University of Avignon at the CLEF 2025 SimpleText Track: Guided Medical Abstract Simplification

Notebook for the Simpletext Lab at CLEF 2025

Ygor Gallina[1], Tania Jiménez[1] and Stéphane Huet[1]

[1]Avignon Université, LIA, France

Abstract

This paper presents the participation of the SimpLIA team to the CLEF2025 SimpleText challenge on Document-level Scientific Text Simplification. The goal of the task is to simplify a scientific abstract in the biomedical domain for lay readers. To achieve this, we tried different approaches such as translation, keyword extraction and summarization. Our best performing approaches used LLM prompting guided by human-defined guidelines. The various approaches employed in this study rely on readily available tools and models.

Keywords

text simplification, biomedical, scientific abstracts, prompting, large language models (LLM)

1. Introduction

Access to high-quality healthcare information, in a language that people can understand, is a major societal challenge. Lack of clarity around their health information can hinder patients' ability to take an active role in their care. Without a solid understanding of their diagnosis, treatment options, and ongoing needs, patients may struggle to follow treatment plans, prioritize self-care, and make informed decisions about their health [1]. This issue is particularly acute for vulnerable groups such as children, non-native speakers, and those with lower educational backgrounds, exacerbating health disparities. In order to reduce the persistent gap in health literacy between medical professionals and the general public, automatic text simplification approaches can be leveraged to present complex health information in a clear and concise manner, ensuring that essential details are retained while maintaining the accuracy of the original content.

The CLEF 2025 SimpleText lab [2] is dedicated to enhancing accessibility to scientific texts for all users. Unlike conventional text, scientific documents have dense specialized vocabulary and complex sentence structures; successfully adapting these texts to a non-specialized audience while maintaining accuracy is still a challenge.

SimpleText2025 consider three tasks:

1. Text Simplification: Simplify scientific text [3]
   - 1.1 Sentence-level Scientific Text Simplification
   - 1.2 Document-level Scientific Text Simplification
2. Controlled Creativity: Identify and Avoid Hallucination [4]
3. SimpleText 2024 Revisited: Selected tasks by popular request

In order to simplify a text, one must understand the whole document as well as its surrounding context. We have decided to only focus on the Subtask 1.2 [5] at the document level which is more challenging than the sentence level. In a nutshell, our approaches focus on prompting Large Language Models (LLMs) because of their capacity of dealing with large context and thus whole documents.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

ygor.gallina@univ-avignon.fr (Y. Gallina); tania.jimenez@univ-avignon.fr (T. Jiménez); stephane.huet@univ-avignon.fr (S. Huet)

Table 1
Datasets Statistics. Length ratio is the mean ratio between the word length of complex document and simple documents. † identifies data sources part of the SimpleText2025 test set.

dataset subset	#doc.	#word complex	#word simple	length ratio
cochrane-auto				
train	849	339	196	1.84
val†	119	350	203	1.83
test†	117	323	187	1.86
Cochrane†	217	535	n/a	
Medline†	110	207	n/a	
SimpleText2024†	103	145	n/a	

2. Data description

Complex	Two randomised clinical trials were eligible for inclusion. One trial compared biliary lavage with hydrocortisone versus saline in 17 patients. Hydrocortisone tended to increase adverse events (pancreatitis, cholangitis with septicaemia, paranoid ideas, fluid retention) (RR 3.43, 95% CI 0.51 to 22.9) and had no cholangiographic improvement, which led to termination of the trial. The other trial compared budesonide versus prednisone in 18 patients. Patients had statistically significant higher serum bilirubin concentration after treatment with prednisone compared with budesonide (MD 10.4 µmol/litre, 95% CI 1.16 to 19.64 µmol/litre). No other statistically significant effects on clinical or biochemical outcomes were reported on any of the evaluated interventions. There is no evidence to support or refute peroral glucocorticosteroids for patients with primary sclerosing cholangitis. The intrabiliary application of corticosteroids via nasobiliary tube seems to induce severe adverse effects.
Simple	One trial compared biliary lavage with hydrocortisone versus saline. The other trial compared oral administration of budesonide versus prednisone. No statistically significant effects were found on mortality, serum activity of alkaline phosphatases, serum bilirubin, and adverse events for any of the evaluated intervention regimens. Two trials on glucocorticosteroids for primary sclerosing cholangitis were identified.

Figure 1: Sample of the cochrane-auto validation set (document CD004036).

The SimpleText2025 test data gathers a variety of document sources. Table 1 describes the number of documents, number of words and length ratio (the length of complex documents over simple documents). The majority of documents comes from the cochrane-auto[1] [6] dataset, which we used as our development dataset before knowing it was part of the test set. It was constructed from systematic reviews of biomedical scientific literature, produced by Cochrane[2] a not-for-profit organization.

A sample is composed of a complex document (the input) and a simple version (the reference). The complex version is the "Main results" part of the Cochrane Review abstract and the simple version is a part of the "Plain language summary" starting from the "What did we find?" paragraph to the end, excluding titles. Figure 1 showcases a document from cochrane-auto[3].

By looking at the validation samples, we found that the simplification relies on the reformulation of scientific wording, rather than the contextualization of the study and the explanation of specialty terms. In particular, text describing results and confidence intervals are generally removed in the simple version. In contrast, we also noticed that specialized terms tend to be untouched in the simple version, since the Cochrane lay summary targets an audience interested and educated in the health domain.

The SimpleText2025 test set contains 4 sources of documents mostly from the scientific medical domain: Cochrane (literature review), cochrane-auto (literature review) and Medline (scientific articles), and also from the computer science domain: SimpleText2024. The documents from Cochrane, Medline

[1] github.com/JanB100/cochrane-auto
[2] cochranelibrary.com
[3] The original document can be found at pubmed.ncbi.nlm.nih.gov/20091555.

and SimpleText2024 do not match the average length of the cochrane-auto dataset; Cochrane documents are longer by $\simeq 200$ words, Medline and SimpleText2024 documents are shorter by $\simeq 150$ and $\simeq 200$ words respectively.

3. Approaches

3.1. Naive prompt

We started to experiment with the most straightforward prompt we can imagine: "Simplify this text in English." (cf. Listing 4). This prompt was tested using the R library `rollama` [7] with several widely available and established LLMs: LLaMA3.3 (70.6B), LLaMA4 (108.6B), Gemma2 (9B), Gemma3 (4.3B), Mistral-Small (22.2B). It was also tested using LLaMA3.2 (3.2B) and qwen (4B) but the scores were too low to be reported.

3.2. Cascade: Extractive summarization followed by prompting

We also considered a two-step approach: we first reduce the size of the document using extractive summarization following [8] and then ask for simplification using the prompt in Listing 4. We also evaluated the extractive summarization alone, but the results were not satisfactory.

3.3. Backtranslation

We hypothesize that by repeatedly translating the same document to a target language, the documents wording will gradually simplify and is likely to shrink in size. We chose the NLLB [9] model for its wide range of available languages, and Spanish as the target language.

We adopted the following pipeline:

- Translate original (English) document to another language (Spanish).
- Translate the resulting text back to English.
- Repeat these two steps n times.

Because the NLLB models only allow inputs of 512 tokens, the document is split by sentences, which are then grouped up to 512 tokens, translated and concatenated to form the full document.

3.4. Keyword simplification

One way to simplify a text involves explaining or replacing specialty terms that lay readers are not familiar with. Following this idea, we identify keywords (used as surrogates for specialty terms) using the MultipartiteRank [10] algorithm[4]. Then a large language model is prompted (cf. Listing 5) to generate a simpler version of the term. Once the simpler versions of the words are generated, we search and replace the keywords with their simpler version. The simplified document is then very similar to the original.

To prevent the LLM from deviating from the expected YAML format, the initial prompt was modified to instruct the generation of a definition of the term before its simplification, which acts like a chain-of-thought.

3.5. Cochrane guidance

Simplification, as well as summarization, is not performed in the same way depending on the goal and the intended audience. In the context of Cochrane's Lay Summaries, the objective is defined in guidebooks destined to help professionals write a Cochrane Review. The "Lay Summary" aims at ensuring that "anyone looking for information about the key points of a Cochrane Review can read and

[4] github.com/boudinfl/pke

understand them." [11]. As such, the "Template and guidance for writing a Cochrane Plain language summary"[5] details how a Lay Summary should be written, and what are the aspects to focus on.

We identified different parts of the handbook to help guide the simplification process.

Guidelines The 5th section of the handbook, namely "General advice on writing in plain language", describes redaction advice regarding language, style and structure. Because the "simple" version of the dataset does not contain headings, bullet lists nor paragraphs, we only considered the language and style advice. We created a prompt that includes all the extracted advice (Guidelines-all), only the language one (Guidelines-lang) and lastly, just the style one (Guidelines-styl) (cf. Listing 2).

The handbook's advice section is formatted as multilevel bullet points list, but we manually rewrote them as sentences to create clearer guidelines (cf. Appendix B). For example:

- Avoid (or, when this is not possible or desirable, explain):
  - long words. For example, use 'blood thinners' as an alternative to 'anticoagulants'.
  - research jargon. Use:
    - 'study' rather than 'trial';
    - 'people with [condition]', 'women', 'children' etc. rather than 'participants';
    - [...]

was manually rewritten to:

- Avoid (or explain)[6] long words. For example, use 'blood thinners' as an alternative to 'anticoagulants'.
- Avoid (or explain) research jargon. For example, use 'study' rather than 'trial'; 'people with [condition]', 'women', 'children' etc. rather than 'participants'; [...]

Fewshot The handbook's Appendices 3 and 4 contain two curated examples of lay summaries to help writers in their work. We hypothesize that these chosen examples can help a language model reproduce the style of human written lay summaries. To match the cochrane-auto dataset and to act as few-shot examples, the two summaries were copied from the "What did we find?" paragraph up to "How up to date is this review?" (excluded) and paragraph names were removed.

In order to validate the curation choice of these examples, two other documents were randomly chosen from the validation set of cochrane-auto[7] (Fewshot-rand) to act as few-shot examples. Fewshot-coch is expected to perform better than Fewshot-rand given the samples' curation. The prompt used is defined in Listing 3.

Prompting lay In order to understand how the guidance affects the lay summaries, we created a simple prompt (distinct from the "Naive prompt" approach described in Section 3.1) that serves as a base for "Guidelines" and "Fewshot" prompts (cf. Listing 1).

4. Results

Results on the validation set of cochrane-auto are reported in Tables 2 and 3. Scores are computed using only 1 run. During development the approaches were evaluated with BLEU [12], SARI [13], BertScore [14] and LENS [15]. Although the "Naive prompt" approach was also assessed through LLaMA4 (109B), these results were not reported in Table 2 due to space constraints; overall, its performances remain lower than LLaMA3.3 70B, with SARI, LENS and length ratio of 33.9, 71.1 and, 0.7 respectively.

All approaches (except "Keyword") surpass "Baseline input"(where the input is left untouched, cf. Table 3) in terms of LENS (41.7), but none in terms of SARI (46.0).

[5] training.cochrane.org/handbook/current/chapter-iii-s2-supplementary-material

[6] Some parts were removed such as ", when this is not possible or desirable,".

[7] The documents ids are CD013028 and CD013733.

Table 2
Performance of LLM-based approaches computed on the validation set of the cochrane-auto dataset. lr stands for length ratio (closer to 1 is better). For each LLM, the best approach is emphasized in **bold**. The submitted runs are highlighted in blue.

	Gemma3 4B			Gemma2 9B			Mistral-Small 24B			LLaMA3.3 70B		
Approach	SARI	LENS	lr	SARI	LENS	lr	SARI	LENS	lr	SARI	LENS	lr
Naive prompt	35.4	60.7	0.9	32.8	67.8	0.5	37.4	56.9	1.0	36.4	74.2	0.8
Cascade	33.4	**68.4**	0.5	32.2	60.5	0.4	34.4	50.9	0.5	32.4	65.0	0.4
Keyword	35.8	21.3	2.0	35.7	13.6	2.3	35.7	20.6	2.0		ø	
Prompting lay	39.7	48.8	2.1	35.8	61.5	1.4	39.1	69.6	1.4	**37.5**	74.2	1.5
Fewshot - coch	38.9	52.0	2.0	35.4	68.6	1.2	39.6	68.9	1.4		ø	
rand	39.7	55.0	1.9	35.3	70.3	1.1	39.2	69.9	1.3		ø	
Guidelines - all	38.6	53.0	2.0	35.7	74.3	1.3	39.7	75.7	1.5		ø	
lang	36.6	59.1	1.9	35.6	73.5	1.3	39.1	**77.3**	1.5		ø	
styl	**39.8**	51.6	1.9	**37.3**	**77.1**	1.1	**40.0**	71.4	1.3	36.7	**79.4**	1.2

Table 3
Performance for backtranslation computed on the validation set of the cochrane-auto dataset. lr stands for length ratio (closer to 1 is better). The best scores are in **bold**, the submitted run is colored in blue.

Approach	SARI	LENS	lr
Baseline input	46.0	41.7	1.8
Backtrans.			
n2	**33.3**	33.8	0.5
n4	32.6	35.9	0.5
n6	32.5	36.4	0.5
n8	32.4	**36.8**	0.5
n10	32.4	36.7	0.5

Comments on LLMs With regard to SARI score, Mistral-Small almost always achieves the higher scores. With respect to LENS, its superiority is not as striking, but it sill obtains the best score most of the time.

LLaMA3.3 with 70B parameters gets the best LENS score in 3 out of 4 runs, but its SARI score never surpasses Mistral-Small. Its execution is very slow due to the large number of parameters, therefore we did not run this model in all settings. Gemma2 with 9B parameters obtains a lower SARI overall, but is comparable to Mistral-Small with 24B parameters on the LENS scores.

Comments on approaches In general, including guidelines inside the prompt, in particular following the "Guidelines-styl" approach, is benefiting to LLMs. Indeed, the "Prompting Lay" is always improved with adding guidance (except for "Fewshot-coch" using Mistral-Small, cf. Table 2) and the highest scores are observed with "Guidelines-styl". However, we observed one exception to this rule, Gemma3 having higher LENS scores using the "Cascade" and "Naive prompt" approaches.

From the "Fewshot" experiments (cf. Table 2), it seems that the examples chosen by Cochrane are less effective than random ones.

Document Length Overall the "Guidelines" approaches produced much longer documents (≃360 words) than the reference (≃195 words). This can be explained by the fixed limit of 500 words stated in the prompt (cf. Appendix 2), which was computed using the longest document of the validation set; this criterion implicitly encourages LLMs to produce lengths close to this limit.

Another set of experiments led with the "Guidelines" approaches revisited the word limit with a new value of 850. The results were consistently better with the original value of 500. This implies that

Table 4
Submitted runs scores

RunId	SARI	Approach	
301392	**44.9**	Guidelines-all	(Mistral-Small)
301393	44.4	Guidelines-lang	(Mistral-Small)
302455	43.6	Guidelines-styl	(Gemma2)
303086	43.2	Guidelines-styl	(LLaMA3.3)
302653	42.3	Prompting (Naive prompt)	(LLaMA3.3)
302680	41.1	Prompting (Naive prompt)	(LLaMA4)
302682	38.4	Cascade	(LLaMA3.3)
302458	36.6	Backtranslation	(NLLB)

generating smaller documents would result in higher performances and we could for example set the word limit dynamically according to the length ratio and the length of the input document.

Run submission The LENS score was used as the primary indicator of performance because it takes into consideration both the input and the reference, while capturing semantics. SARI for its part, only relies on edit distance.

We chose to send runs that obtained a LENS score greater than $\simeq 70$. We thus submitted eight runs and a baseline. Table 4 summarizes our submissions and reports the scores on the SimpleText25 test set.

Due to the long inference time of LLaMA3.3 linked to its large numbers of parameters we were not able to submit all approaches on time for the track. Similarly, the "Guidelines-styl" run with Mistral-Small, while it achieves the best SARI score overall, was not submitted.

By comparing the evaluations performed on the validation set and the SimpleText25 test set, we observed the same ranks of our systems w.r.t. SARI. Our prior findings regarding the benefits of incorporating guidelines, notably with Mistral-Small, are confirmed with the experiments done on these new data.

5. Manual analysis

In an effort to gain more insight on the produced simplified documents, 79 predictions coming from 4 randomly chosen documents[8] were analyzed. To minimize bias, the documents were annotated solely relying on the document ID, the original and the simplified documents, while the approach used to produce the document was not shown. The analytical process involved an initial labeling stage where open-label categories were applied to 20 documents, followed by a second phase where label revisions or consolidations were made to better align with the underlying themes and phenomena. After adjusting the label set, more documents were annotated.

Low quality predictions All analyzed documents from the "Backtranslation" approach were too short and only simplified the start of the document. In the same way, documents from the qwen LLM were undersized and the end of the simplification seemed truncated. On the contrary, simplification from the "Keyword" approach were too long.

Factuality 7 out of the 79 analyzed simplification contained fake information, the majority of these spurious data being mixed facts (In the original document A does B and C does D, but the prediction states that A does D.), and one occurrence of wrong information (with "Guidelines-all", Gemma3 generated "*A vena cava filter is a small, portable electronic device that counts the number of steps you take.*", whereas it is actually a filter catching blood clots).

[8]Examples were taken from the cochrane-auto validation set: CD006212, CD009242, CD011768, CD013792.

Interestingly, 3 predictions contained correctly inferred data from the results described in the complex document. For example, the document CD009242 compares the age at which children can walk independently when using treadmills or not and reported the study results as "(MD *-4.00*, 95% CI -6.96 to -1.04)", which was correctly simplified as "4 months earlier" by the LLM. This shows that the language model was able to understand the information and rephrase it accordingly.

In 15 descriptions, contextual details were incorporated on the experiments, mainly a description of the studied condition and an explanation of the treatments. For example, the original document CD013792 about treating miscarriages does not give context about the compared medicines (different kinds of progesterone), but using "Guidelines-lang", Mistral-Small was able to add context to understand why these medicines are used: "*These medicines are types of progestogens, which are hormones that help maintain pregnancy.*"

Writing Style Some conclusions directly address the reader to use the information with caution and to fact-check such as "*If you are a parent or caregiver, you may want to talk to your child's doctor or therapist about [treatment].*" , "*If you're pregnant and at risk of miscarriage, talk to your doctor about your options.*" Of the 13 predictions containing these conclusions destined to users, 50% of them come from the "Guidelines-styl" approach.

Document structure Every LLM organized its response (in markdown syntax) in at least one of its predictions depending on the approach. The output structure ranges from one bullet point list to multiple sections with titles, and bold or italic text. More precisely, 90% of generated text coming from Gemma3 was structured, 50% from Gemma2, and ≃30% from LLaMA32, LLaMA33, and Mistral-Small.

6. Conclusion

This paper presents the University of Avignon's participation in the SimpleText CLEF 2025 evaluation on Tasks 1.2, using LLMs of various sizes. We tried several approaches, experimenting with backtranslation using a Machine Translation system and with prompting to LLMs to guide them toward the task. The best results were obtained by giving instructions drawn from the Cochrane Guidelines 3.5 on how to simplify documents. Future works should investigate closely the size of the produced summaries and further analyses the outputs to gain insights on explaining the results.

Acknowledgments

This research was funded, in part, by Bpifrance under the PARTAGES project. We thank Avignon University's IUT for allowing us to use their local ollama instance. This work was performed using HPC resources from GENCI-IDRIS (Grant 2025-A0181016171).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. King, Poor health literacy: a 'hidden' risk factor, Nature Reviews Cardiology 7 (2010) 473–474.

[2] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText track: Simplify scientific texts (and nothing more), in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of

the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.
[3] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, in: [16], 2025.
[4] B. Vendeville, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 2: Identify and Avoid Hallucination, in: [16], 2025.
[5] J. Bakker, L. Ermakova, Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025) (2025).
[6] J. Bakker, J. Kamps, Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.
[7] J. B. Gruber, M. Weber, rollama: An R package for using generative large language models through Ollama, 2024. URL: https://arxiv.org/abs/2404.07654v1.
[8] D. Majumdar, Reinforcement Learning in NLP, 2025. URL: tutorials/reinfnlp/reinfnlp.html.
[9] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022. URL: https://arxiv.org/abs/2207.04672. arXiv:2207.04672.
[10] F. Boudin, Unsupervised Keyphrase Extraction with Multipartite Graphs, in: Proceedings of NAACL-HLT 2018, Association for Computational Linguistics, 2018. URL: http://arxiv.org/abs/1803.08721.
[11] N. Pitcher, D. Mitchell, C. Hughes, Template and guidance for writing a Cochrane Plain language summary (2022).
[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2001, p. 311. URL: http://portal.acm.org/citation.cfm?doid=1073083.1073135. doi:10.3115/1073083.1073135.
[13] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://direct.mit.edu/tacl/article/43364. doi:10.1162/tacl_a_00107.
[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTSCORE: EVALUATING TEXT GENERATION WITH (2020).
[15] M. Maddela, Y. Dou, D. Heineman, W. Xu, LENS: A Learnable Evaluation Metric for Text Simplification, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16383–16408. URL: https://aclanthology.org/2023.acl-long.905/. doi:10.18653/v1/2023.acl-long.905.
[16] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

A. Used Prompts

```
Given the following scientific abstract create a lay summary (500
    words maximum).

Abstract:
        {{abstract}}
```

Listing 1: Prompt used for the Prompting lay approach.

```
Given the following scientific abstract create a lay summary (500
    words maximum) using the following guidelines.
{% for g in guideline %}
- {{ g }}
{% endfor %}

Abstract:
        {{ abstract }}
```

Listing 2: Prompt used for the Guideline approach.

```
Given the following scientific abstract create a lay summary (500
    words maximum). Here are some examples of lay summaries.
{% for e in examples %}
Example:
{{e}}

{% endfor %}

Abstract:
        {{abstract}}
```

Listing 3: Prompt used for the Few-shot approach.

```
queryLLama33 <- make_query(
  text = cochraneauto_docs_val$complex,
  prompt = "Simplify this text in english",
  system = "Please do NOT use a first introductory sentence, only
     the simplified text'",
  prefix = "Here is the text : "
)
resp_llama33_val=query(queryLLama33, screen = FALSE, output = "text"
    )
```

Listing 4: Prompt used for the Naive prompt approach.

```
I'm a highschool student interested in healthcare. I want to understand some difficult
    words in a scientific abstract.
After the --- is an abstract and a list of {{ n }} words I find difficult to understand.
Define the words and give me a simpler version that I can understand (it's okay if some
    meaning is lost). The simpler version should be a direct replacement for the words
    in the abstract.
The answer should be formatted as a yaml list of objects with 3 attributes: the word,
    the definition and the simpler version of the word.
```

```
The answer should only contain yaml no other text or comment. The output must be
    readable by a yaml parser.
- word: "WORD 1"
  definition: "definition of WORD 1"
  simple: "simpler WORD 1"
- word: "WORD 2"
  definition: "definition of WORD 2"
  simple: "simpler WORD 2"
- ...


---
**Abstract** : {{ abstract }}
**Words to explain** :
{% for t in terms %}
- {{ t }}
{% endfor %}
```

Listing 5: Prompt used for the Keyword Simplification approach.

```
- word: "AMD"
  definition: "Age-related macular degeneration - a condition that affects the central
      part of the retina and can cause vision loss."
  simple: "macular degeneration"
- word: "Evidence"
  definition: "Scientific proof or data that supports a claim or idea."
  simple: "proof"
- word: "Fatty Acid Supplements"
  definition: "Supplements containing substances like omega-3 that are made of fats."
  simple: "fish oil"
- word: "Low Risk"
  definition: "A situation where something is unlikely to cause harm or problems."
  simple: "safe"
- word: "Omega"
  definition: "A chemical element (Omega) but in this context, it refers to a specific
      type of fatty acid."
  simple: "fish oil"
- word: "People"
  definition: "Individuals or participants in a study."
  simple: "patients"
- word: "Placebo"
  definition: "An inactive substance or treatment given to a control group in a study,
      used to see if the treatment itself has an effect."
  simple: "dummy pill"
- word: "Progression"
  definition: "The process of something developing or worsening over time."
  simple: "worsening"
- word: "Trials"
  definition: "Research studies, particularly clinical trials."
  simple: "studies"
- word: "USA"
  definition: "United States of America"
  simple: "America"
```

Listing 6: Simplified keywords from Gemma2 for document CD010015.

B. Cochrane Writing Advice

This section compiles the tips available in Cochrane's "Template and guidance for writing a Cochrane Plain language summary". The tips were manually edited to form full sentences.

Language advice

- Use everyday language. For example, refer to 'people' instead of 'study participants'.
- Avoid (or explain) long words. For example, use 'blood thinners' as an alternative to 'anticoagulants'.
- Avoid (or explain) research jargon. For example, use 'study' rather than 'trial'; 'people with [condition]', 'women', 'children' etc. rather than 'participants'; the name of the intervention instead of 'intervention'; the name of the control or comparison instead of 'control' or 'comparison'; the name of the outcome instead of 'outcome'.
- Avoid (or explain) words or phrases with dual or nuanced meanings. For example, use 'medicines' instead of 'drugs'. 'Significant' means 'important' for a lay reader.
- Explain 'common' medical words. For example: 'acute condition': a condition or state that develops suddenly and lasts a short time; 'chronic condition': a condition or state that lasts for a long time.
- Explain technical medical terms. Plain language does not always mean 'lay language'. Your reader may know the topic via the technical term – especially if they are a patient or carer, so it might be best to include the technical term and explain it. For example, to explain the action of anticoagulants, you could write: 'Anticoagulants are medicines that stop harmful blood clots forming. However, these medicines may cause unwanted effects such as bleeding.' Or you could write the term in plain language followed by the technical term in brackets. For example, 'blood thinners (anticoagulants)'.
- Avoid acronyms and abbreviations. If you cannot avoid them, make sure you define them when you first mention them. For example, 'nicotine replacement therapy (NRT)'. Use phrases like 'for example', 'such as', 'in other words', 'and so on' instead of 'e.g.', 'i.e.' or 'etc.', as they are not always understood if you are writing for a wide audience.
- Write for an international audience. Avoid regional words or terms; for example, use 'hospital emergency care' instead of 'Accident & Emergency (A&E)' (UK) or 'Emergency Room (ER)' (USA).

Style advice

- Keep paragraphs and sentences short, but vary your sentence length occasionally to keep the readers' attention. Aim for an average of 20 words in a sentence. Break up longer sentences into shorter ones. For example, instead of 'Most people who smoke want to stop, however many find it difficult to do so, even though they may use medicines that are designed to help them stop', you could write 'Most people who smoke want to stop, but many find it difficult. People who smoke may use medicines to help them stop.'.
- Use the active voice. For example, write 'We compared and summarized the results of the studies' instead of 'The results of the studies were compared and summarized'.
- Use pronouns. Write in the first-person plural. For example, use 'we assessed' instead of 'the review authors assessed'. Address your reader using the second-person pronoun 'you'. For example, write 'A pedometer is a small, portable electronic device that counts the number of steps you take.'.
- Use verbs. For example, say 'the students investigated' not 'the students conducted an investigation', or 'we analyzed the data' not 'we carried out an analysis of the data'.
- Write numbers as numerals (1, 2, 3…) rather than words. However, avoid starting a sentence with a numeral. If necessary, rewrite the sentence. For example, write 'The studies included 3260 people' instead of 'Three-thousand, two-hundred and sixty people took part in the studies'.

- Be concise. A Plain language summary can be up to 850 words long, but you do not have to fill the word limit. You should aim to keep it as short as possible while still including the most important information.
- Replace ‘wordy’ phrases with shorter alternatives: use ‘during’ instead of ‘during the course of’; use ‘often’, instead of ‘it was often the case that’; use ‘some’ or ‘many’, instead of ‘a number of’; and o use ‘because’ instead of ‘due to the fact that’.

C. Hyperparameters

The LLMs were used with ollama’s default parameters or the model’s default if provided. Unless stated otherwise, for all LLMs used in this work the temperature, topn and topp are set, respectively, to 0.8, 40 and 0.90. For Gemma3, the same parameters are set to 1.0, 64 and 0.95.

Guidelines - all, Mistral-Small, SARI 34.8, LENS 47.7, length ratio 4.9

We reviewed two studies that looked at different treatments for people with a condition called primary sclerosing cholangitis (PSC). PSC is a chronic condition where the bile ducts in the liver become inflamed and scarred, making it hard for bile to drain properly. This can lead to serious health problems.

The first study compared two treatments: hydrocortisone and saline. Hydrocortisone is a type of steroid medicine that reduces inflammation. Saline is a saltwater solution often used as a comparison treatment in studies. The study involved 17 people with PSC. Unfortunately, the people who received hydrocortisone had more side effects than those who got saline. These side effects included pancreatitis (inflammation of the pancreas), cholangitis with septicaemia (infection and inflammation of the bile ducts that can spread to the bloodstream), paranoid ideas, and fluid retention. The study was stopped early because hydrocortisone did not improve the condition of the bile ducts and caused these serious side effects.

The second study compared two different steroid medicines: budesonide and prednisone. Both are used to reduce inflammation but work in slightly different ways. This study included 18 people with PSC. After treatment, those who received prednisone had higher levels of bilirubin in their blood than those who got budesonide. Bilirubin is a waste product made by the liver that can build up if the bile ducts are blocked. Higher levels of bilirubin can indicate that the liver is not working properly.

However, neither study found any other significant effects on how people felt or on other measures of their health. This means we still do not have clear evidence to support or refute the use of steroid medicines taken by mouth for people with PSC. The studies also suggest that putting corticosteroids directly into the bile ducts through a tube in the nose (intrabiliary application) may cause severe side effects.

In summary, more research is needed to find effective and safe treatments for people with primary sclerosing cholangitis.

Guidelines - lang, Mistral-Small, SARI 37.3, LENS 71.9, length ratio 4.7

Two studies looked at different treatments for people with a condition called primary sclerosing cholangitis (PSC). This is a chronic liver disease where the bile ducts become inflamed and scarred, affecting how bile flows from the liver. The first study compared two treatments: hydrocortisone (a type of steroid) mixed with saline solution versus just saline solution alone. The treatment was given through a process called biliary lavage, which involves flushing out the bile ducts.

The study included 17 people and found that those who received hydrocortisone had more side effects such as pancreatitis (inflammation of the pancreas), cholangitis with septicaemia (infection in the bile ducts spreading to the bloodstream), paranoid ideas, and fluid retention. These side effects were serious enough to stop the study early. There was no improvement in the condition of the bile ducts seen through imaging tests.

The second study compared two different steroids: budesonide versus prednisone. This study included 18 people. The results showed that people who took prednisone had higher levels of bilirubin (a waste product made by the liver) in their blood after treatment compared to those who took budesonide. Bilirubin is a substance produced by the body and processed by the liver, so high levels can indicate problems with liver function.

However, there were no other significant differences in how well people felt or in other measures of health between the two groups. This means that neither steroid showed clear benefits over the other for treating PSC. Overall, these studies do not provide strong evidence to support or reject the use of steroids taken by mouth (peroral glucocorticosteroids) for people with primary sclerosing cholangitis. The study also suggests that using corticosteroids directly in the bile ducts through a nasobiliary tube can cause severe side effects.

In summary, more research is needed to find effective and safe treatments for people with PSC.

Guidelines - styl, Mistral-Small, SARI 37.2, LENS 75.1, length ratio 3.1

We reviewed two clinical trials to understand the effects of different treatments on people with a condition called primary sclerosing cholangitis (PSC). This condition affects the bile ducts in the liver.

The first trial compared hydrocortisone and saline. We found that hydrocortisone increased adverse events, such as pancreatitis, infections, paranoid ideas, and fluid retention. The study had to stop early because of these severe side effects. Hydrocortisone did not improve bile duct imaging results either.

The second trial compared budesonide and prednisone. We found that people treated with prednisone had higher levels of bilirubin in their blood. Bilirubin is a waste product made by the liver. Higher levels can indicate liver problems. However, we did not find any other significant differences between these two treatments.

Overall, our review does not provide clear evidence to support or refute the use of oral glucocorticosteroids for people with PSC. We also found that applying corticosteroids directly into the bile ducts via a nasobiliary tube can cause severe side effects.

You should talk to your healthcare provider about the best treatment options for you if you have PSC. More research is needed to find effective and safe treatments for this condition.

Figure 2: Samples of the document CD004036 comparing the use of different sets of guidelines.