A Multi-Agent Framework with Diagnostic Feedback for Iterative Plain Language Summary Generation from Cochrane Medical Abstracts

Felipe Arias Russi[1,2], Carolina Salazar Lara[3], Rubén Manrique[1]
[1]Systems and Computing Engineering Department, Universidad de los Andes, Bogotá D.C.
[2]Department of Mathematics, Universidad de los Andes, Bogotá D.C.
[3]Department of Biomedical Engineering, Universidad de los Andes, Bogotá D.C.

{af.ariasr, c.salazar499, rf.manrique}@uniandes.edu.co

Abstract

Plain Language Summaries (PLS) improve health literacy and enable informed healthcare decisions, but writing them requires domain expertise and is time-consuming. Automated methods often prioritize efficiency over comprehension, and medical documents' unique simplification requirements challenge generic solutions. We present a multi-agent system for generating PLS using Cochrane PLS as proof of concept. The system uses specialized agents for information extraction, writing, diagnosis, and evaluation, integrating a medical glossary and statistical analyzer to guide revisions. We evaluated three architectural configurations on 100 Cochrane abstracts using six LLMs (both proprietary and open-source). Results reveal model-dependent trade-offs between factuality and readability, with the multi-agent approach showing improvements for smaller models and providing operational advantages in control and interpretability.

1 Introduction

Health literacy is the ability of an individual to access, understand, and apply health information. This is essential for informed decision-making and effective navigation of healthcare systems. Inadequate health literacy remains a global challenge, contributing to poor treatment adherence, higher hospitalization rates, and health disparities (Berkman et al., 2011; Sørensen et al., 2015; Bahador et al., 2020). Plain Language Summaries (PLS) is a way to reduce health literacy gaps by translating medical texts into clear, accurate, and accessible language for non-technical audiences (Bahador et al., 2020). However, producing high-quality PLS manually is resource-intensive and requires expertise in both medical content and health communication.

Recent advances in LLMs offer new opportunities to automate PLS generation. While early efforts showed that LLMs can produce readable and semantically faithful summaries, most approaches relied on single-pass generation and lacked systematic guardrails for factuality, readability, and regulatory compliance (Turbitt et al., 2023; Van Veen et al., 2024). The increasing complexity of biomedical content and the need for domain-specific standards underscore the importance of structured, multi-step workflows over monolithic generation.

This work builds on our previous research in LLM-based PLS generation (Arias-Russi et al., 2025), which showed the potential of single-prompt models to translate Cochrane abstracts into PLS and Clinical Trials into Protocol Plain Language Summaries (PPLS). However, generating these kind of structured PLS that meet professional standards differs from generic text simplification; it needs adherence to specific templates and guidelines, diagnostic feedback, and systematic quality control. Unlike general simplification tasks that focus only on reducing complexity, structured PLS generation requires writing documents with well-defined structures that balance accessibility with medical accuracy.

Current LLMs struggle to balance simplicity with factual accuracy, often oversimplifying complex medical content or preserving meaning at the cost of readability (Li et al., 2024). Our prior work revealed similar domain-specific challenges, requiring distinct prompts for different document types (Cochrane PLS and PPLS). Also, this approach did not provide mechanisms to identify specific problems in the generated PLS drafts or provide targeted corrections.

Based on these limitations, this research aims to: (1) develop tools that allow a better understanding of what makes a text non-compliant with PLS standards and how to systematically address these issues, and (2) create a multi-agent framework supported by diagnostic tools that can both generate structured PLS and evaluate their quality through some iterative refinement.

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 87–104
November 4-9, 2025 ©2025 Association for Computational Linguistics

We propose a framework that decomposes PLS generation into specialized subtasks, each handled by dedicated agents: information extraction, writing, diagnostic, and evaluation (Figure 1). The key component is a diagnostic feedback loop where evaluation agents identify specific complexity issues and guide targeted revisions using verifiable criteria. We instantiate this framework for Cochrane PLS generation, as their guidelines (Pitcher et al., 2022) provide an useful template ideal for testing structured document generation.

2 Related Work

2.1 Plain Language Summaries

Recent work in biomedical text simplification explores different approaches. The BioLaySumm shared task (Xiao et al., 2025) focuses on generating lay summaries from biomedical abstracts. Participants showed various strategies: supervised fine-tuning of T5 and LLaMA models (Zhang et al., 2025); extract-then-summarize pipelines with persona-based prompts and UMLS definitions (Gupta and Krishnamurthy, 2025); structured prompting with dynamic few-shot selection and RAG (Lossio-Ventura et al., 2025); and Tree-of-Thought prompting with hybrid methods (Sivagnanam et al., 2025). Fine-tuning approaches include QLoRA adaptation with iterative refinement (Binti Moriazi and Sung, 2025) and readability-controlled instruction tuning (Tran et al., 2025). Others focus on preprocessing (Dehkordi et al., 2025) or evaluation metrics (Lyu and Pergola, 2024a; Scholz and Wenzel, 2025). More related work can be found in our previous work (Arias-Russi et al., 2025).

Our work addresses a complementary task: generating structured PLS following Cochrane's established template. Unlike lay summaries that prioritize readability alone, structured PLS must adhere to specific section requirements (Title, Key Messages, Background, Methods, Results, Limitations, Currency), maintain professional standards, and balance accessibility with regulatory compliance. We used some evaluation metrics from BioLaySumm and related work to assess both readability and structural conformance.

2.2 Multi-agent Systems for Text Simplification

Multi-agent systems have emerged as a promising approach for text processing. The Society of Medical Simplifiers (Lyu and Pergola, 2024b) simplifies biomedical literature into general plain language text, using five agents in three interaction loops—a Layperson Agent identifies technical terms, a Medical Expert provides clarifications, and a Simplifier Agent edits text, focusing on making content accessible without following specific templates or guidelines. ExpertEase (Mo and Hu, 2024) generates grade-specific simplified documents for educational purposes, using Expert, Teacher, and Student agents that calibrate text complexity for target reading levels like 2nd-3rd grade. For diagnostic applications, MedAgent-Pro (Wang et al., 2025) produces evidence-based medical diagnoses with supporting visual evidence rather than simplified text, employing RAG, Planner, and Tool agents to integrate clinical guidelines for diseases like glaucoma. AgentSimp (Fang et al., 2025) creates general simplified documents focusing on coherence and metaphor handling, using nine agents including a Metaphorical Analyst and Terminology Interpreter, but without adherence to medical communication standards or structured templates.

Rather than generating general simplified text, we aim to create structured PLS that facilitate the work of medical writers, helping to automate the process to get high-quality PLS drafts. Multi-agent systems are particularly suited for this task because structured documents can be decomposed into separate subtasks (extraction, integration, evaluation, and refinement) that align naturally with specialized agents (see the conceptual framework in Figure 1). From this abstract idea of generating structured PLS, we designed a multi-agent approach specifically for Cochrane PLS, as their detailed guidelines provide a well-defined template that serves as an ideal test case for our framework (Pitcher et al., 2022). Our primary approach uses an on-demand evaluator tool that the editor agent calls when needed. Inspired by Self-Refine (Madaan et al., 2023), we also tested an alternative iterative approach where the evaluator runs multiple refinement cycles independently, instead of being invoked by the editor agent.[1]

3 Methodology

We present the methodology for developing and evaluating a multi-agent system for automatic gen-

[1] All materials including agent prompts, datasets, evaluation scripts, and workflow implementation are available at: https://github.com/feliperussi/tsar-2025-medical-writing-agent-cochrane

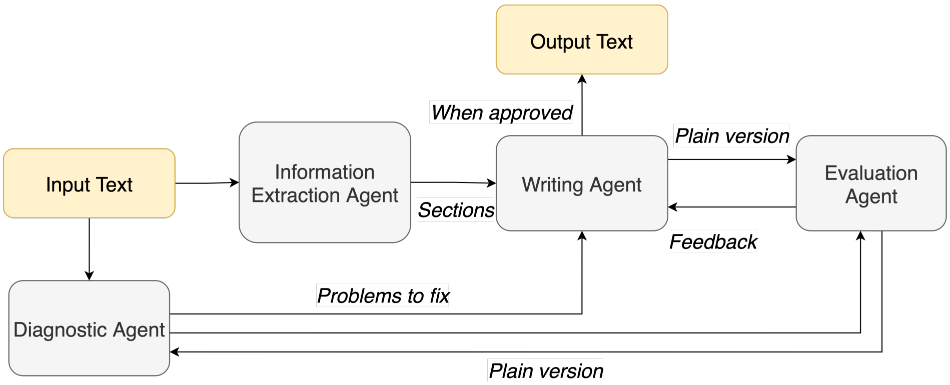

Figure 1: Conceptual framework for generating PLS. The system orchestrates specialized agents for information extraction, writing, terminology simplification, and evaluation. A diagnostic feedback loop enables evaluation agents to identify complexity issues and guide targeted revisions. This modular architecture supports structured, accurate, and readable PLS aligned with Cochrane standards.

eration of PLS from Cochrane medical abstracts. The approach leverages linguistic analysis, medical glossaries, domain-specific glossaries, and advanced language models to produce accessible medical communication. The methodology follows six main steps: (1) Data Collection and Processing, (2) PLS Linguistic Feature Extraction, (3) Percentiles for PLS Diagnosis, (4) Develop Diagnostic Tools, (5) Define Specialization, and (6) Define Architecture for the multi-agent system (see Appendix A for the complete workflow diagram).

3.1 Data Collection and Processing

3.1.1 Data Source

We collected 9,469 Abstract-PLS pairs (18,938 documents) extracted from the Cochrane Database of Systematic Reviews (1996-2025), spanning nearly 30 years of evolving medical communication practices (Cochrane Library, 2025). Prior work has highlighted significant content misalignment between abstracts and their corresponding PLS, where summaries often incorporate information from full-text articles (Bakker and Kamps, 2024). To address this issue, the authors proposed a new dataset (Cochrane-auto) that ensures better alignment between abstracts and PLS. Although we recognize this problem, we preferred to use the original Cochrane data to maintain the integrity of the dataset, and ensure easier evaluations and comparison between different strategies.

3.1.2 Data Processing Pipeline

We filtered and split the dataset into reference and test sets, applying minimum length thresholds (200 words for abstracts, 150 for PLS). After filtering, we retained 16,308 documents from the original 18,938 documents. For our experiments, we used only PLS texts from the reference corpus to compute statistical thresholds for the evaluation agent (ignoring their paired abstracts), and the test set pairs for generation and evaluation. The test set comprised recent publications (2023-2025) to align with the 2022 Cochrane PLS guidelines. Table 1 shows the data used in our study (complete dataset splits are available in the repository).

Data	PLS	Abstracts	Total
Reference	6,754	–	6,754
Test	100	100	200
Total	6,854	100	6,954

Table 1: Dataset distribution. Reference corpus: PLS texts for computing percentile thresholds. Test set: Abstract-PLS pairs for evaluation.

3.2 PLS Linguistic Feature Analysis

3.2.1 Feature Extraction

We extracted 18 linguistic features from each document (see Appendix B.1) comprising 9 readability indices, 4 structural metrics, 3 vocabulary measures, and 2 content density indicators. These features enable the multi-agent system to compare any text against typical PLS patterns using percentile distributions. The Cochrane PLS guidelines (Pitcher et al., 2022) recommend specific criteria: maximum 850 words, active voice, personal pronouns, and 20 words per sentence average.

These metrics answer concrete diagnostic questions: "Where does this text's passive voice usage fall compared to typical PLS?" or "Is this sentence length in the common range?" Using the percentile thresholds from Section 3.2.2, the evaluator agent identifies specific deviations and provides feedback to the editor agent for improving the draft.

3.2.2 Percentiles for PLS Diagnosis

We computed statistical thresholds from our reference corpus of PLS texts based on percentiles to create a reference baseline for evaluation. These thresholds function as an interpretable diagnostic tool for the multi-agent system (or a human evaluator), providing explicit information about where generated text falls within the distribution of each linguistic feature. This approach enables specific improvements based on concrete positional feedback rather than abstract quality scores. We chose percentiles over machine learning approaches (e.g., gradient boosting, decision trees with feature importance) because when communicating diagnostic feedback to an evaluator, percentile distributions provide the most natural and interpretable way to identify which features have atypical values and where they fall relative to typical patterns.

We use a dual percentile system that adapts to each metric's direction. For metrics where lower values are preferred, we label the ranges as P25, P50, P75, and P90, corresponding to the actual percentiles. For metrics where higher values are preferred, we maintain the same labels but apply them to the inverse percentiles (P75, P50, P25, and P10 respectively). This ensures P25 and P75 consistently identify the best quartile regardless of metric direction.

The tool provides the evaluator agent with specific positional information (e.g., "passive voice falls in P90 range") that directly translates to actionable feedback. While deviation from typical patterns (beyond P10 or P90) suggests that a revision may be warranted, such deviations do not automatically indicate poor quality. For instance, a text scoring in the P90 range for complex vocabulary may still be considered plain language if those terms are medically necessary and properly defined. Complex medical procedures may require precise technical terminology that cannot be simplified without losing essential meaning (a limitation discussed in Section 7). The percentile ranges serve as diagnostic indicators rather than absolute quality judgments, guiding targeted improvements while preserving content accuracy. Appendix B.2 presents the complete thresholds used as the diagnostic baseline. These thresholds are then integrated into the PLS Evaluation Tool (Section 3.3.2) to enable automated quality assessment.

3.3 Diagnostic Tools Development

We developed two deterministic tools that emulate the resources and decision-making process of professional medical writers: a medical glossary service and a PLS evaluation tool. These tools provide the editor agent with the same type of guidance a human medical writer would use, including professionally-recommended terminology simplifications and rapid interpretable indicators to identify atypical text patterns. By grounding our tools in professional practices mentioned in the Cochrane PLS guidelines, we enable systematic evaluation and improvement of generated text.

3.3.1 Medical Glossary Tool

We collected 20,637 medical terms with plain language alternatives from 11 professional dictionaries recommended by the Williams (2025) and Cochrane Plain Language Summary Guidelines (Pitcher et al., 2022, page 29, Appendix 1). Table 2 shows the distribution of terms across sources, spanning cancer terminology, public health, diabetes, genetics, clinical trials, and other healthcare domains (see Appendix C for detailed source descriptions). The tool uses a longest-match regex algorithm to identify medical terms in submitted text and returns structured JSON with the term, its plain language alternative, and source; mirroring how a medical writer would consult reference materials during revision.

Source	Focus Area	Terms
NCI-C	Cancer terminology	9,416
NCI-D	Cancer drugs	9,144
CDC-T	Public health	891
ADA-D	Diabetes	247
NCI-G	Genetics	242
UIowa	General	242
MRCT	Clinical trials	187
WA-PH	Immunization	104
WebMD-A	Asthma	75
CCIIO	Insurance	59
Cochrane	Systematic reviews	30
Total		20,637

Table 2: Medical glossary sources with term counts and focus areas.

3.3.2 PLS Evaluation Tool

Using the percentile thresholds computed in Section 3.2.2, the evaluation tool provides rapid, interpretable assessments that a medical writer would typically perform manually. Given a text, it computes 18 linguistic features using standard computational linguistics algorithms (no LLMs involved) and maps each to its percentile range based on the thresholds. The tool generates evaluation reports showing word count compliance, metric-by-metric analysis, percentile assignments, and revision suggestions for atypical patterns. This human-interpretable output allows the editor agent to make informed decisions about which deviations warrant revision and which are contextually justified (see Appendix B.3 for example output).

3.4 Multi-Agent System Architecture

We designed 14 specialized agents and their corresponding prompts organized into four functional groups: Information Extraction, Writer, Diagnostic, and Evaluation. Each prompt was developed based on the Cochrane PLS template (Pitcher et al., 2022), iteratively refined through a combination of Gemini 2.5 Pro and Claude Opus 4.1 generations with human revision to ensure alignment with Cochrane guidelines (see the repository for all the prompts). Figure 2 illustrates the complete multi-agent architecture with all components and their interactions. We first describe the core agent functionalities below; the architectural variants (Baseline, V_1, V_2) are presented in Section 3.4.4.

3.4.1 Information Extraction Agents

These agents work (mostly) in parallel to extract different parts of the abstract simultaneously, with each agent based on a specific section of the Cochrane PLS template (Pitcher et al., 2022).

To begin, the **Plain Titles Agent** reformulates technical review titles into patient-friendly questions, following Cochrane's recommendation to use question-based titles that directly address patient concerns. When complex medical terms appear in titles, they are either replaced with plain language alternatives or clearly defined for patient understanding. Similarly, the **Key Messages Agent** extracts 2–3 main findings as bullet points, ensuring technical terms are either avoided or explained.

For the introductory content, the **Background Agent** creates 2–3 subsections with question-based headings that explain what the health condition is ("What is [condition]?"), why it matters, and what the researchers wanted to find out ("What did we want to find out?"). This output includes the review aims, which are then referenced by the **Methods Agent**. Building on these aims, the Methods Agent writes "What did we do?" in 1–2 sentences, ensuring direct connection to the research goals. It describes three key actions—searching for studies, combining results, and rating confidence in evidence—using standardized phrases like "We searched for studies that compared. . . " while avoiding specific study design mentions unless essential.

The **Results Agent** generates "What did we find?" by coordinating two specialized tool agents: the **Characteristics Agent** extracts study details (number of studies, participants, duration, countries), while the **Findings Agent** translates technical findings into plain language, simplifying narratives and avoiding technical statistical terms.

To complete the extraction pipeline, the **Limitations Agent** identifies constraints from the review findings, and the **Date Extraction Agent** standardizes when the evidence was collected.

3.4.2 Writer Agents

The **Assembly Agent** takes all the pieces from the extraction agents and combines them into one complete summary. It follows the exact order required by Cochrane: title, key messages, background sections, "What did we want to find out?", "What did we do?", "What did we find?", and so on.

The **Editor Agent** improves the assembled draft through revision. It checks for problems like unexplained medical terms, complicated sentences, or forbidden elements (like acronyms or statistical data), working in coordination with evaluation mechanisms to ensure quality standards are met.

3.4.3 Diagnostic and Evaluation Agents

These agents provide specialized diagnostic support and quality assessment throughout the writing process. The **Technical Terms Recognizer Agent** identifies remaining medical terms that require explanation in plain language context.

The **Evaluator Agent** is a hybrid agent with dual functionality. In its diagnostic capacity, it identifies specific issues by leveraging the diagnostic tools developed in the previous section. As an evaluator, it performs comprehensive quality checks by verifying factual accuracy through comparing drafts against original extraction outputs to detect hallucinations, ensuring all required sections are present, and using the PLS Evaluation Tool to assess readability metrics against the thresholds from Section 3.2.2.

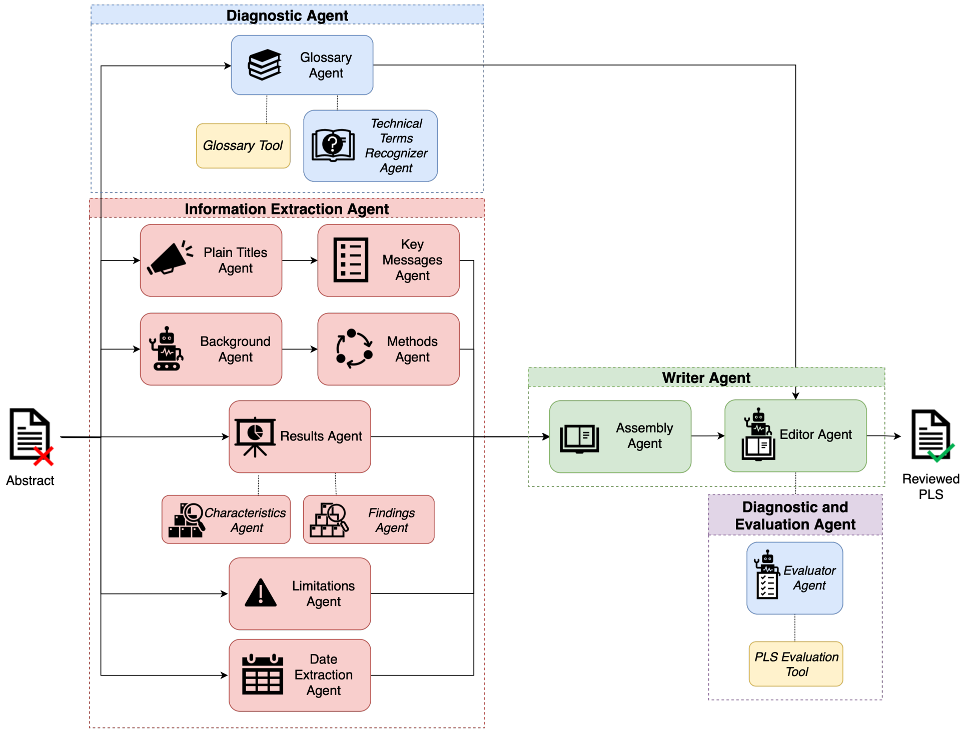

Figure 2: Multi-agent system architecture for PLS generation (V_1). Information Extraction Agents (red) process the abstract in parallel to extract different components. Writer Agents (green) assemble and refine content sequentially. Diagnostic Agents (blue) provide specialized terminology support using deterministic tools (yellow). The hybrid Diagnostic and Evaluation Agent (purple) performs both diagnostic analysis and final quality assessment. The pipeline flows from extraction through assembly and iterative refinement until the PLS is approved.

3.4.4 Architectural Variants

We evaluated three system configurations to assess different approaches to task decomposition and quality control:

- **Baseline:** This configuration consolidates the entire workflow of the agents viewed into a single, complete prompt. This allows the model to generate the complete PLS in a single step, without the need for iterative refinement or specialized agent roles.

- **Version 1 (V_1):** The primary multi-agent architecture shown in Figure 2 that integrates all the agents and tools described in the previous sections. In this configuration, specialized agents handle different subtasks as designed, the Evaluator Agent operates as an on-demand tool that the Editor can invoke when needed, and the Medical Glossary Tool provides terminology support through the Glossary Agent.

- **Version 2 (V_2):** An iterative variant where the Evaluator operates as a standalone agent that systematically evaluates each draft and provides feedback to the Editor. The process terminates when either (1) the Evaluator approves the draft based on quality criteria, or (2) the maximum of 10 iterations is reached. Additionally, V_2 removes the Medical Glossary Tool to assess whether explicit medical dictionaries are necessary. For more details, V_2 iterative approach is presented in Appendix A.1.

These variants allow us to compare monolithic versus multi-agent approaches, tool-based versus iterative evaluation strategies, and assess the impact of explicit glossaries on generation quality.

4 Evaluation and Results

We evaluated the three architectural configurations (Baseline, V_1, V_2) described in Section 3.4.4 across multiple language models to assess the impact of multi-agent decomposition and iterative refinement on PLS generation quality.

4.1 Experimental Setup

We implemented the system using n8n (2025) workflow automation platform and evaluated on 100 Abstract-PLS pairs from 2023-2025. We tested six language models: Gemini 2.5 Pro (Google AI for Developers, 2025), GPT-5 (OpenAI, 2025b) via Azure OpenAI, GPT-OSS-120B and GPT-OSS-20B (OpenAI, 2025a) via Together AI (2025), Llama 3.3 70B, and Llama 3.2 3B.

All models were evaluated with the Baseline configuration. V_1 was tested with Gemini 2.5 Pro, GPT-5, and GPT-OSS-120B. V_2 was tested with Gemini 2.5 Pro, GPT-OSS-120B, GPT-OSS-20B, Llama 3.3 70B, and Llama 3.2 3B (GPT-5 could not be run with V_2 due to implementation constraints). Additionally, we tested a hybrid variant where Llama 3.3 70B (V_2) used Gemini 2.5 Flash as the Evaluator Agent. All models used temperature 0.0 except GPT-5 (temperature 1.0 due to API constraints). For each run, all agents within a configuration used the same underlying model.

4.2 Evaluation Metrics

Generated PLS texts were evaluated across three dimensions:

1. Relevance: Measuring the semantic similarity between the LLM-generated summaries and human-written reference PLS using BERTScore (Zhang et al., 2020), which computes token-level similarity through contextual embeddings. We also calculated similarity against original abstracts to assess information retention.

2. Factuality: Evaluating the consistency of generated content with source abstracts (ensuring no contradictory information is introduced) using AlignScore (Zha et al., 2023) and MeaningBERT (Beauchemin et al., 2023), which measure factual alignment and semantic equivalence respectively.

3. Readability: Assessing grammaticality and ease of comprehension through computational metrics. Additionally, we computed percentile distributions across all 18 linguistic features to measure conformity with typical PLS patterns, with “Best 25%” representing the percentage of features in the optimal quartile.

4.3 Results and Analysis

Our evaluation reveals inconclusive results with mixed patterns across models and configurations (Tables 3 and 4). No single architectural approach consistently outperforms others across all quality dimensions and model types, with results suggesting fundamental trade-offs between factuality and readability that manifest differently depending on base model characteristics.

Gemini 2.5 Pro is the only model where we can validly compare all three configurations (GPT-OSS-120B V_1 had implementation limitations where the evaluator tool could only be invoked once, making it unsuitable for valid comparison). Gemini shows a clear trade-off pattern: the V_2 configuration achieves the best semantic similarity and factuality scores to the reference corpus, but the baseline produces more readable text across most readability indices. The V_1 configuration (with medical glossary and evaluator as tool) achieves readability metrics similar to baseline and better than V_2, suggesting that the medical glossary may help balance factuality and simplicity in agentic workflows, though we cannot confirm this conclusively. This pattern reflects a general trend where agentic configurations tend to improve relevance and factuality metrics, while baseline configurations often produce more readable outputs, though this varies across models.

GPT-5 agentic (V_1) underperforms its baseline across most metrics, with the exception of AlignScore to the original abstract. For GPT-OSS-120B, the baseline outperforms both V_1 and V_2 variants in factuality to reference and readability metrics, though V_2 shows improvements over V_1.

Smaller models exhibit distinct behaviors. GPT-OSS-20B demonstrates considerable improvements with the V_2 architecture in semantic similarity and factuality compared to baseline, with modest impact on readability. Llama 3.2 3B shows an interesting pattern where both baseline and agentic configurations achieve the highest AlignScore to the original abstract among all tested models (with baseline being globally highest), yet both produce the least readable outputs, with the agentic version particularly affected. This suggests smaller models may compensate for limited capabilities by maintaining strict alignment to source material while struggling with linguistic transformation. The hybrid configuration (Llama 3.3 70B with Gemini 2.5 Flash as evaluator) achieves competitive semantic quality while substantially improving readability compared to standard Llama 3.3 70B, demonstrating that evaluator quality impacts generation quality.

Examining conformity to typical PLS patterns (Best 25% in Table 4), results are inconclusive. Gemini 2.5 Pro improves with agentic configurations, but this does not generalize to other models. Most configurations achieve conformity levels comparable to human reference patterns. Ap-

Model	Approach	vs. Original Abstract			vs. Reference PLS		
		BERTScore	MeaningBERT	AlignScore	BERTScore	MeaningBERT	AlignScore
Reference (human)		0.8482	0.6825	0.7551	–	–	–
Gemini 2.5 Pro	*Baseline*	0.8352	0.5957	0.7820	0.8701	**0.7162**	0.7002
	Agentic (V1)	0.8420	0.5928	0.7909	0.8708	0.6907	0.7157
	Agentic (V2)	**0.8469**	**0.6136**	**0.7992**	**0.8736**	0.7153	**0.7219**
GPT-5	*Baseline*	**0.8342**	**0.6075**	0.7692	**0.8619**	**0.6873**	**0.6598**
	Agentic (V1)	0.8278	0.5863	**0.7887**	0.8499	0.6718	0.6522
GPT-OSS-120B	*Baseline*	0.8407	0.6477	0.7696	0.8650	0.7346	**0.6878**
	Agentic (V1)	0.8321	0.6468	**0.7995**	0.8562	0.7304	0.6708
	Agentic (V2)	**0.8464**	**0.6557**	0.7953	**0.8651**	**0.7393**	0.6595
GPT-OSS-20B	*Baseline*	0.8327	0.5857	0.7396	0.8565	0.6697	0.6480
	Agentic (V2)	**0.8422**	**0.6680**	**0.8002**	**0.8615**	**0.7534**	**0.6590**
Llama 3.3 70B	*Baseline*	0.8514	**0.6985**	0.7536	0.8679	0.7158	0.7076
	Agentic (V2)	**0.8549**	0.6818	**0.7823**	0.8708	**0.7446**	**0.7140**
	Agentic (V2 + Gemini 2.5 Flash)	0.8485	0.6514	0.7644	**0.8711**	0.7325	0.6982
Llama 3.2 3B	*Baseline*	0.8477	0.6566	**0.8499**	0.8467	0.6302	**0.6982**
	Agentic (V2)	**0.8551**	**0.6952**	0.8403	**0.8532**	**0.6672**	0.6706

Table 3: Semantic similarity and factuality metrics for all tested models and approaches. **Bold** indicates best performance within each model, underlined indicates worst within each model. Gray shading highlights best global performance, red shading highlights worst global performance. Human reference excluded from comparisons. All metrics averaged across 100 test samples.

Model	Approach	Words	FKGL↓	ARI↓	CLI↓	FRE↑	GFI↓	LIX↓	SMOG↓	RIX↓	DCRS↓	Best 25%	P25%	P50%	P75%	P90%	P10%
Original Abstract		868	13.85	14.07	11.11	42.12	20.39	59.26	17.37	8.60	8.75	27.61	7.39	11.89	20.22	15.89	1.00
Reference PLS (human)		655	11.38	11.46	11.21	49.23	16.21	50.35	14.23	6.07	7.37	52.50	25.89	17.89	26.61	14.78	0.22
Gemini 2.5 Pro	*Baseline*	661	**8.26**	**8.01**	**9.07**	**64.45**	**12.38**	**41.21**	**12.74**	**4.02**	**6.53**	78.44	45.33	5.28	33.11	12.89	0.00
	Agentic (V1)	591	8.56	8.49	9.42	63.16	12.51	42.02	12.16	4.21	6.39	82.00	46.56	9.22	35.44	7.94	0.00
	Agentic (V2)	567	9.40	9.47	10.16	58.78	13.54	44.64	12.42	4.74	6.65	**83.22**	44.67	9.11	38.56	7.06	0.00
GPT-5	*Baseline*	866	**9.62**	**9.89**	**10.46**	**57.97**	**13.49**	**45.08**	**14.62**	**4.87**	**6.74**	**63.83**	41.33	5.56	22.50	9.61	0.00
	Agentic (V1)	879	10.97	11.59	11.71	51.14	15.12	48.34	15.50	5.61	7.71	49.33	26.67	14.00	22.67	8.22	0.00
GPT-OSS-120B	*Baseline*	623	**9.82**	**9.94**	**9.74**	**59.10**	**14.76**	**45.98**	**12.96**	**5.15**	**6.77**	73.44	39.39	10.56	34.06	13.22	0.00
	Agentic (V1)	743	11.05	11.25	11.12	51.08	16.55	49.62	15.03	5.91	8.01	46.72	22.50	17.94	24.22	12.11	0.00
	Agentic (V2)	623	10.83	11.00	10.82	52.66	16.32	49.21	13.87	5.82	7.75	**83.22**	44.67	9.11	38.56	7.06	0.00
GPT-OSS-20B	*Baseline*	541	**10.17**	**8.89**	**7.08**	**58.28**	**13.85**	**44.24**	**11.84**	**5.04**	**6.32**	**71.88**	40.11	15.80	31.76	8.85	0.11
	Agentic (V2)	632	10.50	10.27	10.14	53.76	15.78	47.78	13.84	5.51	7.74	59.32	31.00	13.52	28.32	16.97	0.17
Llama 3.3 70B	*Baseline*	477	**12.01**	**12.31**	**11.05**	**48.55**	16.89	**51.89**	**12.65**	6.57	**6.87**	61.56	29.17	23.33	32.39	9.61	0.44
	Agentic (V2)	497	12.23	12.46	12.59	42.72	16.91	53.08	13.41	6.60	7.54	57.28	26.17	17.61	31.11	16.78	0.00
	Agentic (V2 + Gemini 2.5)	514	9.97	10.13	10.91	54.97	**14.24**	46.69	12.48	**5.11**	6.84	**58.89**	20.78	24.22	38.11	10.39	0.17
Llama 3.2 3B	*Baseline*	477	**9.58**	**8.95**	**10.60**	**53.56**	**14.32**	**45.81**	**12.88**	**4.68**	**6.82**	**69.67**	45.11	19.33	24.56	6.06	0.28
	Agentic (V2)	536	12.93	12.72	12.65	37.95	17.57	54.41	14.06	6.95	7.96	45.69	19.98	21.93	25.71	14.30	1.50

Table 4: Readability metrics and percentile distribution for all tested models. Left: average readability scores (arrows: ↓ lower is better, ↑ higher is better). Right: percentage of linguistic features in each percentile range across 18 selected features. Best 25% represents percentage in optimal quartile. **Bold** indicates best performance within each model, underlined indicates worst within each model. Gray shading highlights best global performance, red shading highlights worst global performance. Original Abstract and Reference PLS (human) excluded from comparisons. Averages computed across 100 test samples.

pendix A.2 presents a detailed example of the V_2 iterative refinement process, illustrating how the Evaluator Agent provides structured feedback that guides draft improvements from 83.33% to 94.44% best quartile conformity.

5 Discussion

Our evaluation reveals inconclusive results regarding which approach is superior, with both baseline and multi-agent configurations showing distinct advantages depending on use case requirements. The baseline proves remarkably effective when properly designed with comprehensive instructions based on Cochrane guidelines, demonstrating that systematic prompt engineering grounded in domain standards can produce high-quality PLS. The multi-agent architecture, while not completely superior, provides specific benefits in certain contexts.

For smaller models, the multi-agent approach shows improvements in relevance and factuality metrics. GPT-OSS-20B improves in semantic similarity and factual alignment when using the V_2 configuration, though with a slight deterioration in readability. Notably, smaller models achieve competitive or higher factuality scores compared to larger models, something we noticed in our prior work (Arias-Russi et al., 2025), suggesting these models may be more conservative in adhering to source material. A potential strategy to address readability limitations would involve using a smaller model for content extraction followed by a larger model for final linguistic refinement, potentially offering cost-effective generation while preserving factual accuracy.

The multi-agent architecture provides operational advantages in terms of control and interpretability. Separating extraction, assembly, and evaluation into distinct agents makes each component transparent and independently modifiable. While the baseline consolidates all instructions into a single comprehensive prompt, the decomposed approach allows for targeted refinement of specific subtasks without affecting the entire pipeline. This modularity also produces intermediate outputs for each section, which are stored separately in our repository and can be inspected individually for diagnostic purposes.

However, the multi-agent approach incurs higher token costs due to multiple agent invocations and memory preservation through context repetition across agents. The V_1 configuration, which includes the medical glossary tool, is particularly token-intensive, with processing costs reaching approximately 10M tokens (input + output combined) per 100-abstract batch. Most tokens are input tokens, which are typically cheaper than output tokens, though iterative configurations may generate similar text multiple times, increasing output costs. The baseline, requiring only a single model invocation per abstract, proves more cost-effective in terms of API usage. This cost-benefit trade-off must be considered when selecting an approach for production deployment.

The effectiveness of smaller models like GPT-OSS-20B with the multi-agent approach could enable local deployment in medical contexts where data privacy is critical. While the current task of simplifying published Cochrane reviews does not involve sensitive information, other medical text simplification scenarios could benefit from local processing. For instance, healthcare providers might need to simplify patient-specific medical reports or treatment explanations without transmitting sensitive data to external APIs. In such contexts, the ability to run smaller models locally while maintaining reasonable quality through multi-agent decomposition could provide a viable solution.

Beyond technical performance metrics, this work addresses the practical need of facilitating medical writers' work and improving health information accessibility. Our framework provides diagnostic tools and automated first drafts that meet professional PLS standards, potentially reducing the manual effort required to produce accessible health communication materials.

6 Future Work

Future work could explore alternative evaluation approaches beyond percentile-based diagnostics, including different metrics and quality assessment methods for iterative refinement. Multi-agent hybrid systems where different agents use specialized models could balance cost and quality. Extending the architecture to other plain language formats such as PPLS or documents following the Federal Plain Language Guidelines (Williams, 2025) would test its generalizability. Additionally, incorporating full-text papers as input sources rather than abstracts alone could address content coverage limitations, potentially leveraging datasets like Cochrane-auto for improved alignment between source and simplified text.

7 Limitations

Our system may perpetuate suboptimal information prioritization patterns from the training data (Bakker and Kamps, 2024). Computational constraints and API rate limits restricted experimentation scope, and we did not conduct formal statistical hypothesis testing. GPT-5 could not be tested with V_2 due to API rate limits, only supports temperature 1.0[2] preventing deterministic generation, and exhibited inconsistent behavior (sometimes the model did not use the evaluation tool for V_1). GPT-OSS-120B V_1 had implementation issues with the evaluator tool. V_1 intermediate outputs were lost due to storage issues; V_2 outputs are available in the repository. The V_2 example in Appendix A.2 represents a single cherry-picked case. Our percentile-based evaluation framework represents statistical conformity rather than absolute quality, and strict percentile ranges could penalize innovative plain language strategies.

8 Lay Summary

Medical research papers often contain complex language that makes them difficult for patients and the general public to understand. Plain Language Summaries help solve this problem by explaining research findings using everyday words. Organizations like Cochrane create these summaries for their systematic reviews, which combine results from many studies to answer health questions. However,

[2] https://web.archive.org/web/20250903093505/https://community.openai.com/t/temperature-in-gpt-5-models/1337133/20

writing plain language summaries requires medical expertise and takes considerable time. This creates a bottleneck in making health information accessible to everyone.

We wanted to find out whether computer systems using AI could automatically generate high-quality plain language summaries. Specifically, we tested whether breaking down the writing task into smaller steps handled by specialized AI agents would work better than using a single comprehensive instruction. We also wanted to know if this approach would be more helpful for some AI models than others.

We built a system that divides summary writing into four stages: extracting information from the medical abstract, assembling it into a draft, checking for medical terms that need simplification, and evaluating readability. The system uses a medical dictionary with over 20,000 terms and their plain language alternatives. It also includes a statistical analyzer that compares the generated text against patterns found in human-written summaries. We tested this system using 100 Cochrane medical abstracts and six different AI models, ranging from large commercial models to smaller open-source ones.

Our results are mixed and there is no clear winner among the designed architectures. The baseline approach, which uses a single well-designed instruction, performed surprisingly well. The multi-agent system did not consistently outperform the baseline, but it showed specific advantages. Smaller AI models improved notably when using the multi-agent approach, achieving better accuracy in preserving medical facts, though sometimes at the cost of readability. We also found that the multi-agent system provides greater control and less black-box effect, allowing users to inspect and modify individual steps separately.

Healthcare organizations and research institutions working with limited computational resources could benefit from these findings. The results suggest that smaller, locally-run AI models combined with the multi-agent approach could generate reasonably accurate summaries while maintaining data privacy. The modular design also makes it easier to adapt the system for different types of medical writing beyond Cochrane summaries. However, more work is needed to improve readability when using smaller models and to reduce the computational costs of the multi-agent approach.

References

American Diabetes Association. 2024. Common Diabetes Terms. Accessed: August 23, 2025.

Jonathan Anderson. 1983. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496. Publisher: [Wiley, International Reading Association].

Andrés Arias-Russi, Carolina Salazar-Lara, and Rubén Manrique. 2025. Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 269–284, Albuquerque, New Mexico. Association for Computational Linguistics.

B. Bahador, S. Baedorf Kassis, H. Gawrylewski, and et al. 2020. Promoting equity in understanding: A cross-organizational plain language glossary for clinical research. *Medical Writing*, 29(4):10–15.

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.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.

N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of Internal Medicine*, 155(2):97–107.

Nur Alya Dania Binti Moriazi and Mujeen Sung. 2025. KHU_LDI at BioLaySumm2025: Fine-tuning and refinement for lay radiology report generation. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 256–268, Vienna, Austria. Association for Computational Linguistics.

Center for Consumer Information & Insurance Oversight. 2024. Uniform glossary of health coverage and medical terms. Updated for plan or policy years beginning on or after January 1, 2024. Accessed: August 23, 2025.

Centers for Disease Control and Prevention. 2011. Plain Language Thesaurus for Health Communications. Accessed: August 23, 2025.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Google-Books-ID: 2nbuAAAAMAAJ.

Cochrane Library. 2025. Cochrane Database of Systematic Reviews. https://www.cochranelibrary.com/cdsr/reviews.

Meri Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284. Place: US Publisher: American Psychological Association.

Mahshad Koohi Habibi Dehkordi, Yehoshua Perl, Fadi P. Deek, Zhe He, Vipina K. Keloth, Hao Liu, Gai Elhanan, and Andrew J. Einstein. 2025. Improving Large Language Models' Summarization Accuracy by Adding Highlights to Discharge Notes: Comparative Evaluation. *JMIR Medical Informatics*, 13(1):e66476. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative Document Simplification Using Multi-Agent Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233. Place: US Publisher: American Psychological Association.

Google AI for Developers. 2025. Gemini models: Gemini API | Google AI for Developers. Latest update: August 2025; archived at `https://web.archive.org/web/20250825013047/https://ai.google.dev/gemini-api/docs/models#previous-experimental-models`.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill. Google-Books-ID: ofI0AAAAMAAJ.

Aaradhya Gupta and Parameswari Krishnamurthy. 2025. Shared task at biolaysumm2025 : Extract then summarize approach augmented with UMLS based definition retrieval for lay summary generation. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 185–189, Vienna, Austria. Association for Computational Linguistics.

Human Subjects Office, University of Iowa. 2021. Medical terms in lay language. Last updated: 05/03/2021. Accessed: August 23, 2025.

Peter Kincaid, Robert P. Fishburne, 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. In *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Large Language Models for Biomedical Text Simplification: Promising But Not There Yet. *arXiv preprint*. ArXiv:2408.03871 [cs].

Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrista-Salas, Francisco Pereira, and Diana Inkpen. 2025. 5cNLP at BioLaySumm2025: Prompts, retrieval, and multimodal fusion. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 215–231, Vienna, Austria. Association for Computational Linguistics.

Chen Lyu and Gabriele Pergola. 2024a. SciGisPy: a novel metric for biomedical text simplification via gist inference score. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 95–106, Miami, Florida, USA. Association for Computational Linguistics.

Chen Lyu and Gabriele Pergola. 2024b. Society of Medical Simplifiers. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

G. Harry Mc Laughlin. 1969. SMOG Grading: A new readability formula. *Journal of Reading*, 12(8):639–646. Publisher: [Wiley, International Reading Association].

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

n8n. 2025. n8n: Workflow automation platform (v1.109.0). `https://github.com/n8n-io/n8n/releases/tag/n8n@1.108.0`. Version 1.109.0, released August 25, 2025.

National Cancer Institute. 2025a. NCI Dictionary of Cancer Terms. Accessed: August 23, 2025.

National Cancer Institute. 2025b. NCI Dictionary of Genetics Terms. Accessed: August 23, 2025.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025b. OpenAI models: Open AI API. Latest update: August 2025; archived at `http://web.archive.org/web/20250817212345/https://platform.openai.com/docs/models/gpt-5`.

Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. *Template and guidance for writing a Cochrane Plain language summary*. Cochrane. Archived at https://web.archive.org/web/20250824000540/https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/guidance-writing-cochrane-plain-language-summary.pdf (Accessed 2025-08-23).

Readability. 2019. Readability 0.3.1. Accessed August 2025.

Karen Scholz and Markus Wenzel. 2025. Evaluating readability metrics for German medical text simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

R. J. Senter and E. A. Smith. 1967. Automated readability index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base, Ohio, USA.

Bhuvaneswari Sivagnanam, Rivo Krishnu C H, Princi Chauhan, and Saranya Rajiakodi. 2025. CUTN_Bio at BioLaySumm: Multi-task prompt tuning with external knowledge and readability adaptation for layman summarization. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 269–274, Vienna, Austria. Association for Computational Linguistics.

SpaCy. 2025. SpaCy 3.8.7. Accessed August 2025.

Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan van den Broucke, and on behalf of the HLS-EU Consortium Brand, Helmut. 2015. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health*, 25(6):1053–1058.

The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard. 2025. Clinical Research Glossary. Accessed: August 23, 2025.

Together AI. 2025. gpt-oss-120B API. Archived at https://web.archive.org/web/20250821140846/https://www.together.ai/models/gpt-oss-120b.

Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025. MedReadCtrl: Personalizing medical text generation with readability-controlled instruction learning. *arXiv preprint*. ArXiv:2507.07419 [cs].

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLaySumm task 1: Evaluating GPT models for biomedical lay summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *Preprint*, arXiv:2503.18968.

Washington State Department of Health. 2023. Glossary of immunization and public health terms. Updated: June 2023. Accessed: August 23, 2025.

WebMD. 2022. Asthma Glossary. Medically Reviewed on September 22, 2022. Accessed: August 23, 2025.

Mary Ann Williams. 2025. Guides: Health Literacy Resources: Plain Language Resources. https://guides.hshsl.umaryland.edu/c.php?g=94026&p=7981462. Accessed: August 23, 2025.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *arXiv preprint*. ArXiv:2305.16739 [cs].

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

Wenjun Zhang, Shekhar Chandra, Bevan Koopman, Jason Dowling, and Aaron Nicolson. 2025. AEHRC at BioLaySumm 2025: Leveraging t5 for lay summarisation of radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 171–178, Vienna, Austria. Association for Computational Linguistics.

A System Architecture and Methodology Workflow

Figure A1 illustrates the complete methodology workflow for developing and evaluating the multi-agent system.

A.1 Iterative Architecture

Initial experiments revealed that the baseline configuration considerably outperformed the V_1 architecture in several metrics. This motivated the development of V_2 as an alternative approach to test different architectural strategies and increase result diversity. The V_2 configuration (Figure A2) addresses two key observations from preliminary testing: (1) smaller models struggle to properly utilize tools, performing significantly better with iterative workflows rather than on-demand tool invocation, and (2) the Medical Glossary Tool, while conceptually useful, consumes substantial tokens without proportional quality gains.

The V_2 architecture maintains the same specialized agents as V_1 but reorganizes their interaction pattern. Instead of the Editor invoking the Evaluator as a tool when needed, the Evaluator operates as an independent agent that systematically reviews each draft and provides structured feedback to the Editor. This iterative loop continues for up to 10 cycles or until the Evaluator approves the output based on quality criteria. When the maximum iteration limit is reached without approval, we select the iteration that achieves the highest percentage of linguistic features in the optimal quartile. For each generated text, the PLS Evaluation Tool computes 18 linguistic features and maps them to percentile ranges (P25, P50, P75, P90).

The optimal quartile corresponds to P25 for lower-is-better metrics and P75 for higher-is-better metrics. We calculate what percentage of all features fall within this optimal range and select the iteration with the highest percentage. While metrics like AlignScore might provide more reliable quality assessment, we opted for this percentile-based approach due to computational efficiency constraints, as computing AlignScore for each iteration would significantly extend processing time. The removal of the Medical Glossary Tool reduces token consumption while testing whether explicit medical dictionaries are necessary for quality PLS generation.

This architectural variant allowed us to compare monolithic (Baseline), tool-based (V_1), and iterative (V_2) approaches using identical underlying models and agents, providing insights into how interaction patterns affect generation quality across different model sizes.

A.2 Example V_2 Iterative Refinement Process

This section illustrates the V_2 iterative refinement process using Cochrane review CD002120.PUB4 (“Combined oral contraceptive pill for primary dysmenorrhoea”) as a representative example.

The Information Extraction Agents processed the abstract and produced structured outputs for each required PLS component:

> **Plain Title:** What are the benefits and harms of the combined birth control pill for treating painful periods (primary dysmenorrhoea)?
>
> **Key Messages:**
>
> - For women with painful periods, the combined birth control pill reduces pain more than a placebo. However, it may cause unwanted effects such as irregular bleeding, headaches, and nausea.
> - Taking the pill continuously may be better at reducing pain than taking it in the standard way.
> - Because the studies were short, the long-term effects are not known. It is also unclear if the pill is more effective than other pain medicines due to a lack of studies directly comparing them.
>
> **Background:** Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. When these painful cramps are not caused by another underlying medical condition in the pelvis, it is called ‘primary dysmenorrhoea’. Combined oral contraceptive pills are a common treatment for this pain.
>
> **Methods:** We searched for studies that compared combined oral contraceptive pills with a placebo, other contraceptive pills, or anti-inflammatory drugs for treating painful periods. We then combined the results and rated our confidence in the evidence.
>
> **Results:** This review included 21 randomised controlled trials involving 3723 women. For women with painful menstrual cramps, OCPs reduce pain more effectively than placebo. However, OCPs increase the risk of irregular bleeding. Continuous use may reduce pain more effectively than the standard schedule...
>
> **Limitations:** Our confidence in the evidence is only moderate because there are not enough studies to be certain about all outcomes. The evidence does not cover long-term effects.
>
> **Currency:** The evidence is up to date to March 2023.

The Assembly Agent constructed the initial draft of 617 words, achieving 83.33% of features in the best quartile. The Evaluator identified three

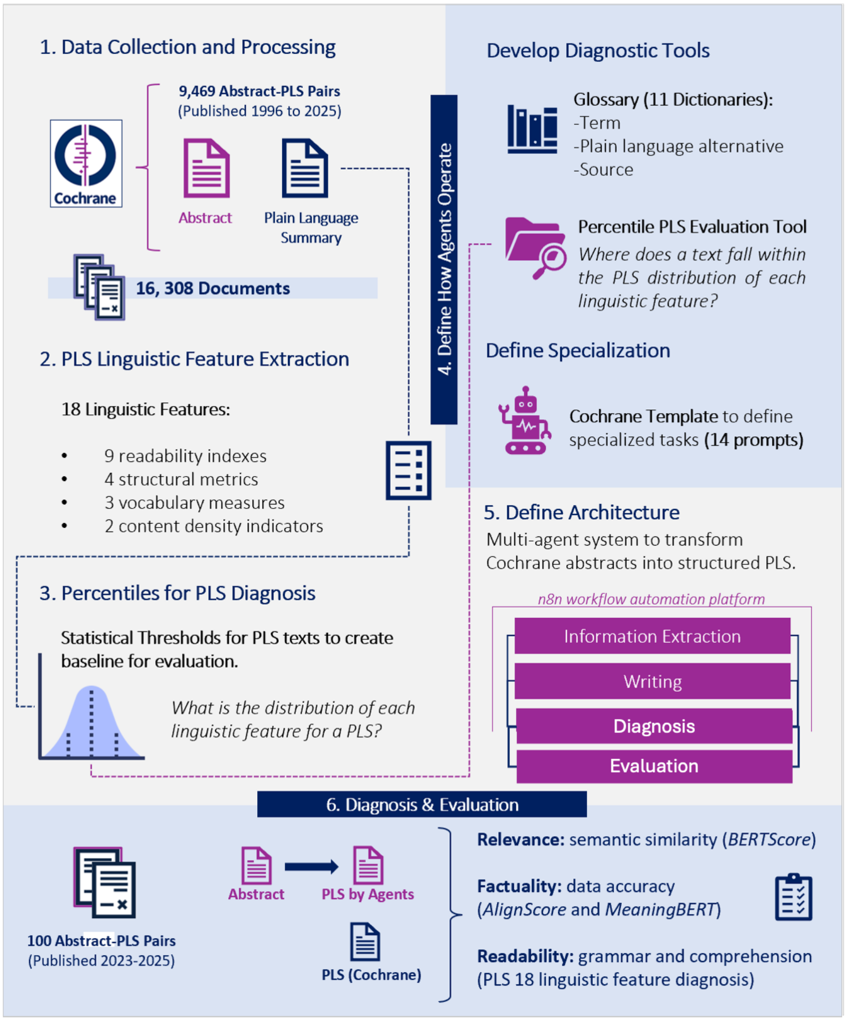
1. Data Collection and Processing
9,469 Abstract-PLS Pairs
(Published 1996 to 2025)
Cochrane
Abstract
Plain Language Summary
16, 308 Documents
2. PLS Linguistic Feature Extraction
18 Linguistic Features:
• 9 readability indexes
• 4 structural metrics
• 3 vocabulary measures
• 2 content density indicators
3. Percentiles for PLS Diagnosis
Statistical Thresholds for PLS texts to create baseline for evaluation.
What is the distribution of each linguistic feature for a PLS?
4. Define How Agents Operate
Develop Diagnostic Tools
Glossary (11 Dictionaries):
-Term
-Plain language alternative
-Source
Percentile PLS Evaluation Tool
Where does a text fall within the PLS distribution of each linguistic feature?
Define Specialization
Cochrane Template to define specialized tasks (14 prompts)
5. Define Architecture
Multi-agent system to transform Cochrane abstracts into structured PLS.
n8n workflow automation platform
Information Extraction
Writing
Diagnosis
Evaluation
6. Diagnosis & Evaluation
100 Abstract-PLS Pairs
(Published 2023-2025)
Abstract
PLS by Agents
PLS (Cochrane)
Relevance: semantic similarity (BERTScore)
Factuality: data accuracy (AlignScore and MeaningBERT)
Readability: grammar and comprehension (PLS 18 linguistic feature diagnosis)

Figure A1: Methodology for developing and evaluating a multi-agent system for automatic generation of PLS from Cochrane medical abstracts.

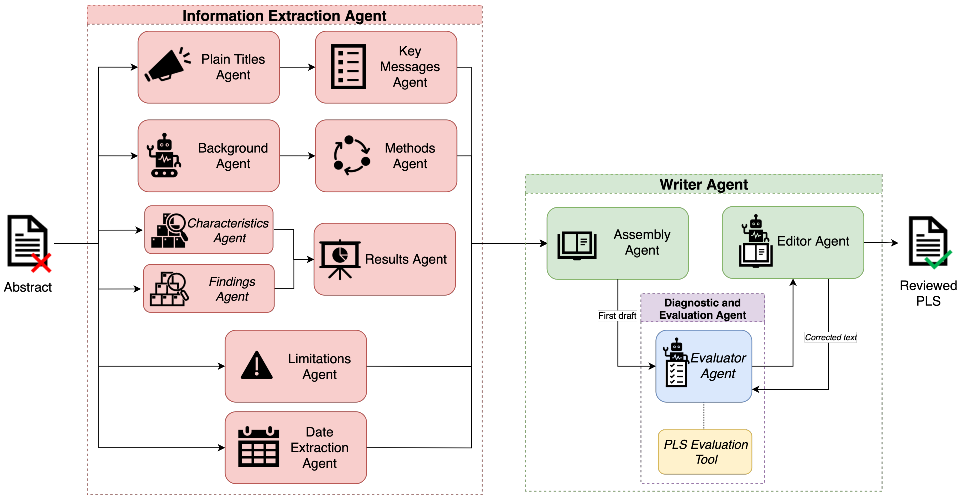

Figure A2: V_2 iterative architecture for PLS generation. Unlike V_1 where the Editor invokes the Evaluator as a tool on-demand, V_2 establishes an independent iterative loop where the Evaluator systematically reviews drafts and provides feedback to the Editor. The process continues for up to 10 iterations or until approval. This architecture removes the Medical Glossary Tool and proves particularly effective for smaller models that struggle with tool invocation.

key issues: (1) a non-question subheading ("Main results: oral contraceptives for painful menstrual cramps"), (2) complex phrasing such as "When these painful cramps are not caused by another underlying medical condition in the pelvis. . . ", and (3) excessive passive voice constructions. Three metrics fell in the P90 range: nouns (158), complex words (241), and long words (165). The draft was marked NOT_APPROVED.

> What are the benefits and harms of the combined birth control pill for treating painful periods. . .
>
> **Key messages** [same as extracted]
>
> **What is primary dysmenorrhoea and what did this review study?**
>
> *What is primary dysmenorrhoea?* Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. It is a common and major cause of pain. When these painful cramps are not caused by another underlying medical condition. . .
>
> *What did we want to find out?* We wanted to find out about the benefits and harms of using combined oral contraceptive pills. . .
>
> **How was the review conducted?** We searched for studies that compared combined oral contraceptive pills with a placebo, other contraceptive pills, or anti-inflammatory drugs. . .
>
> **Main results: oral contraceptives for painful menstrual cramps** For women with painful menstrual cramps, the pill reduces pain more effectively than a placebo. However, the pill comes with unwanted effects. . .
>
> **How reliable is this evidence?** Our confidence in the evidence is only moderate because there are not enough studies to be certain. . .
>
> **How up to date is this evidence?** The evidence is up to date to March 2023.

The Editor incorporated this feedback in Draft 2, reducing word count to 561 and improving the best quartile percentage to 94.44%. The Editor converted the non-question subheading to "What are the main results of the review?", simplified phrasing (e.g., "This is called 'primary dysmenorrhoea' when the pain is not caused by another medical problem" instead of the more complex original wording), and reduced passive voice from 17 to 13 instances. All linguistic metrics moved within P75 or better, with no metrics remaining in the P90 range. The draft was marked PASS.

> What are the benefits and harms of the combined birth control pill for treating painful periods. . .
>
> **Key messages** [same as extracted]
>
> **What is primary dysmenorrhoea?** Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. This is called 'primary dysmenorrhoea' when the pain is not caused by another medical problem in the pelvis. . .
>
> **What did we want to find out?** We wanted to find out about the benefits and harms of using combined oral contraceptive pills. . .
>
> **How did we conduct this review?** We searched for studies that compared combined oral contraceptive pills with a placebo. We found 21 studies, known as randomised controlled trials, with a total of 3723 women. In these studies, researchers randomly put people into one of 2 or more treatment groups. . .
>
> **What are the main results of the review?** For women with painful menstrual cramps, the pill reduces pain more effectively than a placebo. However, the pill comes with unwanted effects. . .
>
> **How reliable is this evidence?** Our confidence in the evidence is only moderate because there are not enough studies to be certain about all health effects. . .
>
> **How up to date is this evidence?** The evidence is up to date to March 2023.

Table A1 quantifies the improvements between iterations.

Metric	**Draft 1**	**Draft 2**
Word count	617	561
FKGL	9.68	9.54
Passive voice	17	13
Nouns	158	142
Complex words (DC)	241	213
Long words	165	147
Best quartile (%)	83.33	**94.44**
P90 metrics	3	0
Decision	FAIL	**PASS**

Table A1: Metric improvements from Draft 1 to Draft 2 for CD002120.PUB4.

B Linguistic Analysis Framework

B.1 Linguistic Features

We computed 20 linguistic features for each document using the Readability (2019) and SpaCy (2025) libraries. Here we describe the 20 features selected for quality assessment (see Table B1 for the percentile thresholds):

1. **Words:** Total word count in the text.
2. **Sentences:** Total sentence count in the text.
3. **Flesch Reading Ease (FRE):** Produces a score where higher values indicate easier readability (Flesch, 1948; Kincaid et al., 1975).
4. **Flesch-Kincaid Grade Level (FKGL):** Estimates the U.S. school grade level needed to comprehend the text (Flesch, 1948; Kincaid et al., 1975).
5. **Gunning Fog Index (GFI):** Estimates the number of years of formal education needed to understand the text (Gunning, 1952).
6. **SMOG Readability Formula (SMOG):** Estimates readability by counting polysyllabic words (Mc Laughlin, 1969).
7. **Dale-Chall Readability Score (DCRS):** Assesses readability by comparing text words against a list of familiar words (Chall and Dale, 1995).

8. **Coleman-Liau Index (CLI):** Measures readability based on letter and word counts per sentence (Coleman and Liau, 1975).

9. **Automated Readability Index (ARI):** Computes readability using characters, words, and sentences (Senter and Smith, 1967).

10. **LIX:** Calculates readability by analyzing the proportion of long words in the text (Anderson, 1983).

11. **RIX:** Computes readability from the number of long words per sentence (Anderson, 1983).

12. **Words per Sentence:** Average number of words per sentence, computed as total words divided by total sentences.

13. **Passive Voice:** Frequency of passive voice constructions, determined via verb forms tagged as VBN (e.g., "was given").

14. **Active Voice:** Frequency of active voice constructions, counted as verbs not tagged as VBN (e.g., "ran", "decided").

15. **Nominalization:** Count of nominalizations, where verbs or adjectives are transformed into nouns (e.g., "development" from "develop").

16. **Complex Words (DC):** Count of complex words according to the Dale-Chall method (unknown polysyllabic words from a list of basic words).

17. **Long Words:** Count of words exceeding 7 letters in length.

18. **Complex Words:** Count of words with three or more syllables (e.g., "inconceivable").

19. **Pronouns:** Count of pronouns in the text, determined by tokens with the part-of-speech PRON (e.g., "him", "she").

20. **Nouns:** Count of nouns in the text, determined by tokens with the part-of-speech NOUN (e.g., "book", "concept").

B.2 Percentile-Based Reference Thresholds

We derived these thresholds from the training corpus of 6,754 Plain Language Summaries. The labeling system adapts to metric directionality as described in Section 3.2.2: for lower-is-better metrics (↓), the labels correspond to actual percentiles; for higher-is-better metrics (↑), the same labels represent inverse percentiles. This ensures P25 and P75 consistently identify what we could consider the "best" quartile across all metrics (having in mind that not being in this quartile does not necessarily mean that the text is bad/not plain).

B.3 Example PLS Evaluation Tool Output

The PLS Evaluation Tool generates structured text output (both JSON and human-readable format) that is directly provided to the LLM agents as input. Table B2 illustrates the tool's analysis of an abstract-PLS pair from the same publication. The original abstract deviates significantly from typical PLS patterns, while its professionally written PLS achieves better conformity. For each metric deviating from typical patterns (P90 or beyond), the tool automatically generates specific feedback suggesting reduction to median values. The actual tool produces formatted text, but we present it here in tabular form for clarity.

C Medical Glossary Sources

We compiled medical glossaries from eleven authoritative sources to support plain language translation (Table 2 in the Methodology section presents the source distribution and term counts). Most of these resources are compiled in the University of Maryland's Williams (2025), which provides comprehensive plain language resources for health communication. Each source provides specialized terminology translations for different healthcare domains:

- **NCI-C** (National Cancer Institute, 2025a): National Cancer Institute's comprehensive cancer dictionary covering types, treatments, procedures, and side effects for patient education.

- **NCI-D** (National Cancer Institute, 2025a): National Cancer Institute's drug database with chemotherapy agents, targeted therapies, and immunotherapy medications.

- **CDC-T** (Centers for Disease Control and Prevention, 2011): CDC's thesaurus providing plain language alternatives for epidemiological and public health terminology.

- **ADA-D** (American Diabetes Association, 2024): American Diabetes Association's glossary covering diabetes types, management, complications, and monitoring terms.

Feature	P25/P75*	P50	P75/P25*	P90/P10*
READABILITY INDICES				
FRE ↑	≥ 48.17	≥ 40.48	≥ 32.68	< 25.27
FKGL ↓	≤ 11.77	≤ 13.16	≤ 14.59	> 16.05
GFI ↓	≤ 16.15	≤ 17.79	≤ 19.39	> 21.05
SMOG ↓	≤ 10.68	≤ 12.11	≤ 13.58	> 14.87
DCRS ↓	≤ 7.19	≤ 7.65	≤ 8.16	> 8.63
CLI ↓	≤ 11.36	≤ 12.66	≤ 13.96	> 15.16
ARI ↓	≤ 12.02	≤ 13.60	≤ 15.29	> 16.95
LIX ↓	≤ 50.25	≤ 54.32	≤ 58.40	> 62.38
RIX ↓	≤ 6.04	≤ 7.04	≤ 8.14	> 9.36
STRUCTURAL COMPLEXITY				
Words/Sent. ↓	≤ 19.81	≤ 22.13	≤ 24.76	> 27.60
Passive Voice ↓	≤ 9	≤ 13	≤ 18	> 23
Active Voice ↑	≥ 41	≥ 29	≥ 20	< 14
Nominalization ↓	≤ 8	≤ 13	≤ 20	> 27
VOCABULARY COMPLEXITY				
Complex Words (DC) ↓	≤ 115	≤ 160	≤ 213	> 277
Long Words ↓	≤ 88	≤ 122	≤ 164	> 208
Complex Words ↓	≤ 60	≤ 84	≤ 115	> 145
CONTENT DENSITY				
Pronouns ↑	≥ 21	≥ 13	≥ 8	< 5
Nouns ↓	≤ 83	≤ 116	≤ 157	> 202

Table B1: Percentile-based reference ranges for 18 linguistic features derived from 6,754 PLS texts. Column headers show actual percentiles for lower-is-better metrics (↓) and with asterisk (*) for higher-is-better metrics (↑). For example, P25/P75* means 25th percentile for ↓ metrics and 75th percentile for ↑ metrics, both representing the best quartile.

- **NCI-G** (National Cancer Institute, 2025b): National Cancer Institute's genetics dictionary explaining hereditary conditions, genetic testing, and molecular biology concepts.
- **UIowa** (Human Subjects Office, University of Iowa, 2021): University of Iowa's general medical term translations designed for informed consent documents and patient communication.
- **MRCT** (The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard, 2025): Multi-Regional Clinical Trials Center's glossary for clinical research protocols, trial phases, and regulatory terminology.
- **WA-PH** (Washington State Department of Health, 2023): Washington State's glossary focused on vaccine types, immunization schedules, and disease prevention terminology.
- **WebMD-A** (WebMD, 2022): WebMD's asthma-specific dictionary covering triggers, medications, devices, and respiratory symptoms.
- **CCIIO** (Center for Consumer Information & Insurance Oversight, 2024): Health insurance glossary explaining coverage terms, benefits, deductibles, and healthcare plan types.
- **Cochrane** (Pitcher et al., 2022): Cochrane's guideline for writing systematic review summaries with standardized plain language templates.

D Complete Linguistic Feature Analysis

Table D1 presents a comprehensive analysis of all 20 linguistic features across all tested models, organized into three categories: Structure (word/sentence metrics), Style (voice and pronoun usage), and Readability & Complexity (traditional readability indices and vocabulary measures).

(a) Original abstract: 1019 words				**(b) Corresponding PLS: 672 words**			
Metric	**Value**	**Target**	**Rating**	**Metric**	**Value**	**Target**	**Rating**
Within typical ranges				*Within typical ranges*			
FRE	38.55	≥ 40.48	P25	Words/Sent.	19.76	≤ 22.13	P25
Active Voice	83	≥ 29	P75	FKGL	11.68	≤ 13.16	P25
Pronouns	30	≥ 13	P75	GFI	15.47	≤ 17.79	P25
CLI	11.95	≤ 12.66	P50	Active Voice	52	≥ 29	P75
				Pronouns	36	≥ 13	P75
				CLI	12.75	≤ 13.96	P75
				FRE	46.53	≥ 40.48	P50
				ARI	12.52	≤ 13.60	P50
				LIX	52.06	≤ 54.32	P50
				RIX	6.38	≤ 7.04	P50
				DCRS	7.30	≤ 7.65	P50
Deviating from typical patterns				*Deviating from typical patterns*			
FKGL	15.11	≤ 13.16	P90	Nominalization	22	≤ 13	P90
ARI	16.17	≤ 13.60	P90	Nouns	198	≤ 116	P90
Words/Sent.	29.11	≤ 22.13	Beyond P90	SMOG	14.18	≤ 12.11	P90
Passive Voice	42	≤ 13	Beyond P90	Complex Words (DC)	269	≤ 160	P90
Nominalization	44	≤ 13	Beyond P90	Complex Words	127	≤ 84	P90
Nouns	326	≤ 116	Beyond P90	Passive Voice	31	≤ 13	Beyond P90
GFI	22.72	≤ 17.79	Beyond P90	Long Words	217	≤ 122	Beyond P90
LIX	64.15	≤ 54.32	Beyond P90				
RIX	10.20	≤ 7.04	Beyond P90				
SMOG	18.97	≤ 12.11	Beyond P90				
DCRS	9.30	≤ 7.65	Beyond P90				
Complex Words (DC)	507	≤ 160	Beyond P90				
Complex Words	282	≤ 84	Beyond P90				
Long Words	357	≤ 122	Beyond P90				

Table B2: Example of PLS Evaluation Tool output comparing an abstract-PLS pair. The original abstract (a) exceeds the word limit and shows poor conformity with 14 of 18 metrics deviating from typical patterns. Its corresponding PLS (b) meets the word limit and achieves moderate conformity with only 7 metrics deviating. The tool automatically generates feedback suggesting median target values for all metrics at P90 or beyond.

Model	**Approach**	**Structure**			**Lexical Features**							
		Words	**Sent.**	**WPS**	**Pass.**	**Act.**	**Pron.**	**Nom.**	**Nouns**	**CW-DC**	**CW**	**LW**
Original Abstract		868	33.82	26.04	24.56	49.85	20.42	43.35	278.62	411.69	215.47	285.83
Reference PLS (human)		655	33.08	20.08	19.64	57.95	35.01	25.02	187.50	263.36	133.12	197.00
Gemini 2.5 Pro	*Baseline*	661	41.64	16.05	14.75	66.43	45.52	19.05	171.80	240.15	98.59	166.07
	Agentic (V1)	591	36.10	16.53	11.49	63.20	40.19	17.68	159.35	208.00	86.35	149.57
	Agentic (V2)	567	32.73	17.48	12.05	59.57	35.07	17.92	153.56	207.12	92.02	152.80
GPT-5	*Baseline*	866	48.69	17.90	18.59	90.29	46.74	22.09	249.45	321.40	136.71	235.15
	Agentic (V1)	879	45.60	19.49	21.67	92.90	44.73	25.89	278.74	375.02	160.88	253.35
GPT-OSS-120B	*Baseline*	623	32.47	19.34	14.17	63.93	35.52	19.92	160.30	229.30	109.29	165.78
	Agentic (V1)	743	37.70	19.80	18.73	75.59	39.10	25.68	218.46	330.97	160.50	221.94
	Agentic (V2)	623	31.62	19.80	14.37	64.15	33.24	21.29	177.80	266.55	130.59	182.99
GPT-OSS-20B	*Baseline*	541	27.52	19.31	12.18	57.55	32.31	16.08	141.26	193.89	87.96	141.78
	Agentic (V2)	632	33.41	18.88	15.47	65.61	32.76	20.92	179.55	275.15	131.01	183.65
Llama 3.3 70B	*Baseline*	477	21.65	22.23	9.51	44.43	29.26	19.56	130.86	173.90	95.13	141.06
	Agentic (V2)	497	25.26	19.85	10.89	46.97	23.49	19.93	148.66	207.16	112.08	165.90
	Agentic (V2 + Gemini 2.5 Flash)	514	29.37	17.62	9.57	54.74	32.64	17.86	143.63	194.28	92.17	148.95
Llama 3.2 3B	*Baseline*	477	31.40	15.28	11.94	38.27	25.36	19.09	127.76	185.00	98.34	145.72
	Agentic (V2)	536	27.83	19.79	15.73	40.84	17.42	25.17	162.41	239.07	130.42	187.85

Table D1: Linguistic feature analysis (Structure and Lexical Features) across all tested models and approaches. Gray shading highlights original abstract and human reference baseline. Abbreviations: Words Per Sentence (WPS), Passive voice (Pass.), Active voice (Act.), Pronouns (Pron.), Nominalization (Nom.), Complex Words Dale-Chall (CW-DC), Complex Words (CW), Long Words (LW). Averages computed across 100 test samples.