

# LLM-Guided Planning and Summary-Based Scientific Text Simplification: DS@GT at CLEF 2025 SimpleText

Krishna Chaitanya Marturi<sup>1,\*</sup>, Heba H. Elwazzan<sup>2</sup>

<sup>1</sup>Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

## Abstract

In this paper, we present our approach for the CLEF 2025 SimpleText Task 1, which addresses both sentence-level and document-level scientific text simplification. For sentence-level simplification, our methodology employs large language models (LLMs) to first generate a structured plan, followed by plan-driven simplification of individual sentences. At the document level, we leverage LLMs to produce concise summaries and subsequently guide the simplification process using these summaries. This two-stage, LLM-based framework enables more coherent and contextually faithful simplifications of scientific text.

## Keywords

LLMs, Text Simplification, CLEF 2025, CEUR-WS

## 1. Introduction

In the past decade or so, a new form of learning has emerged. Instead of science being only accessible through journals or formal education, the general public is now privy to an enormous wealth of material through the Internet. Whether it be directed self-studying, or casual social media perusal, nearly everyone is now able to access scientific information. An important caveat remains, however, and that is when a resource is accessible for free and without rigorous moderation, misinformation will invariably run rampant. This is why now more than ever, the need for reliable, easy-to-understand scientific-based text and content has taken the forefront.

This is where automated text simplification comes in. With the volume of scientific text at hand, rewriting the same exact content in layman terms manually is intractable. Resources have been expended towards automating this task, and over the course of 20 years, automatic text simplification has progressed significantly [1], reaching a critical point with the development of Natural Language processing techniques, and more recently, with the widespread use of LLMs.

The capabilities of Large Language Models have made them a game-changer for automatic text simplification. Unlike previous methods, LLMs can achieve a deeper semantic interpretation of source text, allowing them to not only simplify vocabulary and syntax but also to summarize and restructure information for clarity. Though their internal representations of knowledge are opaque, their practical application as a powerful tool for generating simplified text is undeniable, paving the way for more effective simplification systems.

The SimpleText lab [2] is part of CLEF, and it aims to address the task of text simplification of scientific text. Task 1 in particular [3] involves investigating the performance of text simplification on both the sentence-level and document-level. It uses the Cochrane-Auto dataset [4] and uses standard text simplification metrics such as SARI, BLEU, BERTscore, etc. to evaluate the generated simplified text against the reference ones.

In this paper, we tackle both sentence-level text simplification as well as document-level using a tiered approach. For the sentence-level, an LLM first generates a simplification strategy regarding a sentence, and then the LLM is tasked to perform that strategy to generate a simplified version. For the document-level, the LLM generates a summary of the text as a whole and then uses that summary as part of the prompt that guides the simplification process.

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

\*Corresponding author.

✉ kmarturi3@gatech.edu (K. C. Marturi); helwazzan3@gatech.edu (H. H. Elwazzan)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The paper is organized as follows: Section 2 provides a small literature review of the text simplification task; section 3 provides details of the approach undertaken for each of the subtasks; section 4 showcases the results and provides a brief discussion; section 5 discusses possible future work and the conclusions we have derived from these experiments.

## 2. Related Work

Scientific text simplification aims to enhance the accessibility and comprehensibility of technical content for non-expert audiences, including patients, educators, and the general public. This task has been explored at both sentence and document levels, with increasing interest in using neural and large language model (LLM)-based methods.

Ondov et al. [5] provide a comprehensive survey of automated methods for biomedical text simplification. They categorize approaches into rule-based, statistical, and neural systems, highlighting the trade-offs between linguistic control and generative fluency. Their analysis underscores challenges in maintaining factual consistency and domain-specific accuracy, particularly in biomedical domains. This motivates the need for more grounded, interpretable approaches, such as plan-driven or summary-guided simplification, which we explore in this paper.

Recent work has also examined the role of LLMs in improving user comprehension and reducing cognitive load. Guidroz et al. [6] evaluate how LLM-based simplification affects the understanding of the readers and the mental effort of different audiences. They find that while LLM-generated simplifications generally improve readability, there is a risk of hallucination and over-simplification, especially in scientific and biomedical texts.

Fang et al. [7] propose a hierarchical strategy involving LLMs to simplify document level sentences using a progressive process, that breaks it down to discourse-level, topic-level, and lexical-level simplification. The task is formulated as a conditional generation problem by autoregressively conditioning on the input source document. This approach effectively preserves the content of the document while eliminating ambiguity and subjectivity, and avoids treating the document simplification task as merely document summarization.

These findings support our motivation to investigate structured prompting methods to enhance control, coherence, and factuality in the simplification of scientific texts.

### 2.1. Evaluation Metrics

A variety of automatic evaluation metrics are employed to assess the quality of generated or simplified text. In this work, we consider four commonly used metrics: SARI, BLEU, BERTScore (F1) and Flesch-Kincaid grade level (FKGL), each capturing different aspects of text quality to compare LLM-based simplification strategies.

**SARI** (System output Against References and against the Input sentence) [8] is specifically designed for the text simplification task. Unlike traditional metrics, SARI compares the system output not only to reference simplifications but also to the original input. It evaluates the quality of three operations: *keeping* relevant words, *deleting* unnecessary ones, and *adding* appropriate new content. SARI is computed as the average of F1 scores for these three operations and is typically scaled from 0 to 100, with higher scores indicating better simplification quality.

**BLEU** (Bilingual Evaluation Understudy) [9] is an n-gram precision-based metric widely used in machine translation and text generation. It measures the overlap between system outputs and reference texts, incorporating a brevity penalty to discourage overly short outputs. However, BLEU is less suitable for simplification, as it tends to penalize edits that diverge lexically from the reference even when such changes improve simplicity or meaning.

**BERTScore** [10] evaluates the semantic similarity between the generated text and the reference using contextual embeddings from a pretrained transformer model. The F1 variant computes the harmonic mean of precision and recall based on cosine similarity between token embeddings. BERTScore is particularly useful in capturing semantic adequacy, especially when lexical overlap is low but the meaning is preserved.

**Flesch–Kincaid Grade Level (FKGL)** [11] is a readability metric that estimates the U.S. school grade level required to comprehend the text. It is computed using average sentence length and average syllables per word. Lower FKGL scores indicate simpler text and are often used as a proxy for evaluating readability in simplification tasks. However, FKGL focuses solely on surface-level features and does not consider syntactic or semantic correctness [12].

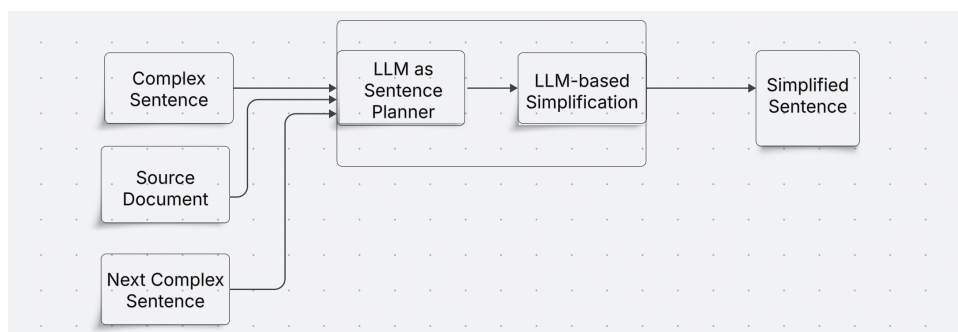
### 3. Methodology

The focus of task 1 is to study the performance of simplification systems in both sentence-level and document-level settings.

#### 3.1. Sentence-level simplification - Task 1.1

Inspired by recent advances in plan-driven sentence simplification [4], we adopt a large language model (llama-3.3-70b-versatile) as a plan-based simplifier. As illustrated in Figure 1, this approach utilizes few-shot prompting with three inputs: a complex sentence, its corresponding source document, and the next complex sentence from the document.

The task is structured into two stages. In the first stage, the model is prompted to select an appropriate simplification strategy from a predefined set: *rephrase*, *delete*, *split*, *ignore*, or *merge*. In the second stage, the model is prompted to generate the corresponding simplified sentence based on the selected strategy [Appendix B.1].

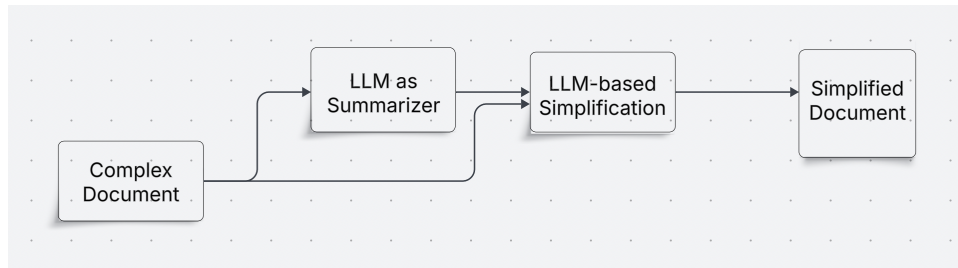


**Figure 1:** LLM based Plan-Driven Sentence Simplification - Task 1.1

#### 3.2. Document-level simplification - Task 1.2

In this task, we leverage large language models (LLMs) for summary-guided document simplification [7]. Figure 2 outlines our two-step pipeline, where a large language model, llama-3.3-70b-versatile, is employed both as a summarizer and a simplifier.

First, the model is prompted to produce a clear and concise summary of the input complex document [Appendix B.2]. This summary serves as a semantic scaffold to guide the simplification process. In the second step, the same model is prompted to simplify the original document using the generated summary as contextual guidance [Appendix B.3]. This strategy enhances coherence and faithfulness while reducing the risk of over-simplification.



**Figure 2:** LLM based Summary-Guided Document Simplification - Task 1.2

## 4. Results

### 4.1. Evaluation of Task 1.1: Sentence-level Scientific Text Simplification

The `plan_guided_llama` system demonstrates effective sentence-level simplification on the 37 aligned Cochrane-auto abstracts, achieving a strong SARI score of 42.33 and reducing lexical complexity (8.52) while maintaining reasonable BLEU and FKGL values. This indicates that the model balances simplification and content preservation. Detailed results are shown in Table 1.

**Table 1**

Results for CLEF 2025 SimpleText Task 1.1 Sentence-Level Text Simplification: Test data on 37 aligned Cochrane-auto abstracts

Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity Score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100.00	100.00	11.73	0.56	0.67	0.50	0.00	0.16	0.60	8.71
<code>plan_guided_llama</code>	37	42.33	10.43	7.77	0.48	0.97	0.47	0.00	0.18	0.70	8.52

The `plan_guided_llama` model demonstrates strong simplification capability on the 217 Plain Language Summaries test set, achieving a SARI score of 42.98, which reflects balanced simplification performance. However, its BLEU score (6.33) is notably low, suggesting limited surface-level overlap with references. The model produces simplified outputs with a significantly lower FKGL (7.82) compared to the source (13.29), indicating improved readability. The compression ratio (0.48) and low exact copy rate (0.00) suggest aggressive simplification, while the deletions proportion (0.71) is higher than additions (0.18), pointing to a tendency to simplify by removal. Full results are shown in Table 2.

Overall, the results demonstrate that the `plan-guided LLaMA` system effectively simplifies complex biomedical text at sentence-level while maintaining readability and informativeness, with a trade-off in exact lexical overlap.

### 4.2. Evaluation of Task 1.2: Document-level Scientific Text Simplification

The results in Table 3 show that the `llama_summary_simplification` system achieves a SARI score of 40.32, indicating moderate simplification quality. While the BLEU score (7.63) is comparatively low—suggesting some divergence from reference phrasing—the FKGL score of 9.56 reflects improved readability relative to the source. The compression ratio of 0.59 and deletion proportion of 0.70 suggest

**Table 2**

Results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 217 Plain Language Summaries

Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity Score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100.00	100.00	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
plan_guided_llama	217	42.98	6.33	7.82	0.48	0.99	0.46	0.00	0.18	0.71	8.50

aggressive content reduction, contributing to simplification but potentially at the cost of semantic fidelity.

**Table 3**

Results for CLEF 2025 SimpleText Task 1.2 document-level text simplification on 37 aligned Cochrane-auto abstracts

Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity Score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100.00	100.00	11.73	0.56	0.67	0.50	0.00	0.16	0.60	8.71
llama_summary_simplification	37	40.32	7.63	9.56	0.59	0.86	0.42	0.00	0.31	0.70	8.49

The llama\_summary\_simplification method demonstrates effective simplification on the 217 Plain Language Summaries, achieving a strong SARI score of 42.92 and a reduced FKGL of 9.94. The lexical complexity score of 8.55 indicates simplification in vocabulary. However, a relatively low BLEU score of 5.32 and Levenshtein similarity of 0.39 suggest reduced semantic similarity to the reference (Table 4).

Overall, the summary guided method achieves consistent simplification across both the datasets, balancing lower complexity with reduced semantic fidelity.

### 4.3. Comparing LLM based Sentence Simplification Strategies

Table 5 presents a comparison of sentence-level text simplification performance using two strategies: basic LLM-based simplification and LLM-based plan-driven simplification. Across all metrics, the plan-driven approach yields marginal improvements, particularly in BLEU and FKGL, indicating better fluency and readability while preserving fidelity to the source.

### 4.4. Comparing LLM based Document Simplification Strategies

Table 6 presents a comparison between direct LLM-based document simplification and summary-guided simplification across a range of evaluation metrics.

**Table 4**

Results for CLEF 2025 SimpleText Task 1.2 Document-Level Text Simplification: Test Data on 217 Plain Language Summaries

Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity Score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100.00	100.00	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
llama_summary_simplification	217	42.92	5.32	9.94	0.49	0.72	0.39	0.00	0.24	0.75	8.55

**Table 5**

Comparison of Sentence-Level Simplification Metrics between Basic and Plan-Driven LLM Approaches (Appendix A.1)

Method	SARI <sup>1</sup>	BLEU	BERTScore_F1	FKGL
Basic LLM Simplification	42.887	26.4049	0.9005	9.5452
Plan-Driven LLM Simplification	<b>42.985</b>	<b>30.5769</b>	<b>0.9014</b>	<b>9.047</b>

**Table 6**

Performance Comparison: Summary-Guided vs. Direct LLM Document Simplification (Appendix A.2)

Method	SARI <sup>1</sup>	BLEU	BERTScore_F1	FKGL	Token Length
Direct LLM Simplification	<b>43.775</b>	<b>44.6724</b>	<b>0.8605</b>	<b>11.0827</b>	257.00
Summary-Guided Simplification	42.916	31.6618	0.8493	11.2496	<b>249.93</b>

While the direct LLM-based simplification approach slightly outperforms the summary-guided method in terms of standard metrics such as SARI, BLEU, and BERTScore, the summary-guided approach offers distinct benefits. By first generating a high-level summary and then using it to steer the simplification process, this method can produce more coherent and purpose-driven simplifications. The separation of summarization and simplification stages allows the model to better focus on key ideas, potentially reducing redundancy and irrelevant elaborations.

Moreover, the reduced token length in the summary-guided method suggests more concise output, which can be beneficial in applications where brevity and focus are important. Although the readability score (FKGL) is slightly higher, indicating marginally more complex language, the summary-guided approach may improve factual alignment and structural cohesion.

## 5. Conclusions

This study explores the effectiveness of large language models (LLMs) in text simplification across both sentence and document levels. At the sentence level, our plan-driven approach, which prompts the model to select an explicit simplification strategy before generation, yields improved performance in fluency and readability compared to direct simplification. At the document level, we propose a summary-guided simplification pipeline that, while slightly underperforming in standard metrics, offers qualitative advantages in conciseness and coherence by leveraging intermediate summarization as contextual scaffolding.

Our work demonstrates that incorporating structural planning and summarization can enhance LLM-based simplification, especially for longer and more complex texts. Future research will focus on

<sup>1</sup>Only SARI is the official evaluation metric for Task 1. Other metrics are reported for supplementary analysis.



improving the structural prompting framework by introducing an iterative loop that refines prompts based on automatic evaluation metrics, serving as a built-in feedback mechanism.

## Acknowledgements

We thank the Data Science at Georgia Tech (DS@GT) CLEF competition group for their support. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA [13].

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini for grammar and spelling check, as well as assistance in the code for the conducted experiments. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

- [1] H. Saggion, Automatic Text Simplification, Springer International Publishing, Cham, 2022. URL: <https://link.springer.com/book/10.1007/978-3-031-02166-4>. doi:10.1007/978-3-031-02166-4.
- [2] L. Ermakova, et al., Overview of clef 2025 simpletext track: Simplify scientific texts (and nothing more), in: J. Carillo de Albornoz, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2025), LNCS, Springer-Verlag, 2025.
- [3] J. Bakker, et al., Overview of the clef 2025 simpletext task 1: Simplify scientific text, in: G. Faggioli, et al. (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
- [4] 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.
- [5] B. Ondov, K. Attal, D. Demner-Fushman, A survey of automated methods for biomedical text simplification, Journal of the American Medical Informatics Association 29 (2022) 1976–1988. doi:10.1093/jamia/ocac149.
- [6] T. Guidroz, D. Ardila, J. Li, A. Mansour, P. Jhun, N. Gonzalez, X. Ji, M. Sanchez, S. Kakarmath, M. M. Bellaiche, M. Ángel Garrido, F. Ahmed, D. Choudhary, J. Hartford, C. Xu, H. J. S. Echeverria, Y. Wang, J. Shaffer, Eric, Cao, Y. Matias, A. Hassidim, D. R. Webster, Y. Liu, S. Fujiwara, P. Bui, Q. Duong, Llm-based text simplification and its effect on user comprehension and cognitive load, 2025. arXiv:2505.01980.
- [7] D. Fang, J. Qiang, Y. Zhu, Y. Yuan, W. Li, Y. Liu, Progressive document-level text simplification via large language models, 2025. URL: <https://arxiv.org/abs/2501.03857>. arXiv:2501.03857, preprint.
- [8] 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://aclanthology.org/Q16-1029/>. doi:10.1162/tac1\_a\_00107.
- [9] 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 of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Philadelphia, PA, USA, 2002, pp. 311–318.

- [10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. ArXiv preprint arXiv:1904.09675.
- [11] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, B. S. Chissom, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Research Branch Report 8-75, Naval Technical Training, U.S. Naval Air Station, Memphis, TN, 1975.
- [12] T. Tanprasert, D. Kauchak, Flesch–kincaid is not a text simplification evaluation metric, in: A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, W. Xu (Eds.), Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), Association for Computational Linguistics, Online, 2021, pp. 1–14. doi:10.18653/v1/2021.gem-1.1.
- [13] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: <http://www.pace.gatech.edu>.

## A. Codabench Competition Submissions

### A.1. Submissions for Task 1.1

**Table 7**

Codabench Submission Details for SimpleText Task 1.1 (Sentence-Level Simplification)

Competition	ID #	File Name	Task
SimpleText Task 1	321007	dsgt_Task11_llama_simplifier.zip	Task 1.1 Sentence Level
SimpleText Task 1	303306	dsgt_Task11_plan_guided_llama.zip	Task 1.1 Sentence Level

### A.2. Submissions for Task 1.2

**Table 8**

Codabench Submission Details for SimpleText Task 1.2 (Document-Level Simplification)

Competition	ID #	File Name	Task
SimpleText Task 1	321021	dsgt_Task12_llama_simplification.zip	Task 1.2 Document Level
SimpleText Task 1	303443	dsgt_Task12_llama_summary_simplific.zip	Task 1.2 Document Level

## B. Prompt Templates

### B.1. LLM Prompt for Plan-Driven Sentence Simplification

You are a sentence simplifier.  
 Given a document, a sentence from that document, and the next sentence for context, choose an internal simplification strategy from the following options:  
 'rephrase', 'delete', 'split', 'ignore', 'merge'.  
 Then output ONLY the simplified sentence, based on your chosen strategy.

Document: The economic report showed a significant downturn in the last quarter.



Sentence: The economic report showed a significant downturn in the last quarter.

Next Sentence: Unemployment rates also rose sharply.

Simplified: The report said the economy got worse last quarter.

Document: Online social media provide users with unprecedented opportunities to engage with diverse opinions.

Sentence: Online social media provide users with unprecedented opportunities to engage with diverse opinions.

Next Sentence: They also enable misinformation to spread quickly.

Simplified: Social media let people easily share their opinions.

Document: We included seven cluster-randomised trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and the Netherlands. Health professional participants (numbers not specified) included nursing, medical and allied health professionals. Interventions in all studies included implementation strategies targeting healthcare workers; three studies included delivery arrangements, no studies used financial arrangements or governance arrangements. Five trials compared a multifaceted implementation intervention to no intervention, two trials compared one multifaceted implementation intervention to another multifaceted implementation intervention. No included studies compared a single implementation intervention to no intervention or to a multifaceted implementation intervention. Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies. All studies had low risks of selection bias and reporting bias, but high risk of performance bias. Three studies had high risks of bias from non-blinding of outcome assessors or due to analyses used.

Sentence: We included seven cluster-randomised trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and the Netherlands.

Next Sentence: Health professional participants (numbers not specified) included nursing, medical and allied health professionals.

Simplified:

## **B.2. LLM Prompt for Document Summarization**

You are given a complex document. Your task is to write a clear and concise summary that captures the essential information, main arguments, and key findings.

Guidelines:

- Do not include minor details or examples unless crucial to the main idea.
- Focus on the overall message and structure of the document.
- Use simple and accessible language.
- The summary should be understandable without reading the original

document.

### Document:  
{document}

### Summary:

### **B.3. LLM Prompt for Summary-Guided Document Simplification**

Listing 1: Prompt for LLM-Based Summary-Guided Document Simplification

You are given a complex document and its summary. Your task is to rewrite the complex document in a simpler, clearer way while ensuring the meaning aligns with the provided summary.

Guidelines:

- Keep the rewritten version faithful to both the original document and its summary.
- Use simple, accessible vocabulary and sentence structures.
- Avoid introducing new information not present in the original document.
- Retain the key ideas, structure, and intent captured in the summary.

### Complex Document:  
{document}

### Summary:  
{summary}

### Simplified Document: