No Reader Left Behind: Multi-Agent Summaries Everyone Can Understand

Anonymous ACL submission

Abstract

The Plain Writing Act in the United States requires government documents to be accessible in clear and simple language that the general public can easily understand, yet existing summarization systems struggle to address diverse linguistic and cognitive barriers among general readers. We present NRLB (No Reader Left Behind), a multi-agent framework for plain language summarization that simulates three representative reader groups: elementary school student readers, non-native readers, and readers with attention deficits. NRLB combines template-based planning with iterative, reader-oriented refinement, enabling systematic detection and resolution of difficult terms, missing contexts, and confusing sentences. Evaluations across multiple datasets demonstrate consistent improvements in readability while preserving factual accuracy. Human evaluation further validates NRLB's impact, with annotator preference rates ranging from 55% to 76%, highlighting NRLB's potential to produce plain language summaries that are both faithful to the source and broadly accessible to the general public.

1 Introduction

During the COVID-19 pandemic, gaps in scientific communication deepened public confusion. Lower science literacy correlated with greater susceptibility to misinformation and lower compliance, whereas a mere 1% increase in public science literacy led to a 14.2% rise in civic engagement with public health measures (Qin et al., 2024). These results indicate that even modest gains in science literacy can directly impact public behavior, highlighting the importance of clear and accessible science communication for effective public health responses (Xu et al., 2015). In response, the U.S. government enacted the Plain Writing Act in 2010, mandating that all official communication be understandable to the general public. Federal guidelines recommend a 6th–8th grade reading level, concise and active sentences, annotated or simplified technical terms, and full accessibility across web and mobile platforms. Agencies are also required to have a dedicated section in plain language on their landing page and to provide a public feedback mechanism for unclear documents[1]. The guidelines specifically emphasize the need to accommodate readers with limited literacy, non-native English proficiency, or cognitive and visual impairments (Gooding, 2022). These principles are typically implemented through concrete strategies such as sentence reduction, vocabulary simplification, and definitions for specialized terms, as illustrated in Figure 1.

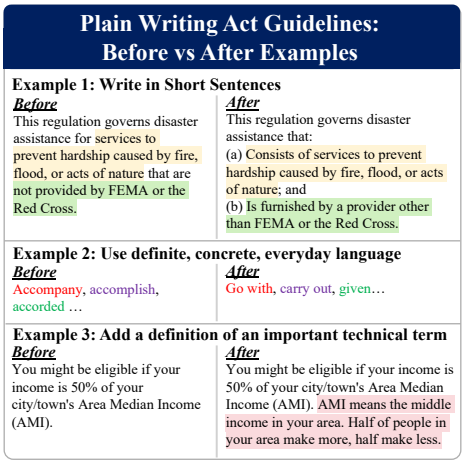

Figure 1: Examples of revision strategies from the U.S. Plain Writing Act: sentence shortening, concrete wording, and defining terms.

However, existing automated summarization and simplification systems often fall short of the Plain Writing Act principles. First, they typically overlook key reader-specific factors such as background knowledge, attention span, and syntactic processing, leading to outputs that retain technical jargon and complex sentence structures. While some mod-

[1] https://www.plainlanguage.gov/law/

els succeed in surface-level simplification, they rarely enhance comprehension for general audiences. Second, many prior approaches generate multiple summary versions customized for specific subgroups (e.g., K–12 students or non-native adults). This strategy contradicts the Act's core principle of clarity for all readers (Mo and Hu, 2024; Xu et al., 2015). These limitations point to the need for a unified framework that enables broader accessibility through adaptive simplification.

To address this gap, we introduce NRLB (No Reader Left Behind), an automated framework for plain language generation that accounts for commonly overlooked reading barriers. Grounded in prior research on literacy and cognitive processing (Smith et al., 2021; Tighe and Schatschneider, 2016; Shero et al., 2021), NRLB simulates three representative reader types: an Elementary School Student reader, a Non-Native Reader, and an Attention-Deficit Reader. Each agent identifies distinct comprehension challenges, to enable targeted revisions that enhance accessibility across reader populations.

The NRLB framework comprises two core modules. **Module 1** - Content Planning and Drafting begins with a Planner Agent that classifies the input document and selects a Domain Expert Agent, who produces a template-based initial summary reflecting the document's key content. **Module 2** - Feedback-Guided Simplification refines this draft through a structured loop of simulated feedback and expert revision. Reader Agents flag the summary with segments of lexical complexity, missing context, or syntactic ambiguity. The Checklist Agent aggregates and prioritizes these issues, which the Domain Expert Agent then uses to propose targeted edits. Finally, the Editor Agent applies these revisions contextually to produce the final plain language summary, evaluated via metrics for Readability, Relevance, and Factuality.

Our key contributions are as follows: (1) We present NRLB, a unified plain language generation framework that addresses real-world reading challenges by simulating three representative reader types. (2) We propose a role-based multi-agent architecture that enables structured, iterative revision based on simulated reader feedback. (3) We demonstrate, through both automatic metrics and human evaluation, that NRLB substantially improves comprehension, relevance, and factual consistency over existing methods.

2 Related Work

2.1 Plain Language Summarization

Plain language summarization aims to simplify sentence structure while preserving key content and meaning. Prior work has concentrated on lexical or syntactic rewriting, often using Simple Wikipedia (Xu et al., 2015), readability-controlled fine-tuning (Luo et al., 2022), or task-specific evaluation metrics (Urlana et al., 2023). These approaches typically target narrow audiences or specific domains, often focusing on surface-level readability at the expense of content fidelity.

In contrast, NRLB generates plain language summaries of full-length documents by integrating structured feedback from simulated readers with diverse comprehension needs. Rather than simple text reduction, it addresses barriers such as background gaps, complex wording, and attention limits while ensuring accuracy and readability.

2.2 Text Simplification via Multi-Agent Collaboration

Recent work has explored multi-agent architectures where role-specific agents collaborate to overcome the limitations of single large language models (Du et al., 2024; Chen et al., 2023). In text simplification, **AgentSimp** (Fang et al., 2025) adopts a stage-wise pipeline in which agents such as Metaphor Analysis and Article Logic Analyst operate sequentially. While effective for simplifying news and web content, it lacks iterative feedback and adaptability and often produces over-simplified outputs.

ExpertEase (Mo and Hu, 2024) assigns teacher, student, and expert roles to generate grade-specific summaries for different educational levels. However, its audience-segregated design diverges from the goal of the Plain Writing Act, which seeks universally understandable documents.

In contrast, **NRLB** introduces an iterative feedback loop across simulated reader types, combining reader simulation with expert-guided revision to promote inclusivity and refinement in plain language summarization.

2.3 Reader-Centered Approaches

Text simplification has traditionally aimed to support underserved readers, such as individuals with dyslexia, non-native speakers, and those with cognitive challenges. For example, (Rello et al., 2013) showed that replacing unfamiliar words with shorter and more frequent alternatives improved

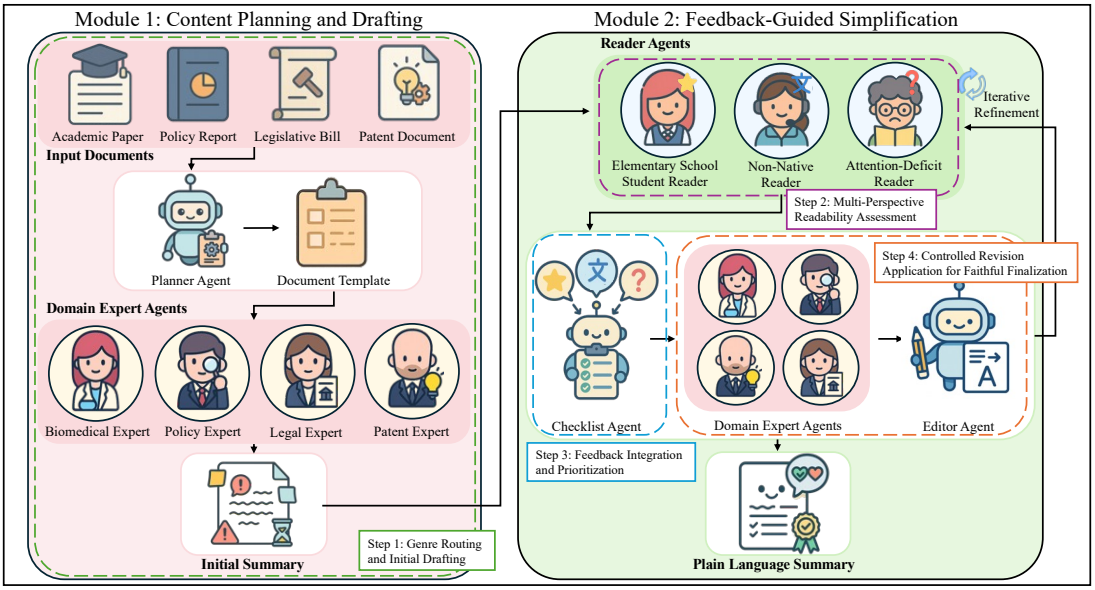

Figure 2: Overview of the NRLB framework. Module 1 generates an initial draft via content planning and domain expertise, while Module 2 applies reader feedback and expert edits to produce a plain language summary.

reading speed and comprehension among adults with dyslexia. Similarly, (Wilkens and Todirascu, 2020) found that vocabulary-based simplification strategies were effective for readers with cognitive challenges in multilingual settings. However, much of this work has focused on lexical simplification, often overlooking broader barriers such as limited background knowledge or syntactic complexity.

Extending this line of research, we focus on three underrepresented reader groups: elementary school students, non-native readers, and readers with attention difficulties. Each group exhibits distinct cognitive characteristics and faces challenges such as unfamiliar vocabulary, missing contexts, or complex sentence structures. While these issues have been noted in prior work (Xu et al., 2015; Dadu et al., 2021), few studies have addressed them within an integrated framework. NRLB fills this gap by simulating feedback from Reader Agents representing these profiles. By incorporating their perspectives into the simplification loop, the system adopts a genuinely reader-centered approach. The design and integration of these agents are described in Section 3.2.

3 Methodology

As illustrated in Figure 2, the proposed NRLB (No Reader Left Behind) framework generates plain language summaries for general audiences through two sequential modules. **Module 1. Content Planning and Drafting** begins with the Planner Agent classifying the genre of the input document. Based on this classification, the system selects an appropriate Domain Expert Agent and assigns a predefined template corresponding to the identified document type. The Domain Expert Agent then completes the template by following genre-specific guidelines, producing an initial summary that emphasizes informational completeness and logical flow. **Module 2. Feedback-Guided Simplification** delivers the initial summary to three simulated Reader Agents: Elementary School Student Reader, Non-Native Reader, and Attention-Deficit Reader. These agents independently identify readability issues in three categories including unknown terms, missing contexts, and confusing sentences. The Checklist Agent integrates and prioritizes the feedback based on frequency and severity. The resulting checklist is returned to the Domain Expert Agent, who drafts specific revision suggestions. The Editor Agent then applies these suggestions to produce a revised summary. This feedback and revision process is repeated up to N times, following the draft, review, and revision cycle outlined in the U.S. Plain Writing Act (2010). For additional implementation details, please refer to Appendix A.

3.1 Genre Routing and Initial Drafting

The NRLB pipeline begins with genre classification. A LLM-based classifier predicts the document type and assigns the corresponding Domain Expert Agent. For example, a policy report is routed to a

policy expert. Based on the genre label, the Planner Agent selects a slot-based summary template aligned with standard writing conventions for that document type. Each template consists of labeled sections with recommended sentence lengths. For instance, a policy report from the U.S. Government Accountability Office (GAO) typically includes "Why GAO Did This Study," "What GAO Found," and "What GAO Recommends." A list of genre-specific templates is provided in Appendix A.1.

The Domain Expert Agent, implemented as an LLM guided by a domain-specific persona, completes the template using example summaries and writing instructions (Appendix A.2). The agent aims to ensure informational completeness and logical flow while adhering to length constraints. For example, in biomedical documents, the Methods slot briefly summarizes the experimental design, while the Results slot highlights key findings. Although the initial summary effectively conveys the core content, it may still include barriers to comprehension such as technical jargon, complex syntax, or missing background information. These issues are systematically addressed in the next module.

3.2 Multi-Perspective Readability Assessment

To cover a broad spectrum of common comprehension barriers faced by the general public, we identified three representative reader profiles in Module 2. Such profiles were grounded in prior research on literacy and cognitive processing. These include the Elementary School Student Reader Agent, the Non-Native Reader Agent, and the Attention-Deficit Reader Agent, each representing a distinct reader population. All three agents evaluate the same summary according to the linguistic and cognitive characteristics of the group they simulate. They are instantiated via LLM using prompts that reflect the reader's abilities at three levels: word-level (e.g., limited vocabulary), knowledge-level (e.g., lack of background context), and sentence-level (e.g., difficulty parsing complex syntax). Each agent is instructed to detect readability issues across all three predefined categories: unknown terms, missing contexts, and confusing sentences. This simulation-based approach enables consistent and reproducible feedback across diverse reader perspectives without requiring human annotators. Full prompt specifications are provided in Appendix A.3.

- **The Elementary School Student Reader Agent** simulates the perspective of a fourth-grade student with fluent command of approximately 3,000 basic English words (Laufer and Ravenhorst-Kalovski, 2010). It flags unfamiliar or technical terms beyond this range as unknown terms. Institutional references or scientific concepts that require background knowledge are classified as missing contexts. Sentences longer than 15 words with multiple subordinate clauses are marked as confusing sentences. Typical feedback includes phrases such as "too many difficult words," "background explanation is needed," and "the sentence is too long."

- **The Non-Native Reader Agent** simulates an adult reader familiar with high-frequency vocabulary but sensitive to low-frequency academic terms, idioms, and cultural expressions (Ha, 2022). It flags rare or idiomatic phrases as unknown terms, and identifies culturally specific institutions or historical references lacking explanation as missing contexts. Sentences with garden-path constructions, reduced relative clauses, or repeated passive voice are labeled as confusing sentences. Common feedback includes statements like "I don't know this word," "the cultural context is missing," and "the sentence structure is too complicated."

- **The Attention-Deficit Reader Agent** reflects the cognitive profile of individuals with limited working memory and reduced attention span (Jacobson et al., 2011). It marks visually or phonetically complex words, compounds, and low-frequency vocabulary as unknown terms. Concepts introduced without sufficient background are flagged as missing contexts. Sentences exceeding 15 words or containing multiple relative clauses are categorized as confusing sentences. Representative feedback includes comments such as "the word is hard to decode" and "the long sentence makes it hard to stay focused."

Each Reader Agent provides structured feedback, which is integrated and prioritized in the next stage. This consolidated view reveals comprehension barriers general readers may face. Although we focus on three representative reader types, NRLB is designed to easily plug in additional Reader Agents, ensuring scalability to broader comprehension needs.

3.3 Feedback Integration and Prioritization

The Checklist Agent collects feedback from the Reader Agents, each classifying issues as unknown terms, missing contexts, or confusing sentences. Each issue's priority equals the number of agents that flagged it. Issues flagged by all three agents receive top priority. When multiple items share the same frequency, the agent applies the Automated Readability Index (ARI) to assess sentence complexity (Senter and Smith, 1967). The ARI estimates U.S. grade-level difficulty by combining average characters per word and words per sentence. For each category, the Checklist Agent selects up to three highest-difficulty items and compiles them into a unified checklist. This refined list is forwarded to the Domain Expert Agent to guide targeted revisions.

3.4 Controlled Revision Application for Faithful Finalization

The Domain Expert Agent generates revision proposals based on the checklist received from the Checklist Agent (Appendix A.2). As the same domain-specific expert assigned in Module 1, it ensures continuity between initial drafting and revision. It replaces unknown terms with simple synonyms or five-word definitions, adds missing context with up to fifteen-word explanations from the source, rewrites confusing sentences into shorter active voice, and marks items as "insufficient information" if no source support exists.

The Editor Agent applies these revisions while maintaining sentence-level coherence and contextual flow. When edits involve splitting or replacing parts of the same sentence, the agent determines a conflict-free order of application. Revisions are only applied when the original sentence exists in the current summary, and replacements are executed sequentially to preserve consistency. The Editor Agent integrates Reader feedback with expert-suggested revisions to enable coherent multi-perspective refinement. This revision process can be repeated up to N times; in our experiments, we used two iterations by default. This collaborative approach improves readability while preserving factuality. Ablation results in Appendix B.1 show that omitting the Editor Agent reduces the factual consistency of the final summary.

4 Experiments

4.1 Datasets

We evaluate NRLB on four representative datasets across diverse domains. PLOS consists of biomedical academic papers, GovReport comprises policy reports from U.S. government agencies, BillSum includes legislative bills with expert-authored summaries, and BigPatent covers patent documents with rich technical language. From each dataset, we randomly sampled 500 test examples to ensure both statistical robustness and computational efficiency, following recent best practices in long-document summarization research (Xu et al., 2025). All datasets include human-written reference summaries, and further dataset-specific details are provided in Appendix C.1.

4.2 Models

We evaluate NRLB using three foundation models with distinct characteristics. Llama-3.1-8B-Instruct, an open-source model licensed under Apache 2.0, serves as a reproducible baseline (Dubey et al., 2024). GPT-4o represents OpenAI's latest commercial model with multimodal capabilities (Hurst et al., 2024). Qwen3-8B (Reasoning mode) is included for its strong performance on complex reasoning tasks. Full configuration details are provided in Appendix C.2 (Yang et al., 2025).

4.3 Metrics

We adopt the three core evaluation dimensions defined in the Plain Summarization Shared Task: Relevance, Readability, and Factuality (Goldsack et al., 2024). From the available metrics, we select seven that are most aligned with our plain language generation task: ROUGE1, BERTScore, FKGL, DCRS, CLI, LENS, and SummaC.

Relevance is measured using ROUGE1, which captures unigram overlap between summaries and references (Liu and Liu, 2008), and BERTScore, which quantifies semantic similarity using contextual embeddings (Zhang et al., 2019). Readability is assessed through traditional formula based metrics (FKGL, DCRS, and CLI) that reflect sentence and word complexity (Tanprasert and Kauchak, 2021; Dale and Chall, 1948; Coleman and Liau, 1975), complemented by LENS, a learned metric that better captures qualitative improvements from lexical and structural simplification (Maddela et al., 2022). Factuality is evaluated using SummaC, which applies natural language inference to

		PLOS			GovReport			BillSum			BigPatent		
		GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3
ROUGE-1	Initial Summary	**46.85**	**47.27**	**41.62**	**37.93**	**46.93**	**46.19**	**46.70**	**45.81**	**41.08**	**41.87**	**44.94**	**34.21**
	Round 1	45.12	44.43	37.55	36.35	45.33	41.41	43.28	42.66	37.08	37.74	40.09	30.55
	Round 2	42.77	41.64	36.03	34.73	44.19	40.04	40.88	39.70	34.62	35.40	36.87	29.18
BERTScore	Initial Summary	**86.91**	**86.51**	**85.78**	**86.14**	**86.30**	**85.59**	**87.07**	**86.83**	**85.61**	**86.52**	**86.68**	**85.55**
	Round 1	86.61	86.00	85.33	85.89	85.84	85.04	86.43	86.20	85.03	85.82	85.73	84.89
	Round 2	86.29	85.43	84.96	85.63	85.48	84.74	85.99	85.59	84.62	85.44	85.11	84.50
FKGL ↓	Initial Summary	17.73	18.74	19.12	17.79	19.95	19.81	17.29	18.04	18.07	18.60	21.13	20.88
	Round 1	13.37	16.55	14.12	13.92	17.51	15.51	12.64	15.66	13.08	12.57	16.88	14.48
	Round 2	**10.94**	**15.21**	**12.46**	**11.58**	**15.98**	**13.72**	**10.39**	**14.60**	**11.16**	**10.27**	**14.62**	**12.38**
DCRS ↓	Initial Summary	14.25	13.69	15.04	13.33	13.43	14.24	13.11	12.77	14.14	13.76	13.36	15.10
	Round 1	13.21	12.67	13.61	12.49	12.51	13.31	12.06	11.88	12.68	12.31	12.14	13.32
	Round 2	**12.56**	**12.06**	**13.15**	**11.96**	**11.96**	**12.90**	**11.50**	**11.40**	**12.24**	**11.70**	**11.47**	**12.74**
CLI ↓	Initial Summary	18.72	17.05	20.86	17.77	16.60	19.38	16.56	15.13	17.38	18.32	16.45	21.52
	Round 1	15.86	14.54	16.54	15.71	15.09	17.06	14.07	13.59	14.22	14.69	13.76	16.29
	Round 2	**13.96**	**13.20**	**15.14**	**14.20**	**14.23**	**16.01**	**12.56**	**12.62**	**13.09**	**12.84**	**12.42**	**14.71**
LENS	Initial Summary	65.38	64.80	62.49	60.28	55.37	56.04	60.91	58.38	56.29	59.62	55.28	55.86
	Round 1	71.35	**67.21**	71.62	66.32	59.20	63.37	68.20	62.58	66.68	70.51	61.08	69.34
	Round 2	**74.67**	66.68	**72.60**	**70.00**	**60.60**	**64.61**	**71.63**	**62.68**	**68.30**	**74.44**	**62.62**	**71.56**
SummaC	Initial Summary	55.48	49.67	44.48	42.36	43.63	38.59	34.41	33.48	30.69	53.19	54.73	42.87
	Round 1	67.12	49.17	48.88	53.14	45.70	45.38	39.41	34.51	36.02	69.53	57.95	49.87
	Round 2	**72.93**	**50.86**	**53.81**	**61.27**	**47.76**	**49.00**	**41.56**	**34.55**	**38.50**	**73.49**	**58.94**	**54.58**

Table 1: Performance of three models (GPT-4o, Llama-3.1-8B-Instruct, Qwen3-8B) on four benchmarks across initial and two refinement stages; best scores are bolded. ↓ indicates that lower is better for readability metrics.

assess consistency between the source document and summary (Laban et al., 2022). Detailed descriptions are in Appendix C.3.

5 Results

5.1 Improving Readability via Iterative Revision

Table 1 summarizes the effects of two rounds of revision (Initial, Round 1, Round 2) using the NRLB across four datasets and three models. Iterative refinement consistently improved readability across all datasets. In contrast, ROUGE-1 and BERTScore declined slightly due to reduced lexical overlap, reflecting a common tradeoff in plain summarization where clarity often reduces surface similarity (Dreyer et al., 2021; Goldsack et al., 2024). Meanwhile, factuality improved across all settings, with SummaC scores increasing from 55.48 to 72.93 on PLOS and from 53.19 to 73.49 on BigPatent, confirming that iterative revision enhanced both logical consistency and source alignment.

Among models, GPT-4o showed the best overall balance, maintaining strong initial relevance and achieving the highest readability and factuality by Round 2. Llama-3.1-8B-Instruct retained high ROUGE but made modest gains elsewhere. Qwen3-8B, though initially less consistent, achieved the largest improvements after revision, likely due to its reasoning-oriented architecture. While Round 3 results (Appendix D) brought minor further improvements in readability and factuality, relevance dropped by up to 10 points compared to the initial summary, validating Round 2 as the optimal stopping point.

5.2 Ablation Study

We evaluated the effect of using one, two, or three Reader Agents. As shown in Table 2, all configurations were evaluated under the same two-round revision pipeline using four readability metrics. Even with a single agent, readability showed some improvement; however, using two agents led to further reductions in FKGL and DCRS and an increase in LENS, resulting in clearer and more accessible summaries. For instance, on GovReport, FKGL improved from 17.03 to 16.09, and LENS rose from 58.20 to 60.12.

With three Reader Agents, FKGL decreased further and LENS reached its highest scores on PLOS and BigPatent, indicating the most consistent improvements. In contrast, BillSum achieved near-optimal readability with two agents, and the third added only marginal gains. These findings suggest that two or more Reader Agents improve readability, with the three-agent setup performing best overall. We therefore adopt the three-agent configuration as default. Additional results with four and five agents are reported in Appendix B.2.

Method	PLOS				GovReport				BillSum				BigPatent			
	FKGL ↓	DCRS ↓	CLI ↓	LENS	FKGL ↓	DCRS ↓	CLI ↓	LENS	FKGL ↓	DCRS ↓	CLI ↓	LENS	FKGL ↓	DCRS ↓	CLI ↓	LENS
All	15.40	12.06	13.19	**67.64**	**16.09**	**12.01**	**14.29**	59.08	15.00	11.68	13.09	62.63	**14.68**	**11.51**	**12.93**	**64.40**
w/o ele	15.73	12.19	13.45	66.18	16.56	12.05	14.68	59.94	**14.48**	**11.35**	**12.52**	62.26	15.23	11.67	12.79	61.90
w/o non	**15.20**	**12.01**	13.10	66.36	16.19	**12.01**	14.41	**60.12**	14.52	11.47	12.60	62.10	16.83	11.91	13.31	59.82
w/o att	15.40	12.06	13.21	67.31	16.34	12.06	14.51	59.27	14.49	11.42	12.58	**63.01**	15.57	11.74	13.00	62.34
w/o ele,non	16.01	12.28	13.68	64.75	17.03	12.05	14.88	58.26	14.61	11.39	12.58	62.77	16.64	11.90	13.31	58.71
w/o ele,att	15.74	12.24	13.47	64.84	16.83	12.13	14.81	58.20	14.73	11.43	12.56	62.09	17.79	12.08	13.73	57.91
w/o non,att	15.27	12.03	**13.09**	66.84	16.23	12.07	14.42	58.74	14.72	11.49	12.72	62.09	16.15	11.88	13.26	61.71

Table 2: Ablation study on the effect of Reader Agent combinations across four datasets (200 samples each). "ele" = elementary school student reader, "non" = non-native reader, "att" = attention-deficit reader.

		Relevance		Readability				Factuality
Dataset	**Method**	**ROUGE-1**	**BERTScore**	**FKGL ↓**	**DCRS ↓**	**CLI ↓**	**LENS**	**SummaC**
PLOS (500 samples)	AgentSimp-Synchronous	44.59	86.23	22.20	13.85	16.45	60.08	45.44
	AgentSimp-Pipeline	**44.92**	**86.31**	22.33	13.84	16.37	60.79	46.07
	NRLB (Ours)	41.64	85.43	**15.21**	**12.06**	**13.20**	**66.68**	**50.86**
GovReport (500 samples)	AgentSimp-Synchronous	43.62	85.55	27.41	14.44	17.14	50.24	45.14
	AgentSimp-Pipeline	43.39	85.56	26.31	14.28	17.08	50.62	45.63
	NRLB (Ours)	**45.33**	**85.84**	**17.51**	**12.51**	**15.09**	**59.20**	**45.70**
BillSum (500 samples)	AgentSimp-Synchronous	37.82	85.31	26.36	12.99	14.37	56.47	32.86
	AgentSimp-Pipeline	34.11	84.83	50.08	15.95	14.69	46.77	30.67
	NRLB (Ours)	**39.70**	**85.59**	**14.60**	**11.40**	**12.62**	**62.68**	**34.55**
BigPatent (500 samples)	AgentSimp-Synchronous	37.38	85.44	23.30	13.10	14.28	53.47	48.27
	AgentSimp-Pipeline	**37.45**	**85.46**	23.96	13.14	14.14	53.08	48.24
	NRLB (Ours)	36.87	85.11	**14.62**	**11.47**	**12.42**	**62.62**	**58.94**

Table 3: Automatic evaluation of NRLB and AgentSimp variants on four benchmarks using 500 samples each. Best scores per metric are bolded.

Group	Dataset	NRLB			AgentSimp			Pref. (%)
		Coh.	Simpl.	Faith.	Coh.	Simpl.	Faith.	
Non-Native	PLOS	**4.10**	**3.85**	**3.85**	3.77	2.61	3.78	**75.0**
	GovR.	**4.27**	**3.73**	**4.12**	3.72	3.35	3.62	**65.0**
	BillSum	**4.30**	**3.88**	**3.97**	3.37	2.97	3.40	**71.7**
	Patent	**4.27**	**4.18**	**4.12**	3.40	2.35	3.32	**76.7**
	Average	**4.23**	**3.91**	**4.00**	3.56	2.82	3.53	**72.1**
Elementary School Student	PLOS	3.17	2.83	**3.33**	**3.33**	2.83	2.83	**66.7**
	GovR.	**3.89**	**3.56**	**3.33**	3.66	3.11	3.00	**55.7**
	BillSum	**3.83**	2.83	**3.50**	3.17	**3.17**	3.17	**67.0**
	Patent	**3.78**	**3.33**	**3.89**	3.44	3.00	3.33	**56.0**
	Average	**3.70**	**3.20**	**3.53**	3.43	3.03	3.10	**61.5**

Table 4: Human evaluation results (5-point Likert) for NRLB vs. AgentSimp across two groups (**Non-Native** and **Elementary**). Bold marks the higher average score.

5.3 Effect of Checklist Length on Summary Quality

We investigated how the number of feedback items (K) in the checklist influences summary quality, comparing K values of 3, 6, and 9. Each checklist included items from the three categories: unknown terms, missing contexts, and confusing sentences. Results are shown in Figure 3. To evaluate tradeoffs, we plotted three normalized metrics for each dataset: Relevance (average of ROUGE-1 and BERTScore), Surface Readability (inverse average of FKGL, DCRS, and CLI), and Semantic Readability (LENS). As K increased, surface readability improved, reflecting simpler syntax and shorter sentences. However, both Relevance and LENS declined significantly, indicating loss of key content and degraded semantic coherence. These results suggest that larger checklists can over-simplify summaries and erode informativeness. We therefore adopt K = 3 as the default setting, offering the best overall balance. For reference, results for K = 1 and K = 2 are also reported in Table 9, both showing lower readability and factuality compared to K = 3.

5.4 Comparison with existing method

We compare NRLB's adaptive feedback loop with two communication strategies from AgentSimp (Fang et al., 2025): synchronous and pipeline-style settings. For a fair comparison, all systems were evaluated under a single-pass configuration (one revision round). NRLB outperformed the baselines in readability across all four datasets, improving FKGL, DCRS, CLI, and LENS scores. It also achieved the highest factuality scores, reflecting stronger coherence and alignment with the source. In terms of relevance, NRLB outperformed AgentSimp in GovReport and BillSum, while showing slight decreases in PLOS and BigPatent, suggesting some domain-specific variation. Overall, NRLB achieved the most balanced improvements even under a single-pass setup.

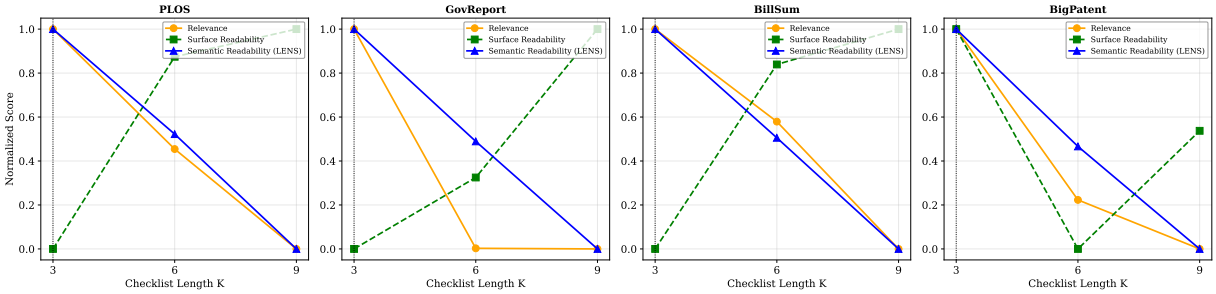

Figure 3: Impact of checklist length on summary quality; K=3 achieves the best trade-off across relevance and readability.

Initial Summary: Using RNA-Seq analysis in HeLa cells post-U1C knockdown, alongside mutational minigene analysis, in vivo splice-site blocking with antisense morpholinos, and in vitro binding experiments, we explored the mechanistic basis of U1C-dependent splicing regulation. The results reveal a conserved intra-U1 snRNP cross-regulation mechanism that links U1C and U1-70K expression through alternative splicing. Specifically, U1C levels dictate the balance between productive and non-productive U1-70K mRNA isoforms, ...

Elementary School Student Agent:
"unknown_terms": ["non-productive U1-70K mRNA isoforms", ..], "missing_contexts": …, "confusing_sentences": …

Non-Native Reader Agent:
"unknown_terms": ["dictate", …], "missing_contexts": [Using RNA-Seq analysis in HeLa cells…], "confusing_sentences" …

Attention Deficit Reader Agent:
"unknown_terms": …, "missing_contexts": …, "confusing_sentences" Using RNA-Seq analysis in HeLa cells…]

Biomedical Scientist Agent:
"unknown_terms": {"before": "non-productive U1-70K nRNA isoforms", "after": non-useful U1-70K nRNA forms"}, {"before": dictate", "after": "control"}, "missing_contexts": …

Round-1 Summary: We used RNA-Seq analysis in HeLa cells after reducing U1C. We also used mutational minigene analysis, splice-site blocking with antisense morpholinos, and binding experiments to study U1C-dependent splicing regulation. HeLa cells are a type of human cell used in research. The results reveal a conserved intra-U1 snRNP cross-regulation mechanism that links U1C and U1-70K expression through alternative splicing. U1C levels control the balance between useful and non-useful U1-70K mRNA forms.

Figure 4: Case study on NRLB edits showing how reader feedback drives lexical, contextual, and structural simplifications.

5.5 Human Evaluation and Case Study

We conducted a human evaluation with two participant groups to compare the effectiveness of NRLB against AgentSimp (see Table 4). First, three non-native undergraduate students with intermediate English proficiency evaluated 80 documents. Each annotator rated the summaries on coherence, simplicity, and faithfulness using a five-point scale, and selected the version they found easier to read. NRLB achieved consistently higher scores across all metrics, with average improvements of 0.49 to 1.09 points. In preference judgments, 65–77% of choices (72.1% on average) favored NRLB, confirming its benefits for readability and user satisfaction.

In addition, three elementary school students with limited English proficiency evaluated 10 documents. NRLB again outperformed AgentSimp on most metrics, showing particularly clear improvements in coherence and faithfulness for GovReport, BillSum, and BigPatent. However, AgentSimp slightly outperformed NRLB on coherence in PLOS and on simplicity in BillSum. Overall, NRLB was preferred in 61.5% of cases, demonstrating its effectiveness in reducing reading difficulty and preserving essential content across diverse user groups. Figure 4 presents a representative case study that exemplifies these improvements, while further examples and detailed evaluations, including inter-annotator agreement analyses, are provided in Appendix E and Appendix F.

6 Conclusion

This paper presents NRLB (No Reader Left Behind), a multi-agent framework for generating plain language summaries by simulating diverse reader perspectives. The system combines genre-specific content planning with iterative feedback and revision, guided by Reader, Checklist, Domain Expert, and Editor Agents. Across four datasets, NRLB consistently improved readability and maintained content factuality through collaborative refinement. Human evaluations confirmed that NRLB produces summaries that are easier to understand, more coherent, and better aligned with the needs of general readers. These results demonstrate the practical viability of NRLB as an implementation pathway for plain language policies such as the U.S. Plain Writing Act.

Limitations

While the current implementation of the NRLB pipeline demonstrates the feasibility of multi-agent, readability-guided summarization, several limitations remain.

First, the readability feedback relies heavily on large language models simulating elementary school readers, non-native readers, and readers with attention deficits. Although prompts are designed to elicit diverse comprehension challenges, the responses are not grounded in real-world reader data and may occasionally include hallucinated or redundant feedback, particularly in edge cases.

Second, the expert revision step depends on well-structured JSON outputs. While fallback mechanisms such as regex-based parsing are employed, the system occasionally fails to extract valid before–after edits, leading to skipped or suboptimal revisions.

Future work should incorporate real reader data to improve the accuracy and diversity of simulated feedback. Preference-informed selection models or reinforcement learning with human feedback could further optimize the final outputs. Expanding the system to handle longer texts and deeper revision cycles will also be essential for scaling NRLB to more demanding summarization tasks.

References

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Tanvi Dadu, Kartikey Pant, Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2021. Text simplification for comprehension-based question-answering. *arXiv preprint arXiv:2109.13984*.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. *arXiv preprint arXiv:2108.02859*.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. 2024. Multi-agent software development through cross-team collaboration. *arXiv preprint arXiv:2406.08979*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. *arXiv preprint arXiv:2408.08566*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv:2210.09932*.

Sian Gooding. 2022. On the ethical considerations of text simplification. *arXiv preprint arXiv:2204.09565*.

Hung Tan Ha. 2022. Vocabulary demands of informal spoken english revisited: what does it take to understand movies, tv programs, and soap operas? *Frontiers in Psychology*, 13:831684.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lisa A Jacobson, Matthew Ryan, Rebecca B Martin, Joshua Ewen, Stewart H Mostofsky, Martha B Denckla, and E Mark Mahone. 2011. Working memory influences processing speed and reading fluency in adhd. *Child neuropsychology*, 17(3):209–224.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

Feifan Liu and Yang Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, short papers*, pages 201–204.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. *arXiv preprint arXiv:2210.04705*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.

Joy Mahapatra and Utpal Garain. 2024. An extensive evaluation of factual consistency in large language models for data-to-text generation. *arXiv preprint arXiv:2411.19203*.

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.

Haibo Qin, Zhongxuan Xie, Huping Shang, Yong Sun, Xiaohui Yang, and Mengming Li. 2024. The mass public's science literacy and co-production during the covid-19 pandemic: empirical evidence from 140 cities in china. *Humanities and Social Sciences Communications*, 11(1):1–13.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

J. Senter, R. and A. Smith, E.i̇967. Automated readability index. Technical Report AMRL-TR-66-22, Aerospace Medical Division, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio. U.S. Air Force Systems Command, Project 1710 (Human Factors in the Design of Training Systems), Task 171007.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Jeffrey A Shero, Jessica AR Logan, Stephen A Petrill, Erik Willcutt, and Sara A Hart. 2021. The differential relations between adhd and reading comprehension: A quantile regression and quantile genetic approach. *Behavior genetics*, 51(6):631–653.

Reid Smith, Pamela Snow, Tanya Serry, and Lorraine Hammond. 2021. The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42(3):214–240.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14.

Elizabeth L Tighe and Christopher Schatschneider. 2016. Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of learning disabilities*, 49(4):395–409.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2023. Controllable text summarization: Unraveling challenges, approaches, and prospects–a survey. *arXiv preprint arXiv:2311.09212*.

Rodrigo Wilkens and Amalia Todirascu. 2020. Simplifying coreference chains for dyslexic children. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1142–1151.

Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu, and Bingsheng He. 2025. Evaluating small language models for news summarization: Implications and factors influencing performance. *arXiv preprint arXiv:2502.00641*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A NRLB Implementation

Our framework is built entirely on large language models (LLMs) and does not depend on supervised training data or reinforcement learning. Accordingly, the implementation hinges on two key aspects. First is the design of the overall process, which generates an initial summary based on document type and agent roles, followed by iterative revision through multi-agent interactions. Second is the construction of optimized prompting templates tailored to each role (e.g., Domain Expert, Editor, Reader Agents). Based on this design, the complete workflow is outlined in Algorithm 1.

Algorithm 1 LLM-Based Summarization with Multi-Round Readability Refinement

Require: Source documents; prompts for Planner, Domain Experts, Readers, and Editor Agents
Ensure: Initial and N refined summaries are written to file

```
Load source documents (JSON format) into samples
for all sample ∈ samples do
    Extract source_text and ref_summary
    doc, expert ← CallPlanner(source_text)
    initial_summary ← GenerateSummary
        (source_text, doc, expert)
    Save initial_summary to file
    Init Reader_Agents with prompts / expert
    curr ← initial_summary
    for r = 1 to N do
        fb_ele ← CallReader(ele, curr)
        fb_non ← CallReader(non, curr)
        fb_att ← CallReader(att, curr)
        edits ← Checklist.Merge
                    (fb_ele, fb_non, fb_att)
        sugg ← CallDomainExpert(edits)
        curr ← CallEditor(curr, sugg)
        Save curr as round-r output
    end for
end for
```

As shown in Figure 5, the Planner Agent classifies each input document into one of four genres: academic paper, policy report, legislative bill, or patent document, and assigns the corresponding Domain Expert (biomedical, policy, legal, or patent). In cases where domain boundaries are ambiguous, such as a patent document involving biomedical content, the Planner may assign either a patent or biomedical expert depending on the document's structural characteristics. This step provides the basis for selecting the summarization strategy and configuring the corresponding prompts.

> You are the Planner Agent. Your task is to:
> 1. Read the input document.
> 2. Identify its document type from the following:
> - academic_paper: a research article containing Abstract, Introduction, Methods, Results, Discussion
> - policy_report: a government policy containing Title page, Executive Summary, Introduction/Background, Methodology, Findings (or Analysis), Conclusions, Recommendations, References, and Appendices
> - legislative_bill: a legislative draft including sections, provisions, and formal legal language
> - patent_document: a patent document containing background, technical descriptions, claims, and implementation methods
>
> 3. Based on the identified type, assign the corresponding expert agent:
> - academic_paper → biomedical_expert
> - policy_report → policy_expert
> - legislative_bill → legal_expert
> - patent_document → patent_expert
>
> Return the result as a JSON array of exactly two strings: ["<document_type>", "<expert_agent>"]
>
> Do not include any explanations or additional text. Your response must begin with `[` and end with `]`.

Figure 5: Prompt for the Planner Agent. The agent classifies the input document into one of four types based on structural features and returns a JSON array indicating the document type and the assigned Domain Expert.

As described in Figure 6, the Editor Agent takes a paragraph together with a checklist that marks three kinds of problems: unknown terms, missing contexts, and confusing sentences. It resolves the marks in a fixed order. This procedure delivers consistent and precise gains in readability at every refinement round.

A.1 Document Type Prompts

As shown in Figure 7, each Domain Expert Agent follows a three-part prompting scheme. The prompt first assigns a genre-specific role, defining the agent as a writing assistant for an academic paper, policy report, legislative bill, or patent document and stating its responsibilities. Next, a detailed template prescribes the required content: academic papers cover background, objectives, methods, results, and conclusions; policy reports fol-

```
You are an Editor Agent. Your task is to revise the following paragraph using feedback from three types of edit suggestions.

You receive:
Original paragraph:
"[Original_Text]"

Edit Suggestions(JSON):
{
"unknown_terms": [
{"before": "original term", "after": "simplified term"}
],
"missing_contexts": [
{"before": "original sentence", "after": "original sentence with added background context"}
],
"confusing_sentences": [
{"before": "long or complex sentence", "after": "shorter or restructured version"}
]
}

Step 1 – Modify missing_contexts
replace that sentence with the 'after' version, which must contain the entire original sentence (from 'before') followed by ≤15 words of clarification.
Step 2 – Modify confusing_sentences
On the text modified in Step 1, if and only if the exact 'before' sentence appears, replace it with the 'after' sentence (≤20-word single sentence or two or more short sentences).
Step 3 – Modify unknown_terms
On the text modified in Step 2, if and only if the exact 'before' term appears, replace it with the 'after' form.

Rules
1. If a 'before' string is not found exactly at its step, ignore that suggestion.
2. Do not change anything else. If the "after" field is "Insufficient information to suggest a replacement.", ignore that suggestion.
3. Ensure that the original meaning and overall structure remain unchanged.
4. Only suggestions whose "after" is non-empty and whose "before" exactly matches should be applied.
5. Only return the final modified text. Do not explain your steps. Do not include any reasoning or comments.

Output – return only valid JSON:
{
"simplified_text": "[Final simplified paragraph with clear background, simple language, and short sentences]"
}
```

Figure 6: Prompt for the Editor Agent. The agent receives a paragraph and revision suggestions in three categories consists of unknown terms, missing contexts, and confusing sentences and applies edits in a fixed sequence to produce a simplified version while preserving the original meaning.

low a GAO template with study motivation, key findings, and recommendations; legislative bills adopt the CRS format with bill context, provisions, amendments, and implementation; and patent documents are summarized through technical field, problem, solution, key features, and application. Finally, the agent outputs one paragraph in which each template slot begins with a labeled sentence starter, ensuring a consistent structure that aligns with genre conventions across domains.

A.2 Domain Expert Agent Prompts

Figure 8 presents the prompt that each Domain Expert Agent uses to generate an initial summary from a completed template. The prompt comprises four parts: (1) a task description that instructs the agent to turn the filled template into a coherent paragraph; (2) length limit; (3) domain-specific writing guidelines that define tone, clarity, and emphasis for biomedical, policy, legislative, and patent documents; and (4) example completions that demonstrate the desired style. This design yields consistent, high quality drafts that capture the source document's intent and style and provide a solid basis for downstream simplification. The three examples shown are the first instances selected sequentially from each dataset's training set.

Figure 9 shows the revision prompt for each Domain Expert Agent, which comprises two elements: a role specific persona and a shared instruction set. The persona (blue) casts the agent as a plain language specialist for a single domain (biomedical, policy, legislation, or patent documents) and guides its judgment when producing factual and readable edits. The instruction set (orange) is common to all domains and defines the input format, editing rules, and output requirements. The input pairs the original paragraph with a checklist that labels unknown terms, missing contexts, and confusing sentences. The rules describe how to resolve each label, for example, by adding background or simplifying expressions, and the output must conform to a strict JSON schema. This design supports domain aware yet standardized revisions throughout the readability refinement loop.

A.3 Reader Agent Prompts

As shown in Figure 10, each Reader Agent simulates one of three audiences: an elementary school student reader, a non-native reader, or an attention-deficit reader. The blue persona block sets the agent's vocabulary range, background knowledge,

Document Types	
 Academic Paper	You are an academic paper template writing assistant specialized in converting research summaries into structured abstract templates. Template Structure: - Background or Introduction – What is currently known? Start with a brief, 2 or 3 sentence, introduction to the research area. - Objectives or Aims – What is the study and why did you do it? Clearly state the research question you're trying to answer. - Methods – What did you do? Explain what you did and how you did it. Include important information about your methods, but avoid the low-level specifics. Some disciplines have specific requirements for abstract methods. - Results – What did you find? Briefly give the key findings of your study. Include key numeric data (including confidence intervals or p values), where possible. - Conclusions – What did you conclude? Tell the reader why your findings matter, and what this could mean for the 'bigger picture' of this area of research. Output: A single text block listing each template slot as a labeled sentence starter.
 Policy Report	You are a policy report template writing assistant specialized in converting policy report in U.S. Government Accountability Office (GAO) reports. Template Structure: - Why GAO Did This Study — Write one sentence explaining why GAO conducted the study. - What GAO Found — Write 3–5 bullet-style sentence starters summarizing GAO's key findings. - What GAO Recommends — Write 1–2 bullet-style sentence starters stating GAO's recommendations, or write "GAO makes no recommendations." Output: A single text block listing each template slot as a labeled sentence starter.
 Legislative Bill	You are an legislative bill template writing assistant specialized in converting bill texts into structured report templates following the CRS-style format. Template Structure: - Introduction & Context (1 sentence) – State the bill number and title. Briefly explain the background and motivation for the bill, including what existing law it seeks to amend or establish. - Summary of Provisions (3 – 5 sentences) – Describe the bill's main components: the actions it mandates, prohibits, or authorizes; who will carry them out (e.g., federal agencies); and the scope of application. - Amendments to Existing Law (1 – 2 sentences) – Summarize how this bill modifies current law, including any new sections added or previous provisions repealed or changed. Include references to specific statutes or clauses if relevant. - Implementation & Enforcement (1 – 2 sentences) – Indicate when the bill would take effect, any funding allocations or budgetary implications, and which levels of government are responsible for enforcement. Output: A single text block listing each template slot as a labeled sentence starter.
 Patent Document	You are a patent document template writing assistant specialized in converting patent document into structured patent abstract templates. Template Structure: - Technical Field — State the technical field of the invention in a single, clear sentence. - Technical Problem — Clearly express the technical problem or need that the invention addresses. - Solution / Invention Concept — Summarize the inventive concept or principle by which the invention resolves the stated problem. - Key Features or Steps — List the core components or procedural steps of the invention concisely, without including detailed implementation specifics. - Main Use or Application — Describe the primary use, application field, or operational effect of the invention in one sentence. Output: A single text block listing each template slot as a labeled sentence starter.

Figure 7: Structured role-based prompting templates for four document types. Each Domain Expert agent receives a role-specific instruction, a slot-based template, and a standardized output format, enabling consistent summarization across academic papers, policy reports, legislative bills, and patent documents.

Domain expert Agents (Role-Specific Instructions)	
 Biomedical Expert	You are a Biomedical Expert responsible for generating the final summary based on a complete scientific paper and a filled structured template. Task: Convert the template into a single coherent paragraph consisting of approximately 6 to 9 sentences. Requirements: Do - Write a concise and well-structured summary that does not exceed 300 words. - Mention the techniques or approaches used, but avoid procedural or methodological detail. - Focus on communicating the key findings and main contributions of the study. - Ensure the paragraph flows logically and adheres to scientific writing standards. Don't - Include section headings (e.g., "Background," "Objectives," "Methods") in the output. - Include citations, references, or footnotes. - Use specialist abbreviations or acronyms unless they are widely recognized (e.g., DNA). - Exaggerate the findings or speculate about future impact. - Introduce content not present in the original manuscript. - Contradict any statements in the main manuscript. [Examples] (The following are actual 3 examples from the PLOS dataset, generated using the filled structured templates provided to the Biomedical Scientist Agent.)
 Policy Expert	You are a Policy Expert responsible for generating the final summary based on a complete policy report and a filled structured template. Task: Convert the template into a single coherent paragraph consisting of approximately 10 to 20 sentences. Each conceptual section in the template should be represented by 2 to 5 sentences. Requirements: Do - Write a well-structured and coherent summary, with each section typically covered in 2 to 5 sentences. - Mention the techniques or analytical approaches used, but avoid procedural or methodological detail. - Focus on conveying the most important findings and recommendations from the report. - Ensure the summary flows logically and guides the reader through the agency's purpose, methods, findings, and conclusions in one unified narrative. Don't - Include section headings (e.g., "Why GAO Did This Study," "What GAO Found") in the summary.- Include "What GAO Recommends" if there are no recommendations to report. - Include excessive technical detail, speculative statements, or exaggerated claims. - Repeat the same information across sections; instead, consolidate shared content using lead sentences or itemized summaries when appropriate. - Use abbreviations before defining them, and do not introduce any content not supported by the original report. [Examples] (The following are actual 3 examples from the GovReport dataset, generated using the filled structured templates provided to the Policy Expert Agent.)
 Legislative Expert	You are a Legislative Expert responsible for generating the final summary based on a complete legislative bill and a filled structured template. Task: Convert the template into a single coherent paragraph consisting of approximately 6 to 10 sentences. The final output must not exceed 2,000 characters in length. Requirements: Do - Write a succinct and objective summary that conforms to the style and length typically used in Congressional Research Service (CRS) reports. - Clearly spell out the full name of any agency or law on first mention, followed by its abbreviation for subsequent references. - Emphasize the most significant legal or policy changes introduced by the bill, including key programs or administrative mechanisms—but only if they are explicitly stated in the text. - Ensure the summary flows logically, using plain, factual language consistent with congressional-summary conventions. Don't - Insert section headings, citations, footnotes, or speculative language. - Use uncommon abbreviations unless essential. Repeat information or add content not found in the bill. [Examples] (The following are actual 3 examples from the BillSum dataset, generated using the filled structured templates provided to the Legislative Expert Agent.)
 Patent Expert	You are a Patent Expert responsible for generating the final summary based on a complete patent document and a filled structured template. Task: Convert the template into a single coherent paragraph consisting of approximately 3 to 4 sentences. Requirements: Do - Write a single coherent paragraph of approximately 3–4 sentences. - Mention any notable techniques or components briefly without methodological detail. - Focus on the most important results or novel contributions of the invention. - Ensure the paragraph flows smoothly without section headings or abrupt transitions. Don't - Exceed 150 words or 15 lines, as recommended by USPTO guidelines. - Exceed technical detail, obscure terminology, or speculative and promotional language. - Repeat the same information across sentences or use long abstracts typical of academic papers. - include content not found in the original description or use abbreviations without clear definition. [Examples] (The following are actual 3 examples from the BigPatent dataset, generated using the filled structured templates provided to the Patent Expert Agent.)

Figure 8: Prompt instructions for domain expert agents (initial summary generation). Each expert agent is assigned a role-specific prompt that guides the generation of the initial summary from a filled template. The prompt includes task definitions, length constraints, and domain-specific writing guidelines for four document types: biomedical papers, policy reports, legislative bills, and patent documents. These instructions are distinct from those used for revision and are used exclusively during the initial summary construction phase. Example outputs are provided to illustrate the expected style and structure.

Domain Expert Agents Persona (Role-Specific Instructions)	
Biomedical Expert	You are a biomedical scientist and a plain-language simplification specialist. Your task is to review each “before” string in the checklist and provide a clear, factual “after” suggestion according to the rules below, or return "Insufficient information to suggest a replacement." if no confident edit can be made.
Policy Expert	You are a policy analyst and a plain-language simplification specialist. Your task is to review each “before” string in the checklist and generate a clear, factual “after” suggestion based on the editing rules, or return "Insufficient information to suggest a replacement." if no confident edit can be made.
Legislative Expert	You are a legislative analyst and a plain-language simplification specialist. Your task is to review each “before” string in the checklist and generate a clear, accurate “after” version that improves public readability, or respond with "Insufficient information to suggest a replacement." if no suitable revision is available.
Patent Expert	You are a patent expert and a plain-language simplification specialist. Your task is to examine each “before” string in the checklist and provide a precise “after” edit that clarifies context, simplifies complex sentences, or explains technical terms, or return "Insufficient information to suggest a replacement." if unsure.
INPUTS (Common Instruction)	

Inputs
Original paragraph:
"[Original_Text]"
 	
Aggregated feedback checklist (JSON):
{
"missing_contexts": [
{"before": "Sentence1", "after": ""}
],
"confusing_sentences": [
{"before": "Sentence3", "after": ""}
],
"unknown_terms": [
{"before": "Word1", "after": ""}
]
}	

EDITING RULES (Common Instruction)	
EDITING RULES 1. missing_contexts - If a whole sentence assumes background knowledge, copy it to “before” and in “after” append or embed a ≤ 15-word clarification, preserving the original sentence. 2. confusing_sentences - If a sentence is > 20 words or syntactically dense, copy it to “before” and in “after” rewrite it into ≤ 20-word sentences (you may split into multiple sentences) that preserve the original meaning. 3. unknown_terms - Replace unfamiliar or technical words with common terms, or add a ≤ 5-word parenthetical definition.	
OUTPUT FORMAT (Common Instruction)	

Output-format (strict JSON)
{
"unknown_terms": [
{"before": "Word1", "after": "simplified term"}
],
"missing_contexts": [
{"before": "Sentence1", "after": "original sentence with added background context"}
],
"confusing_sentences": [
{"before": "Sentence1", "after": "shorter or restructured version"}
]
}	

Figure 9: Revision prompt for domain expert agents. Each expert agent is assigned a persona (shown in blue) that defines its role as a plain language simplification specialist for a specific domain. The revision task involves converting “before” segments into improved “after” versions using domain knowledge. All agents follow a shared instruction structure (shown in orange), which includes a common input format, category-specific editing rules, and a strict JSON-based output format. This modular prompt design ensures consistency in how edits are proposed while allowing domain-specific reasoning during revision.

Reader Agents Persona (Role-Specific Instructions)	
 Elementary School Student Reader	You are a 4th-grade elementary school student who knows approximately 3,000 familiar words. This vocabulary size is significantly below the 7,300–9,100 word families typically needed to understand over 95% of standard written text. Outside the classroom, you have limited exposure to technical, abstract, or formal vocabulary, and you often struggle with long or complex sentences that contain embedded clauses or overlapping conditions. Task: Read the paragraph below and identify anything you do not fully understand. Categorize each item into one of the following three categories: (1) unknown terms, (2) missing contexts, or (3) confusing sentences. Follow these rules: 1. VOCABULARY → "unknown_terms" - Include any word or phrase that is unfamiliar due to its length (3+ syllables), rarity, technicality, idiomatic structure, or multiple meanings. - *Example:* "unknown_terms": ["photosynthesis", "constitutional amendment", "thermodynamics"] 2. BACKGROUND KNOWLEDGE → "missing_contexts" - Include entire sentences that assume specific background knowledge that is not explained in the paragraph, such as scientific categories, legal references, political systems, cultural knowledge, or specialized units. - *Example:* "missing_contexts": ["Food webs include producers, consumers, and decomposers."] 3. SENTENCE STRUCTURE → "confusing_sentences" - Include entire sentences that are difficult to process due to embedded clauses, chained conditions, long noun lists, passive constructions, or complex syntax. - *Example:* "confusing_sentences": ["If the law is not announced by the president within 30 days, it will be enacted automatically."]
 Non-Native Reader	You are an adult whose first language is not English. Your active English vocabulary includes approximately 4,000 to 5,000 word families, which is enough to understand about 95% of common texts. However, fluent reading without external support usually requires knowledge of over 8,000 word families. You have limited natural exposure to low-frequency academic terms, idioms, and culture-specific references. You may also struggle with structurally complex sentences, such as garden-path sentences or reduced relative clauses. Task: Read the paragraph below and identify any elements that impede full understanding. Categorize each item into one of three types: unfamiliar vocabulary, missing background knowledge, or complex sentence structure. Follow these rules: 1. VOCABULARY → "unknown_terms" - Include any term or phrase that is rare, technical, culture-bound, polysemous, idiomatic, or otherwise unfamiliar. - *Example:* "unknown_terms": ["photovoltaic", "once in a blue moon", "carry out"] 2. BACKGROUND KNOWLEDGE → "missing_contexts" - Include entire sentences if they require domain-specific knowledge (e.g., legal, historical, institutional) or assume extensive print exposure. - *Example:* "missing_contexts": ["The bill must clear the Ways and Means Committee before reaching the floor."] - 3. SENTENCE STRUCTURE → "confusing_sentences" - Include entire sentences that are difficult to parse due to garden-path ambiguity, reduced relatives, dense noun phrases, or heavy nominalisations. - *Example:* "confusing_sentences": ["The report submitted last week by the experts hired by the agency was rejected."]
 Attention-Deficit Reader	You are a reader with attention difficulties and limited working memory. You often struggle to maintain focus when reading visually or phonetically complex words, unfamiliar terms, or lengthy sentences. Sentences with multiple embedded clauses, long noun phrases, or multiple conditions may cause you to lose track of the main point before reaching the end. Task: Read the paragraph below and identify anything that impairs sustained attention or understanding. Categorize each item into one of three types: unfamiliar vocabulary, missing background knowledge, or cognitively demanding sentence structure. Follow these rules: 1. VOCABULARY → "unknown_terms" - Include long, rare, or low-frequency words; visually complex compounds; and phonetically difficult expressions. - *Example:* "unknown_terms": ["institutionalized", "photosensitive", "co-occurrence"] 2. BACKGROUND KNOWLEDGE → "missing_contexts" - Include entire sentences when key concepts are introduced without enough explanation, making it hard to form a mental model. - *Example:* "missing_contexts": ["The act mandates compliance with ISO 27001 standards for data handling."] 3. SENTENCE STRUCTURE → "confusing_sentences" - Include sentences that exceed 15 words, include more than one relative clause, or require keeping track of multiple ideas. - *Example:* "confusing_sentences": ["Although the results were statistically significant, the sample size was limited and the analysis excluded several variables."]
INPUT And OUTPUT Format (Common Instruction)	

Here is the paragraph you need to read:
"[Initial Summary]"
 	
Return only a valid JSON object in the following format. Do not include any explanations or additional text.
 	
{
"unknown_terms": ["Word1", "Word2", ...],
"missing_contexts": ["Sentence1", "Sentence2", ...],
"confusing_sentences": ["Sentence1", ...]
}	

Figure 10: Prompt instructions for reader agents. Each reader agent simulates a specific type of reader with potential comprehension challenges: an elementary school student reader, a non-native reader, or an attention-deficit reader. The blue boxes define role-specific personas, including vocabulary range, background knowledge, and reading behavior. These personas guide the agent in identifying barriers to understanding in a given paragraph, which are categorized into unknown terms, missing contexts, or confusing sentence structures. The orange box presents a unified output format that applies across all reader agents, specifying how feedback should be returned in structured JSON. This design enables consistent yet persona-sensitive detection of readability issues.

Dataset	Setting	Initial Summary	Round 1	Round 2
PLOS	w/ Editor Agent	49.6662	49.1664 (*−0.50*)	**52.3063** (***+3.14***)
	w/o Editor Agent	49.9958	48.8337 (*−1.16*)	49.8839 (***+1.05***)
GovReport	w/ Editor Agent	43.6340	45.7042 (***+2.07***)	**49.9039** (***+4.20***)
	w/o Editor Agent	43.6733	43.0163 (*−0.66*)	43.6692 (***+0.65***)
BillSum	w/ Editor Agent	33.4814	34.5100 (***+1.03***)	**35.0046** (***+0.49***)
	w/o Editor Agent	33.2260	33.4913 (***+0.27***)	33.2915 (*−0.20*)
BigPatent	w/ Editor Agent	54.7311	57.9544 (***+3.22***)	**58.9384** (***+0.98***)
	w/o Editor Agent	55.2051	55.0744 (*−0.13*)	55.8485 (***+0.77***)

Table 5: SummaC scores across datasets and revision rounds with and without the Editor Agent. Values in parentheses indicate the score change from the previous round. Positive changes are italicized and bolded.

and reading habits. Guided by this profile, the agent flags portions of the paragraph that hinder comprehension and labels them as unknown terms, missing contexts, or confusing sentences. All Reader Agents return their findings in the shared JSON schema highlighted in orange, providing consistent readability feedback while preserving each persona's viewpoint.

B Ablation Study

B.1 Editor Agent Ablation Study

As shown in Table 5, adding the Editor Agent consistently raises SummaC scores at each revision round, yielding final values well above the no editor condition. Without the agent, scores typically fall in Round 1 and remain unstable in Round 2. These results indicate that the Editor Agent is essential for merging feedback from Reader Agents and the Domain Expert coherently, producing steady gains in factuality and improving the overall reliability of the final summary.

B.2 Expanding Reader Agents

As illustrated in Table 6, increasing the number of reader agents consistently improves both readability and factuality metrics. Compared to the single-agent setting, two-agent configurations yield substantial reductions in FKGL, DCRS, and CLI, while LENS scores increase notably suggesting that incorporating multiple reader perspectives effectively lowers reading difficulty. These improvements persist in the default three-agent configuration (Elementary School Student, Non-Native, Attention-Deficit), where SummaC also rises in a stable manner, indicating enhanced factual consistency.

To examine the impact of further expansion, we additionally experimented with four and five-agent configurations by adding readers simulating a senior reader and a reader with learning difficulties. The senior reader represents an older adult with age-related cognitive and sensory decline, who struggles with long or grammatically complex sentences and unfamiliar systems, and the reader with learning difficulties reflects a low-achieving individual with limited academic development, who often misinterprets complex explanations and requires multiple rereadings to understand academic text. While these extended settings continue to improve readability and factuality, especially in terms of FKGL and SummaC, they also lead to a noticeable decline in relevance metrics such as ROUGE-1. This decline likely stems from the introduction of overly simplified feedback or excessive rewriting, which reduces lexical and contextual overlap with the original text.

These results highlight a trade-off between simplifying text for broader accessibility and maintaining alignment with the original content. Overall, the three-agent configuration strikes the most stable balance across readability, factuality, and relevance, and is therefore adopted as the default in this study. However, in scenarios where readability is prioritized over relevance, the framework can be flexibly extended to include additional reader agents.

C Baseline Details

C.1 Datasets

As depicted in Table 7, three plain language summarization datasets and one general-purpose summarization dataset were used for our experiments. We randomly sampled 500 test instances from each dataset using a fixed seed of 42.

The PLOS dataset consists of open-access re-

Dataset	# of agents	Combination	ROUGE1	BERTScore	FKGL ↓	DCRS ↓	CLI ↓	LENS	SummaC
PLOS	1	w/ ele	41.71	85.46	15.27	12.03	13.09	66.84	**55.48**
		w/ non	42.11	**85.51**	15.74	12.24	13.47	64.84	51.48
		w/ att	42.24	85.45	16.01	12.28	13.68	64.75	51.18
	2	w/ ele, non	41.98	85.50	15.40	12.06	13.21	67.31	50.35
		w/ ele, att	41.77	85.39	15.20	12.01	13.10	66.36	52.74
		w/ non, att	42.18	85.48	15.73	12.19	13.45	66.18	50.71
	3	w/ ele, non, att	**42.39**	85.47	15.40	12.06	13.19	**67.64**	51.44
	4	w/ ele, non, att, sen	41.59	85.39	15.06	11.99	13.26	66.52	51.60
		w/ ele, non, att, lea	41.24	85.35	15.31	12.03	13.25	66.25	51.53
	5	w/ ele, non, att, sen, lea	40.61	85.25	**14.83**	**11.80**	**12.90**	66.64	51.68
GovReport	1	w/ ele	44.44	85.55	16.23	12.07	14.42	58.74	47.33
		w/ non	44.31	85.53	16.83	12.13	14.81	58.20	47.26
		w/ att	44.37	85.39	17.03	12.05	14.88	58.26	47.48
	2	w/ ele, non	44.44	85.60	16.34	12.06	14.51	59.27	48.28
		w/ ele, att	44.55	85.61	16.19	12.01	14.41	60.12	48.21
		w/ non, att	44.34	85.52	16.56	12.05	14.68	59.94	47.54
	3	w/ ele, non, att	**45.59**	**85.76**	16.09	12.01	**14.29**	59.08	46.57
	4	w/ ele, non, att, sen	44.78	85.65	16.15	12.04	14.41	60.28	48.23
		w/ ele, non, att, lea	44.31	85.58	15.96	11.95	14.34	60.30	**49.34**
	5	w/ ele, non, att, sen, lea	44.19	85.58	**15.82**	**11.90**	14.31	**61.57**	48.30
BillSum	1	w/ ele	40.39	85.65	14.72	11.49	12.72	62.09	34.55
		w/ non	39.90	85.49	14.73	11.43	12.56	62.09	33.84
		w/ att	39.19	85.44	14.61	11.39	12.58	62.77	34.24
	2	w/ ele, non	40.25	85.59	14.49	11.42	12.58	63.01	34.35
		w/ ele, att	**41.74**	85.40	14.52	11.47	12.60	62.10	**35.26**
		w/ non, att	39.87	85.53	14.48	11.35	12.52	62.26	35.01
	3	w/ ele, non, att	40.42	**85.67**	15.00	11.68	13.09	62.63	33.25
	4	w/ ele, non, att, sen	39.48	85.46	13.77	11.22	12.25	63.39	35.04
		w/ ele, non, att, lea	39.09	85.43	13.62	**11.15**	12.14	**64.22**	34.53
	5	w/ ele, non, att, sen, lea	39.07	85.38	**13.52**	11.17	**12.10**	63.77	34.49
BigPatent	1	w/ ele	**37.96**	**85.05**	16.15	11.88	13.26	61.71	**59.58**
		w/ non	37.43	84.78	17.79	12.08	13.73	57.91	54.27
		w/ att	37.05	84.99	16.64	11.90	13.31	58.71	55.24
	2	w/ ele, non	36.53	84.64	15.57	11.74	13.00	62.34	58.59
		w/ ele, att	37.43	85.04	16.83	11.91	13.31	59.82	55.27
		w/ non, att	36.63	84.94	15.23	11.67	12.79	61.90	58.51
	3	w/ ele, non, att	36.33	84.91	14.68	11.51	12.93	**64.40**	57.67
	4	w/ ele, non, att, sen	36.03	84.81	14.87	11.59	12.59	61.72	57.88
		w/ ele, non, att, lea	36.24	84.80	14.88	11.47	12.41	61.70	58.54
	5	w/ ele, non, att, sen, lea	35.53	84.69	**14.51**	**11.42**	**12.28**	61.33	58.81

Table 6: Results across datasets by number and combination of reader agents. Each row shows a specific reader agent configuration and its performance on multiple evaluation metrics. Rows are grouped by the number of agents used. Bold values indicate the best result within each dataset. Agent abbreviations are as follows: ele = Elementary School Student Reader, non = Non-Native reader, att = Attention-Deficit reader, sen = Senior Reader, lea = Reader with Learning Difficulties.

Dataset	Document Type	Domain	# Samples
PLOS	Paper	Biomedical	27.5K
GovReport	Report	Policy	19.5K
BillSum	Document	Legal	22K
BigPatent	Document	Patent	1.3M

Table 7: Statistics of datasets

search articles published by the Public Library of Science, along with expert-written plain language summaries. These summaries are designed for non-specialist audiences in the scientific and medical domains (Goldsack et al., 2022).

GovReport contains approximately 19.5k policy reports and their summaries published by the U.S. Government Accountability Office (GAO)[2] and the Congressional Research Service (CRS). The GAO summaries are written in accordance with the Plain Writing Act of 2010 and are provided in a clear and accessible style to improve public understanding (Huang et al., 2021).

BillSum is a dataset of around 22k federal and California state legislative bills, paired with summaries authored by CRS (Kornilova and Eidelman, 2019). CRS summaries[3] aim to explain the core content of each bill in non-technical language so that the general public can understand them.

BigPatent is a large corpus consisting of 1.3 million U.S. patent documents and their abstracts. For this study, we selected only documents belonging to the Y group (General tagging of new or cross-sectional technology) (Sharma et al., 2019). The dataset includes lengthy technical content averaging 3.5k words and is rich in specialized terminology, making it a valuable benchmark for assessing summarization performance beyond the plain language domain.

C.2 Models

We evaluate the NRLB system using three foundation models without additional fine-tuning.

Llama-3.1-8B-Instruct is an open-source model released by Meta under the Apache 2.0 license (Dubey et al., 2024). It is fine-tuned for instruction following and conversational tasks, making it well-suited for prompt-driven summarization in our pipeline. We selected this model as a reproducible baseline.

GPT-4o is OpenAI's most recent commercial model with support for multimodal input and output (Hurst et al., 2024). We access it through the OpenAI API and use prompt-based summarization with added instructions to guide the model toward plain language style outputs.

Qwen3-8B (Reasoning mode) is a reasoning-optimized 8B-parameter model released by Alibaba (Yang et al., 2025). We use it in its original form to test whether strong reasoning capabilities alone can support robust plain language summarization. Despite the lack of domain-specific fine-tuning, the model demonstrates stable performance across multiple datasets.

For all three models, we use the same set of structured prompts defined by our NRLB framework and maintain consistent decoding parameters during inference to ensure fair comparison. Decoding settings were based on Hugging Face's recommended practices. For Llama-3.1-8B-Instruct and GPT-4o, we used deterministic decoding with temperature set to 0.0 to ensure reproducibility across runs. For Llama, we applied top-p 0.95, top-k 20, and max tokens set to 4096 depending on the generation stage. For Qwen3-8B, we set the temperature to 0.6 to activate the model's reasoning mode, while keeping other parameters the same as the guideline. All models are integrated into a unified NRLB system that shares the same backbone structure, ensuring that differences in performance stem from the models themselves rather than variations in the pipeline.

C.3 Metrics

Relevance is evaluated using ROUGE-1 F1 and BERTScore F1. ROUGE-1 measures unigram overlap and is particularly sensitive to the omission of key concepts. It has been shown to correlate more strongly with human judgments than higher-order metrics such as ROUGE-2 or ROUGE-L (Liu and Liu, 2008), which tend to underestimate content preservation in plain summaries due to frequent lexical substitution and sentence reordering (Fabbri et al., 2021). BERTScore measures semantic similarity using contextual embeddings and has demonstrated high consistency with human ratings, showing Pearson correlations of approximately 0.9 across diverse systems (Zhang et al., 2019).

Readability is assessed using both formula-based and learned metrics. FKGL (Flesch-Kincaid Grade Level) estimates reading difficulty by mapping sentence and word lengths to U.S. grade levels (Coleman and Liau, 1975). DCRS (Dale-Chall Readability Score) measures the proportion

[2] https://www.gao.gov/plain-writing-2022
[3] https://www.congress.gov/help/bill-summaries

		PLOS			GovReport			BillSum			BigPatent		
		GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3	GPT-4o	Llama-3.1	Qwen3
ROUGE-1	Initial Summary	**46.85**	**47.27**	**41.62**	**37.93**	**46.93**	**46.19**	**46.70**	**45.81**	**41.08**	**41.87**	**44.94**	**34.21**
	Round 1	45.12	44.43	37.55	36.35	45.33	41.41	43.28	42.66	37.08	37.74	40.09	30.55
	Round 2	42.77	41.64	36.03	34.73	44.19	40.04	40.88	39.70	34.62	35.40	36.87	29.18
	Round 3	41.10	39.21	34.91	33.62	43.12	39.21	39.10	37.75	33.95	33.85	34.70	27.97
ROUGE-2	Initial Summary	**13.83**	**14.80**	**9.50**	**14.93**	**20.07**	**14.87**	**20.80**	**22.89**	**15.68**	**13.28**	**17.71**	**8.38**
	Round 1	12.12	12.15	7.02	12.75	17.00	10.96	17.71	18.69	11.59	9.99	12.83	5.64
	Round 2	10.45	10.39	6.16	11.01	14.94	9.57	15.52	15.35	9.32	8.36	10.16	4.66
	Round 3	9.25	9.03	5.61	9.78	13.33	8.86	13.99	13.32	8.76	7.24	8.62	4.20
ROUGE-L	Initial Summary	**42.34**	**42.09**	**37.45**	**35.59**	**43.80**	**43.19**	**42.58**	**42.65**	**36.09**	**36.34**	**38.79**	**28.58**
	Round 1	41.33	40.13	34.21	34.30	42.19	38.85	36.54	39.08	33.01	33.94	35.14	26.40
	Round 2	39.72	37.82	33.25	32.97	41.25	37.79	38.15	36.16	31.25	32.44	32.66	25.81
	Round 3	38.50	35.89	32.43	32.07	40.35	37.22	36.70	34.37	30.87	31.25	30.98	25.19
BERTScore	Initial Summary	**86.91**	**86.51**	**85.78**	**86.14**	**86.30**	**85.59**	**87.07**	**86.83**	**85.61**	**86.52**	**86.68**	**85.55**
	Round 1	86.61	86.00	85.33	85.89	85.84	85.04	86.43	86.20	85.03	85.82	85.73	84.89
	Round 2	86.29	85.43	84.96	85.63	85.48	84.74	85.99	85.59	84.62	85.44	85.11	84.50
	Round 3	85.98	84.93	84.68	85.36	85.15	84.57	85.65	85.14	84.47	85.16	84.68	84.20
FKGL ↓	Initial Summary	17.73	18.74	19.12	17.79	19.95	19.81	17.29	18.04	18.07	18.60	21.13	20.88
	Round 1	13.37	16.55	14.12	13.92	17.51	15.51	12.64	15.66	13.08	12.57	16.88	14.48
	Round 2	10.94	15.21	12.46	11.58	15.98	13.72	10.39	14.60	11.16	10.27	14.62	12.38
	Round 3	**9.65**	**14.20**	**11.42**	**10.32**	**14.94**	**12.64**	**9.30**	**13.44**	**10.66**	**9.21**	**13.28**	**11.24**
DCRS ↓	Initial Summary	14.25	13.69	15.04	13.33	13.43	14.24	13.11	12.77	14.14	13.76	13.36	15.10
	Round 1	13.21	12.67	13.61	12.49	12.51	13.31	12.06	11.88	12.68	12.31	12.14	13.32
	Round 2	12.56	12.06	13.15	11.96	11.96	12.90	11.50	11.40	12.24	11.70	11.47	12.74
	Round 3	**12.20**	**11.61**	**12.88**	**11.68**	**11.56**	**12.66**	**11.17**	**11.00**	**12.06**	**11.37**	**11.04**	**12.45**
CLI ↓	Initial Summary	18.72	17.05	20.86	17.77	16.60	19.38	16.56	15.13	17.38	18.32	16.45	21.52
	Round 1	15.86	14.54	16.54	15.71	15.09	17.06	14.07	13.59	14.22	14.69	13.76	16.29
	Round 2	13.96	13.20	15.14	14.20	14.23	16.01	12.56	12.62	13.09	12.84	12.42	14.71
	Round 3	**12.84**	**12.30**	**14.38**	**13.29**	**13.68**	**15.39**	**11.74**	**11.98**	**12.78**	**11.86**	**11.62**	**13.98**
LENS	Initial Summary	65.38	64.80	62.49	60.28	55.37	56.04	60.91	58.38	56.29	59.62	55.28	55.86
	Round 1	71.35	**67.21**	71.62	66.32	59.20	63.37	68.20	62.58	66.68	70.51	61.08	69.34
	Round 2	74.67	66.68	72.60	70.00	60.60	64.61	71.63	62.68	68.30	74.44	62.62	71.56
	Round 3	**75.91**	66.19	**72.70**	**71.05**	**60.97**	**65.12**	**72.61**	**63.86**	**68.44**	**75.53**	**64.64**	**71.95**
AlignScore	Initial Summary	**86.91**	**73.74**	**85.78**	**69.75**	**65.44**	**55.70**	35.49	**33.03**	24.18	**71.25**	**72.14**	**59.16**
	Round 1	79.40	65.49	61.28	66.07	60.94	54.80	36.54	31.40	28.07	65.65	63.24	52.34
	Round 2	76.22	59.97	59.64	64.77	57.46	53.78	**36.67**	29.52	28.70	63.00	58.28	51.08
	Round 3	74.11	56.53	58.48	63.45	55.60	52.94	36.50	29.19	**28.59**	61.51	55.21	50.12
SummaC	Initial Summary	55.48	49.67	44.48	42.36	43.63	38.59	34.41	33.48	30.69	53.19	54.73	42.87
	Round 1	67.12	49.17	48.88	53.14	45.70	45.38	39.41	34.51	36.02	69.53	57.95	49.87
	Round 2	72.93	50.86	53.81	61.27	47.76	49.00	41.56	34.55	38.50	73.49	58.94	54.58
	Round 3	**74.70**	**52.31**	**57.15**	**64.64**	**49.90**	**51.66**	**42.70**	**35.00**	**38.63**	**75.71**	**59.24**	**57.30**

Table 8: Performance of three models (GPT-4o, Llama-3.1-8B-Instruct, Qwen3-8B) on four benchmarks across initial and three refinement stages; best scores are bolded.

of words not found in a familiar word list, making it especially sensitive to technical vocabulary (Dale and Chall, 1948). CLI (Coleman-Liau Index) is based on character and sentence counts and is computationally efficient (Coleman and Liau, 1975). However, traditional metrics like FKGL have been criticized as unsuitable for text simplification tasks (Tanprasert and Kauchak, 2021). To address these limitations, we additionally employ LENS (Maddela et al., 2022), a learned metric trained on human-labeled simplifications, which showed the highest correlation with human judgments in the SIMPEVAL 2022 benchmark.

Factuality is evaluated using SummaC Conv, a lightweight natural language inference (NLI)-based model that assesses the consistency between each summary sentence and the source document. SummaC achieves over 74% balanced accuracy across multiple benchmark datasets while maintaining high efficiency (Fabbri et al., 2021). AlignScore-large is excluded from our main results, as its improvements were limited in short or heavily simplified summaries where sentence structure differs substantially from the source (Mahapatra and Garain, 2024).

D Additional Effects from Round 3

While the main paper reports results from the first two revision rounds, we present the full Round 3 outcomes in Table 8. Round 3 yielded modest yet consistent improvements in readability and factuality across models and datasets. For instance, GPT-4o's FKGL score on PLOS improved further to 9.65, and its SummaC rose from 44.48 to 57.15. On BigPatent, LENS scores peaked in Round 3, with GPT-4o and Qwen3-8B reaching 75.53 and 71.95 respectively, marking the highest readability observed.

However, these gains often came at the expense of relevance. ROUGE-1 and BERTScore declined

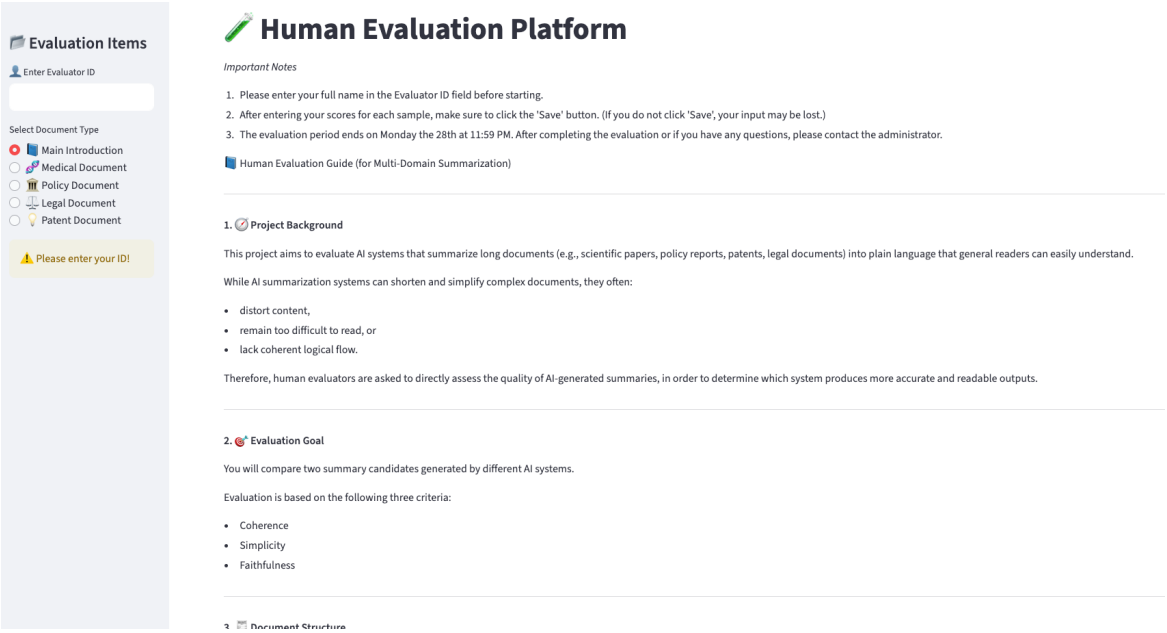

Figure 11: Overview of the web-based human evaluation platform. The main interface presents project background, evaluation goals, and detailed criteria explanations. Evaluators are guided through instructions before scoring summaries based on coherence, simplicity, and faithfulness.

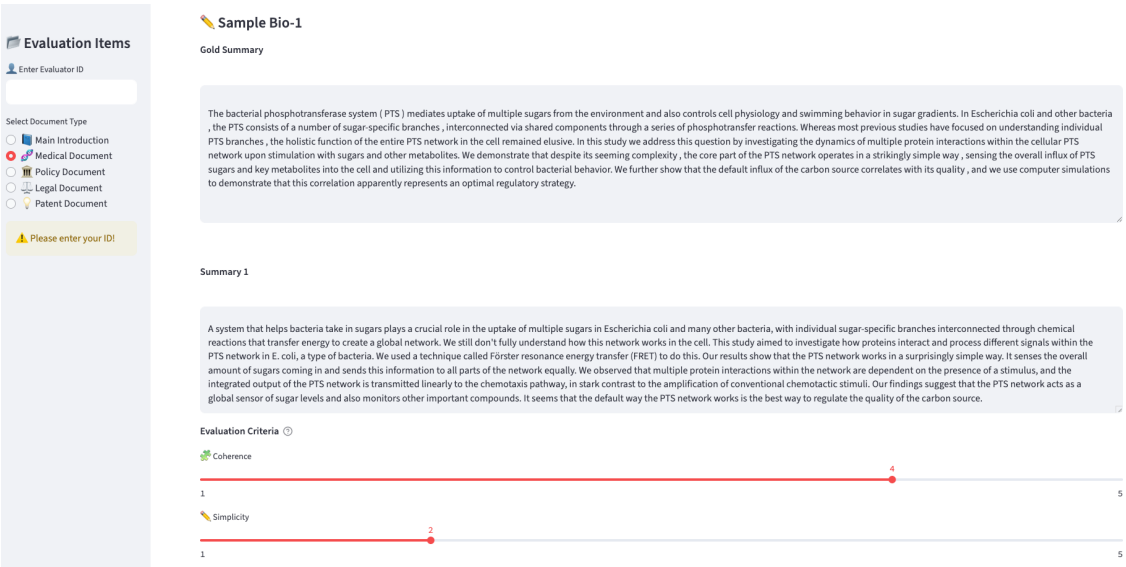

Figure 12: Example evaluation screen for a medical document. Evaluators compare two candidate summaries and assign independent scores on three criteria. A final preference is selected to reflect the overall readability and quality for a general audience.

Initial Summary

The U1 small nuclear ribonucleoprotein (snRNP)-specific U1C protein plays a crucial role in 5′ splice site recognition and regulation of pre-mRNA splicing. To investigate the role of U1C in alternative splicing regulation in the human system, we performed siRNA-mediated knockdown of U1C in HeLa cells and analyzed alternative splicing patterns by high-throughput RNA sequencing (RNA-Seq). Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly. We found that U1C knockdown leads to specific alternative splicing alterations, including increased exon skipping and changes in the usage of alternative 5′ splice sites. Furthermore, we discovered a novel feedback-loop that controls U1-70K and U1C homeostasis, linking the expression of these two U1 snRNP-specific factors via alternative splicing…

Reader Feedback - Elementary School Student Agent

"**unknown_terms**": ["nuclear ribonucleoprotein", "pre-mRNA splicing", "siRNA-mediated knockdown", "high-throughput RNA sequencing", "intra-U1 snRNP cross-regulation", "exon skipping", "stoichiometrically correct"],
"missing_contexts": ["Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly."],
"confusing_sentences": ["Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly."]

Reader Feedback - Non-Native Agent

"**unknown_terms"**: ["U1 small nuclear ribonucleoprotein", "snRNP", "pre-mRNA splicing", "siRNA-mediated knockdown", "HeLa cells", "high-throughput RNA sequencing", "RNA-Seq", "intra-U1 snRNP cross-regulation", "U1-70K", "exon skipping", "stoichiometrically correct"],
"missing_contexts": [],
"confusing_sentences": ["Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly."]

Reader Feedback - Attention-Deficit Agent

"**unknown_terms**": ["snRNP", "pre-mRNA", "siRNA", "RNA-Seq", "intra-U1 snRNP cross-regulation", "U1-70K", "stoichiometrically correct"],
"missing_contexts": ["Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly."],
"confusing_sentences": ["Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly."]

Biomedical Expert Agent's Revisions Based on Reader Agents' Feedback

{
"unknown_terms": [
{"before": "intra-U1 snRNP cross-regulation", "after": "how parts of the U1 snRNP control each other inside the complex"},
{"before": "nuclear ribonucleoprotein", "after": "nRNP-specific U1C protein"},
{"before": "siRNA-mediated knockdown", "after": "a technique that reduces the amount of a protein (siRNA-mediated knockdown)"}
],
"missing_contexts": [
{"before": "Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly.", "after": "Our results showed that U1C and U1-70K work together in a complex way to help assemble the U1 snRNP, which is a crucial step in RNA processing."}
],
"confusing_sentences": ["before": "Our results revealed a conserved, intra-U1 snRNP cross-regulation that links U1C and U1-70K expression through alternative splicing and U1 snRNP assembly", "after": "Our results showed that U1C and U1-70K work together in a complex way to help assemble how parts of the U1 snRNP control each other inside the complex.
This complex, called an snRNP, is a type of protein involved in RNA processing, and this interaction is crucial for proper RNA processing."]
}

Editor Output - Final Simplified Paragraph

The U1 small nRNP-specific U1C protein plays a crucial role in 5′ splice site recognition and regulation of pre-mRNA splicing. To investigate the role of U1C in alternative splicing regulation in the human system, we performed a technique that reduces the amount of a protein (siRNA-mediated knockdown) of U1C in HeLa cells and analyzed alternative splicing patterns by high-throughput RNA sequencing (RNA-Seq). Our results showed that U1C and U1-70K work together in a complex way to help assemble how parts of the U1 snRNP control each other inside the complex.
This complex, called an snRNP, is a type of protein involved in RNA processing, and this interaction is crucial for proper RNA processing. We found that U1C knockdown leads to specific alternative splicing alterations, including increased exon skipping and changes in the usage of alternative 5′ splice sites. Furthermore, we discovered a novel feedback-loop that controls U1-70K and U1C homeostasis, linking the expression of these two U1 snRNP-specific factors via alternative splicing…

Figure 13: NRLB revision process on a biomedical paper. Reader agents identify readability barriers, domain experts propose targeted edits, and the Editor Agent integrates them into a simplified summary. Color, underline, and highlight are used to visualize unknown terms, missing contexts, and confusing sentences, respectively.

Initial Summary

The General Services Administration (GSA) and the Office of Management and Budget (OMB) have been working to reduce the federal government's footprint through the Reduce the Footprint (RTF) policy, which aims to decrease office space usage and associated costs. To assess the effectiveness of this policy, the GAO conducted a study to evaluate the extent to which the RTF performance measures reflect changes in civilian Chief Financial Officers Act agencies' (CFO Act agencies) office space costs. Our analysis found that the RTF policy has led to a significant reduction in office space, with over 6 million square feet of space eliminated from fiscal year 2015 through fiscal year 2018. While 23 agencies reduced their office space, 7 agencies actually increased their space, with 16 agencies reducing space and 7 increasing space. The General Services Administration (GSA) reported that the federal government has saved millions of dollars due to reduced office and warehouse space, but also noted that the average cost per square foot for office space has increased. GSA's and OMB's method for estimating the cost avoidance associated with agencies' real property changes is a reasonable approach given the current limitations...

Reader Feedback - Elementary School Student Agent

"**unknown_terms:** ["General Services Administration", "Office of Management and Budget", "Reduce the Footprint", "Chief Financial Officers Act", "GAO", "fiscal year"],
"missing_contexts": ["The General Services Administration (GSA) and the Office of Management and Budget (OMB) have been working to reduce the federal government's footprint through the Reduce the Footprint (RTF) policy, which aims to decrease office space usage and associated costs"],
"confusing_sentences": ["While 23 agencies reduced their office space, 7 agencies actually increased their space, with 16 agencies reducing space and 7 increasing space."]

Reader Feedback - Non-Native Agent

"**unknown_terms**": ["footprint", "Chief Financial Officers Act", "GAO", "civic", "fiscal year", "cost avoidance", "GSA-managed space"],
"missing_contexts": [],
"confusing_sentences": ["While 23 agencies reduced their office space, 7 agencies actually increased their space, with 16 agencies reducing space and 7 increasing space."]

Reader Feedback - Attention-Deficit Agent

"**unknown_terms**": ["footprint", "GAO"],
"missing_contexts": [],
"confusing_sentences": []

Policy Expert Agent's Revisions Based on Reader Agents' Feedback

{
"**unknown_terms**": [
{"before": "GAO", "after": "Government Accountability Office (GAO)"},
{"before": "fiscal year", "after": "budget year"},
{"before": "footprint", "after": "physical presence"}
],
"**missing_contexts**": [
{"before": "The General Services Administration (GSA) and the Office of Management and Budget (OMB) have been working to reduce the federal government's footprint through the Reduce the Footprint (RTF) policy, which aims to decrease office space usage and associated costs", "after": "The federal government is the largest employer in the United States, with many offices and buildings across the country. The General Services Administration (GSA) and the Office of Management and Budget (OMB) have been working to reduce the federal government's physical presence through the Reduce the Footprint (RTF) policy, which aims to decrease office space usage and associated costs."}
],
"**confusing_sentences**": [
{"before": "While 23 agencies reduced their office space, 7 agencies actually increased their space, with 16 agencies reducing space and 7 increasing space.", "after": "Twenty-three agencies reduced their office space, but seven agencies actually increased their space, and 16 agencies reduced space while 7 increased it."}
]
}

Editor Output - Final Simplified Paragraph

The General Services Administration (GSA) and the Office of Management and Budget (OMB) have been working to reduce the federal government's physical presence through the Reduce the Footprint (RTF) policy, which aims to decrease office space usage and associated costs. The federal government is the largest employer in the United States, with many offices and buildings across the country. To assess the effectiveness of this policy, the Government Accountability Office (GAO) conducted a study to evaluate the extent to which the RTF performance measures reflect changes in civilian Chief Financial Officers Act agencies' (CFO Act agencies) office space costs. Our analysis found that the RTF policy has led to a significant reduction in office space, with over 6 million square feet of space eliminated from budget year 2015 through budget year 2018. Twenty-three agencies reduced their office space, but seven agencies actually increased their space, and 16 agencies reduced space while 7 increased it. The General Services Administration (GSA) reported that the federal government has saved millions of dollars due to reduced office and warehouse space, but also noted that the average cost per square foot for office space has increased. GSA's and OMB's method for estimating the cost avoidance associated with agencies' real property changes is a reasonable approach given the current limitations...

Figure 14: NRLB revision process on a policy document. Reader agents identify readability barriers, domain experts propose targeted edits, and the Editor Agent integrates them into a simplified summary. Color, underline, and highlight are used to visualize unknown terms, missing contexts, and confusing sentences, respectively.

Initial Summary

This bill, the Civil Liability Protection Act (CLPA), aims to provide liability protection for businesses that allow non-profit organizations to use their facilities, while maintaining accountability for gross negligence or intentional misconduct. The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations. The bill prohibits liability for such incidents unless the business entity's actions constitute gross negligence or intentional misconduct, including hate crimes, violent acts, or international terrorism. The bill preempts state laws that are inconsistent with its provisions, except for state laws providing additional protection from liability. The CLPA also specifies that it does not supersede federal or state health and safety laws. The CLPA modifies state laws by preempting inconsistent provisions, while allowing state laws that provide additional protection from liability to remain in effect. The bill does not amend any federal laws…

Reader Feedback - Elementary School Student Agent

"unknown_terms": ["Civil Liability Protection Act", "gross negligence", "inconsistent provisions", "preempts", "supersedes"],
"missing_contexts": ["The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations. "],
"confusing_sentences": ["The bill prohibits liability for such incidents unless the business entity's actions constitute gross negligence or intentional misconduct, including hate crimes, violent acts, or international terrorism."]

Reader Feedback - Non-Native Agent

"unknown_terms": ["gross negligence", "preempts", "preempting", "supersede"],
"missing_contexts": [],
"confusing_sentences": ["The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations."]

Reader Feedback - Attention-Deficit Agent

"unknown_terms": ["Civil Liability Protection Act", "gross negligence", "inconsistent", "supersedes", "inconsistent provisions"],
"missing_contexts": [],
"confusing_sentences": ["The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations."]

Legislative Expert Agent's Revisions Based on Reader Agents' Feedback

{
"unknown_terms": [
{"before": "inconsistent", "after": "not in line with this law"},
{"before": "inconsistent provisions", "after": "provisions that do not match this law"},
{"before": "supersedes", "after": "takes precedence over"}
],
"missing_contexts": [
{"before": "The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations.", "after": "This means that businesses can allow non-profit organizations to use their facilities without worrying about being sued, as long as they are not being extremely careless or intentionally causing harm."}],
"confusing_sentences": [
{"before": "The bill prohibits liability for such incidents unless the business entity's actions constitute gross negligence or intentional misconduct, including hate crimes, violent acts, or international terrorism.", "after": "The bill does not hold businesses responsible for accidents unless they were extremely careless or did something on purpose, such as committing a hate crime or violent act."},
{"before": "The CLPA authorizes businesses to shield themselves from civil liability for injuries or deaths occurring at their facilities when used by non-profit organizations outside the scope of their business operations.", "after": "Businesses can protect themselves from being sued if someone gets hurt at their facility when a non-profit organization is using it for something not related to the business."}
]
}

Editor Output - Final Simplified Paragraph

This bill, the Civil Liability Protection Act (CLPA), aims to provide liability protection for businesses that allow non-profit organizations to use their facilities, while maintaining accountability for gross negligence or intentional misconduct. Businesses can protect themselves from being sued if someone gets hurt at their facility when a non-profit organization is using it for something not related to the business. This means that businesses can allow non-profit organizations to use their facilities without worrying about being sued, as long as they are not being extremely careless or intentionally causing harm. The bill does not hold businesses responsible for accidents unless they were extremely careless or did something on purpose, such as committing a hate crime or violent act. The bill preempts state laws that are not in line with its provisions, except for state laws providing additional protection from liability. The CLPA also specifies that it does not take precedence over federal or state health and safety laws. The CLPA modifies state laws by preempting provisions that do not match this law, while allowing state laws that provide additional protection from liability to remain in effect. The bill does not amend any federal laws…

Figure 15: NRLB revision process on a legislative bill. Reader agents identify readability barriers, domain experts propose targeted edits, and the Editor Agent integrates them into a simplified summary. Color, underline, and highlight are used to visualize unknown terms, missing contexts, and confusing sentences, respectively.

Initial Summary

The present invention relates to an energetically efficient device for moving debris on a surface, such as leaves on a lawn, by addressing the need for an efficient and cost-effective method for removing debris from lawns in an environmentally friendly manner. The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered exergetically from a solar cell array. The key features of the invention include a blower and rotary rake mounted on a motor-driven chassis, a solar cell array for exergetic power, a controller for regulating power and navigation, and a location sensor for navigation, all working together to efficiently and effectively remove debris from lawns, particularly leaves, using a combination of air flow and mechanical action.

Reader Feedback - Elementary School Student Agent

"**unknown_terms**": ["energetically", "exergetically", "exergetic"],
"**missing_contexts**": ["["The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered exergetically from a solar cell array."],
"**confusing_sentences**": ["The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered exergetically from a solar cell array."]

Reader Feedback - Non-Native Agent

"**unknown_terms**": ["energetically", "exergetically", "exergetic"],
"**missing_contexts**": [],
"**confusing_sentences**": []

Reader Feedback - Attention-Deficit Agent

"**unknown_terms**": ["energetically", "exergetically", "exergetic"],
"**missing_contexts**": [],
"**confusing_sentences**": []

Patent Expert Agent's Revision Based on Reader Agents' Feedback

{
"**unknown_terms**": [
{"before": "energetically", "after": "efficiently"},
{"before": "exergetically", "after": "using renewable energy from a solar panel"},
{"before": "exergetic", "after": "energy-producing"}
],
"**missing_contexts**": [
{"before": "The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered exergetically from a solar cell array.",
"after": "The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered using renewable energy from a solar panel."}
],
"**confusing_sentences**": [
{"before": "The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered exergetically from a solar cell array.",
"after": "The invention uses a combination of an air blower and a tilted rotary rake to remove debris. The blower and rake are mounted on a motor-driven chassis that can be powered by a solar panel."}
]
}

Editor Output - Final Simplified Paragraph

The present invention relates to an efficiently efficient device for moving debris on a surface, such as leaves on a lawn, by addressing the need for an efficient and cost-effective method for removing debris from lawns in an environmentally friendly manner. The invention resolves this problem by combining an air blower with a tilted rotary rake to dislodge and impel debris, and providing a motor-driven chassis to transport the blower and rake across the lawn, which can be powered using renewable energy from a solar panel. The key features of the invention include a blower and rotary rake mounted on a motor-driven chassis, a solar cell array for energy-producing power, a controller for regulating power and navigation, and a location sensor for navigation, all working together to efficiently and effectively remove debris from lawns, particularly leaves, using a combination of air flow and mechanical action.

Figure 16: NRLB revision process on a patent document. Reader agents identify readability barriers, domain experts propose targeted edits, and the Editor Agent integrates them into a simplified summary. Color, underline, and highlight are used to visualize unknown terms, missing contexts, and confusing sentences, respectively.

Table 9: Evaluation results across datasets and checklist lengths (K=1,2,3,6,9).

# K	Metric	PLOS	GovReport	BillSum	BigPatent
1	ROUGE1	42.94	44.90	42.79	39.91
	FKGL ↓	16.95	17.92	15.66	16.96
	DCRS ↓	12.58	12.53	11.87	12.21
	CLI ↓	14.39	15.43	13.49	14.08
	LENS	65.28	59.29	62.46	61.93
	SummaC	49.65	45.06	33.99	57.29
2	ROUGE1	44.31	45.21	43.34	40.21
	FKGL ↓	16.52	17.63	15.59	16.95
	DCRS ↓	12.67	12.59	11.94	12.29
	CLI ↓	14.60	15.34	13.63	14.28
	LENS	67.53	59.06	62.35	61.55
	SummaC	49.42	46.23	33.97	57.60
3	ROUGE1	41.64	44.19	39.70	36.87
	FKGL ↓	15.21	15.98	14.60	14.62
	DCRS ↓	12.06	11.96	11.40	11.47
	CLI ↓	13.20	14.23	12.62	12.42
	LENS	66.68	60.60	62.68	62.62
	SummaC	50.86	47.76	34.55	58.94
6	ROUGE1	39.62	43.36	38.86	35.22
	FKGL ↓	14.99	15.93	13.78	15.12
	DCRS ↓	11.83	11.90	11.09	11.41
	CLI ↓	12.63	14.15	12.03	12.33
	LENS	63.57	59.16	61.51	61.63
	SummaC	52.42	49.20	34.31	58.23
9	ROUGE1	42.39	45.59	40.42	36.33
	FKGL ↓	15.40	16.09	15.00	14.68
	DCRS ↓	12.06	12.01	11.68	11.51
	CLI ↓	13.19	14.29	13.09	12.93
	LENS ↓	67.64	59.08	62.63	64.40
	SummaC	51.44	46.56	33.25	57.67

in most cases, particularly for Llama and Qwen. On GovReport, Qwen's ROUGE-1 dropped from 46.19 to 39.21, and BERTScore decreased by more than one point. These trends suggest that continued rewriting may over-simplify or rephrase key expressions, reducing lexical and semantic alignment with the reference.

Interestingly, models optimized for reasoning, such as Qwen3-8B, continued to show sharper gains in readability and factuality during Round 3, whereas GPT-4o maintained a more balanced tradeoff across all metrics. These findings support the optional use of a third revision when maximum clarity is prioritized over strict fidelity. We therefore adopt Round 2 as the default, as it offers the most stable balance across relevance, readability, and factuality.

E Human Evaluation Details

The human evaluation was conducted with two primary participant groups. All evaluators provided informed consent prior to participation. The evaluation was carried out through a Streamlit-based platform, which included detailed information about the study background, evaluation objectives, summary formats, scoring criteria, and privacy policy. Non-Native evaluators participated as unpaid volunteers, while recruiting Elementary School Student evaluators was more challenging, so they received a small financial compensation of approximately $3.7 per document.

The first group consisted of **Non-Native** readers, represented by three undergraduate students with moderate English proficiency. Each evaluator assessed a total of 80 randomly selected documents, comparing two summaries per document. They rated each summary independently on three criteria—coherence, simplicity, and faithfulness—using a five-point scale. In addition, they indicated their overall preference by selecting which of the two summaries they considered more readable and useful for a general audience.

The second group consisted of **Elementary School Students**, represented by three participants with limited English proficiency (two fifth graders and one sixth grader). As an ethical measure, these participants were assigned a filtered subset of summaries covering familiar or educationally accessible topics (e.g., general science or daily-life contexts), since the original materials were considered too complex for their age group. Given their age and reading level, the number of evaluation items was also intentionally limited to reduce fatigue and maintain engagement. Each evaluator reviewed 10 documents from this subset, providing ratings on the same three criteria and selecting their preferred summary in each case. Due to ethical considerations, we did not conduct evaluations with the **Attention Deficit** group.

To further assess annotator consistency, we computed Krippendorff's α within each group. The results showed low agreement overall (mostly between –0.3 and 0.5), with relatively higher consistency only on the Simplification dimension for the Elementary group (α = 0.52). Such low inter-annotator agreement is not unusual in summarization tasks, especially with only three evaluators and inherently subjective criteria. The difference in language proficiency between non-native undergraduates and elementary-level readers likely increased variability in judgment, and given this heterogeneous evaluator composition and the subjective nature of the task, achieving high inter-rater agreement values such as Krippendorff's α is inherently challenging (Fabbri et al., 2021). Nevertheless, an analysis of the standard deviations across evaluators revealed that their ratings followed broadly similar tendencies, with average deviations of 0.40 for non-native readers and 0.54 for

elementary school student readers (ranging from 0.30–0.69 across datasets), indicating substantial consistency in practice. In other words, even if the numerical agreement metrics appear low, the evaluators largely shared similar intuitions about which summaries were clearer and more faithful.

Figure 11 and Figure 12 show the evaluation platform used in the study, which offered a structured and user-friendly interface for the entire evaluation workflow, including both summary comparison and preference selection.

F Case Study

To illustrate how the NRLB framework operates across different document types, we present four representative case studies selected from the PLOS, GovReport, BillSum, and BigPatent datasets. Figures 13 through 16 provide a step-by-step visualization of the full revision pipeline applied to one summary from each domain. These examples highlight how NRLB identifies and addresses readability barriers specific to each domain.

Each case begins with the generation of an initial summary, which is then evaluated by three Agents representing an Elementary School Student Reader, a Non-Native reader, and an Attention-Deficit reader. These agents independently annotate problematic expressions based on their simulated reading experiences. Feedback is categorized into unknown terms, missing contextual information, and confusing sentence structures. These issues often involve technical vocabulary, implicit references, or overly complex syntax that can hinder comprehension.

Based on this feedback, the Domain Expert Agent proposes targeted revisions in a clear before and after format. Suggested edits may involve replacing technical terms with simpler alternatives, adding brief explanations for unfamiliar entities, or restructuring convoluted sentences. Each edit is explicitly linked to a specific comprehension barrier identified by the Reader Agents, ensuring that the revisions address real accessibility challenges rather than arbitrary simplification.

The Editor Agent synthesizes the proposed edits into a revised summary that improves both clarity and factual consistency. The visualization highlights the original problem areas using colors for unknown terms, underlines for missing contexts, and highlights for confusing structures. Among the four domains, PLOS summaries required an additional round of revision due to the density of biomedical terminology. In contrast, summaries from GovReport, BillSum, and BigPatent typically achieved satisfactory readability within one or two rounds. These results demonstrate how the NRLB framework supports adaptive, agent-guided revision tailored to domain-specific complexity and reader needs.