgical training has traditionally been one of apprenticeship, where the surgical trainee learns to perform surgery under the supervision of a trained surgeon. This is time consuming, costly, and of variable effectiveness. Training using a box model physical simulator – either a video box or a mirrored box – is an option to supplement standard training. However, the impact of this modality on trainees with no prior laparoscopic experience is unknown.	Traditionally, surgeons learn by practicing under the watchful eye of experienced surgeons. This traditional method can be time-consuming, expensive, and inconsistent in its results. Using a box model simulator, which can be a video box or a mirrored box, is an alternative way to help train surgeons besides traditional methods. We don't know how well these simulators help people who have never done laparoscopic surgery before.	Surgical training has traditionally been done through apprenticeship, where new surgeons learn by doing surgeries under the guidance of experienced surgeons. This approach is costly, time-consuming, and varies in its effectiveness. Training using box models (physical simulators) is an option to supplement standard surgical training. The impact of box model training on surgical trainees with no prior laparoscopic experience is unknown.
CD003825 (Musini et al., 2015) – Objectives – Blood-pressure-lowering efficacy of loop diuretics To determine the dose-related decrease in systolic or diastolic blood pressure, or both, as well as adverse events leading to participant withdrawal and adverse biochemical effects (serum potassium, uric acid, creatinine, glucose and lipids profile) due to loop diuretics versus placebo control in the treatment of people with primary hypertension.	This review looked at how loop diuretics compare to placebos in lowering blood pressure and causing side effects in people with high blood pressure.	We wanted to see if loop diuretics could lower blood pressure more than a dummy pill in people with high blood pressure, and if they caused side effects that made people stop taking the medication or changed their blood test results.
CD009170 (Saarto et al., 2011) – Selection criteria – Blunt versus sharp suture needles Randomised controlled trials (RCTs) of blunt versus sharp suture needles for preventing needle stick injuries among surgical staff measured as glove perforations or self-reported needle stick injuries.	This review looked at studies comparing blunt and sharp suture needles to see which ones help prevent needle stick injuries in surgeons.	We looked at studies comparing blunt suture needles to sharp ones to see if blunt needles help prevent needle stick injuries among surgical staff. Needle stick injuries were measured by checking for holes in gloves or by asking staff if they had an injury.
CD003583 (Rajaratnam et al., 2010) – Main results – Interventions for melasma We included 20 studies with a total of 2125 participants covering 23 different treatments. Statistical pooling of the data was not possible due to the heterogeneity of treatments. Each study involved a different set of interventions. They can be grouped into those including a bleaching agent such as hydroquinone, triple-combination creams (hydroquinone, tretinoin, and fluocinolone acetonide), and combination therapies (hydroquinone cream and glycolic acid peels), as well as less conventional therapies including rucinol, vitamin C iontophoresis, and skin-lightening complexes. . .	This review included 20 studies with 2125 people looking at 23 different treatments. The studies looked at many different treatments. The treatments include bleaching agents like hydroquinone, triple-combination creams, combination therapies, and less common options like rucinol, vitamin C iontophoresis, and skin-lightening complexes. Triple-combination cream worked better at lightening melasma than hydroquinone alone or any of its dual combinations. . .	We included 20 studies with a total of 2125 participants looking at 23 different treatments. The treatments included bleaching agents like hydroquinone, triple-combination creams, combination therapies, and some less common options like rucinol and vitamin C iontophoresis. Triple-combination cream was more effective at lightening melasma than hydroquinone alone or other dual combinations of treatments. . .
CD010266 (Brignardello-Petersen et al., 2015) – Authors' conclusions – Antibiotic prophylaxis for orthognathic surgery For people undergoing orthognathic surgery, long-term antibiotic prophylaxis decreases the risk of SSI compared with short-term antibiotic prophylaxis and there is uncertainty of whether short-term antibiotic prophylaxis decreases SSI risk relative to a single pre-operative dose of prophylactic antibiotics.	Long-term use of preventive antibiotics lowers the chance of surgical site infections after orthognathic surgery more than short-term use does. It's unclear if short-term use is better than a single dose before surgery.	Long-term antibiotics seem to lower the risk of surgical site infections more than short-term antibiotics for people having jaw surgery. We are unsure if short-term antibiotics work better than a single dose of antibiotics before surgery.
CD011364 (Aouad-Maroun et al., 2016) – Background – Ultrasound for paediatric arterial line placement Arterial line cannulation in paediatric patients is traditionally performed by palpation or with Doppler auditory assistance in locating the artery before catheterization. It is not clear whether ultrasound guidance offers benefits over these methods.	Doctors usually find the artery in children by feeling it or using a device that listens for the artery's sound before inserting a tube. We found no clear evidence that using ultrasound to guide arterial line placement in children is better than other methods.	Traditionally, doctors find the right spot to put an arterial line in children by feeling the artery or using a device that makes sounds when it finds the artery. It's not clear if using ultrasound helps more than the traditional methods.
CD010467 (Prabhakar et al., 2016) – Main results – IV vs. inhalational anaesthesia for brain tumour surgery We included 15 RCTs with 1833 participants. We determined that none of the RCTs were of high methodological quality. For our primary outcomes, pooled results from two trials suggest that time to emergence from anaesthesia was longer with isoflurane than with propofol, and time to emergence from anaesthesia was not different with sevoflurane compared with propofol. . .	We looked at 15 studies involving 1833 people. People took longer to wake up after using isoflurane compared to propofol, but there was no clear difference in waking times between sevoflurane and propofol. People had less nausea and vomiting with propofol compared to sevoflurane and isoflurane. . .	We included 15 studies with 1833 participants in this review. The studies included in this review were not of high quality. People took longer to wake up after anesthesia with isoflurane compared to propofol. There was no clear difference in waking up time between sevoflurane and propofol. People had less nausea and vomiting with propofol compared to sevoflurane and isoflurane. . .

Table 10: Seven qualitative examples of LLM baseline vs. RAG retrieved-split outputs.