ASIM: An Iterative Self-Correcting Agent Framework for Medical Text Simplification

Ziming Yang
University at Buffalo
zimingya@buffalo.edu

Jinjun Xiong
University at Buffalo
jinjun@buffalo.edu

Abstract

Medical text simplification is crucial for democratizing healthcare information. We present ASIM (Adaptive Self-correcting Iterative Medical simplifier), an agent-based framework employing iterative self-correction and retrieval-augmented generation to make medical reports accessible to non-specialists. Our system integrates five autonomous agents: terminology extraction, planning, knowledge-enhanced generation, self-correction, and memory-based learning. Unlike static approaches, ASIM learns from both successful and failed attempts through a four-stage evaluation framework. Experimental results show ASIM achieves 79.70% human preference compared to 46.70% for context-aware and 43.50% for optimization-based baselines, while maintaining competitive readability metrics. The iterative correction mechanism significantly improves comprehension for non-experts while preserving medical accuracy.

1 Introduction

Over 80% of American adults search for health information online (Fox and Duggan, 2013; Rainie and Fox, 2000), yet nearly half have inadequate health literacy skills (Kutner et al., 2006). This health literacy crisis has profound implications: medical texts typically require college-level reading proficiency while the average American adult reads at an eighth-grade level (Joseph et al., 2023). Poor health literacy is associated with worse health outcomes, increased hospital readmissions, and higher healthcare costs (Schillinger et al., 2002), creating substantial barriers to effective patient-provider communication and informed medical decision-making.

Previous approaches focused on either preserving contextual coherence (Cripwell et al., 2023) or optimizing readability (Flores et al., 2023), but suffer from fundamental limitations: context-aware methods retain excessive terminology while optimization-based methods sacrifice content relevance. Existing systems lack iterative refinement capabilities essential for medical applications.

We propose ASIM, an agent-based framework treating medical text simplification as an iterative problem-solving task. ASIM provides contextual explanations and analogies through multiple specialized agents collaborating via structured interaction loops, learning from both successful and failed attempts.

Our contributions: (1) the first agent-based framework for medical text simplification with iterative self-correction; (2) a four-stage evaluation system enabling fine-grained error analysis; (3) significant outperformance of existing methods in human evaluation while maintaining competitive readability metrics.

2 Related Work

2.1 Medical Text Simplification and Agent-Based NLP

(Cripwell et al., 2023) introduced ConBART for context-aware document simplification, effective on general datasets but retaining excessive medical terminology. (Flores et al., 2023) proposed unlikelihood training with readability optimization, achieving improved scores but risking over-simplification compromising completeness and accuracy. Traditional neural approaches operate as static translation models without feedback loops, limiting their ability to handle the nuanced balance between accessibility and medical precision required for healthcare communication.

LLM-based agents (Xi et al., 2023; Yao et al., 2023) and retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2023) have shown promise for complex reasoning and knowledge-intensive tasks. However, agent frameworks for medical text simplification remain largely unexplored. Our work bridges this gap by integrating domain knowl-

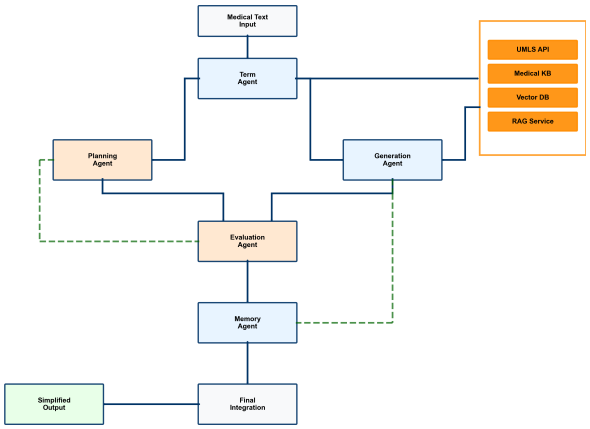

Figure 1: ASIM System Architecture with five-agent framework and iterative self-correction mechanisms.

edge through RAG and enabling iterative improvement through memory-driven learning.

2.2 Self-Correction in Language Models

Self-correction mechanisms (Madaan et al., 2023; Wang et al., 2023) enable iterative refinement but face limitations in specialized domains, lacking domain-specific knowledge integration and evaluation criteria. Our framework addresses these limitations through external knowledge sources, specialized evaluation for medical text quality, and memory-driven learning.

3 Methodology

3.1 System Architecture Overview

ASIM decomposes medical text simplification into specialized subtasks handled by five autonomous agents collaborating through structured protocols (Figure 1). The system operates through iterative cycles with dynamic interactions enabling mid-process corrections and adaptive strategy adjustment.

3.2 Agent Framework

Our framework integrates five specialized agents, each contributing distinct expertise to the simplification process through carefully designed interaction protocols:

Term Agent operates through a multi-stage medical terminology processing pipeline. First, it extracts medical concepts using domain-specific prompts tuned for anatomical terms, clinical procedures, pharmaceutical names, and diagnostic terminology. It then scores complexity using SUBTLEX-US word frequency norms, UMLS semantic type classification, and syllable count. The agent prioritizes terms for simplification by combining complexity scores with term centrality to document meaning. Finally, it queries UMLS API for standardized definitions, synonyms, and lay language alternatives, outputting an enriched terminology list with complexity annotations.

Planning Agent develops comprehensive simplification strategies tailored to specific content types and user requirements. It analyzes text structure to identify logical segments, determines appropriate simplification granularity (sentence-level vs. paragraph-level), and formulates step-by-step plans balancing accessibility with completeness. The planning process incorporates complexity assessment, strategy selection (explanation-focused, analogy-based, or restructuring-oriented), resource allocation determining which concepts require RAG-enhanced explanation versus simple lexical substitution, and quality constraints establishing thresholds for acceptable information loss and readability improvements.

Generation Agent executes simplification plans using retrieval-augmented generation to access comprehensive medical knowledge bases. It queries vector databases for relevant medical definitions, previously successful explanations, and effective analogies. The agent integrates retrieved knowledge with source text context to generate accurate, coherent explanations, adjusting language complexity to match 8th-9th grade reading levels while preserving medical precision. The generation process emphasizes contextual explanations and accessible analogies over simple term substitution, enabling genuine comprehension.

Evaluation Agent assesses output quality through multi-dimensional analysis. It verifies medical accuracy via comparison with authoritative knowledge base definitions, computes readability metrics (FKGL, FRE, GFI), checks for preservation of essential medical information and critical details, and analyzes logical flow and explanation adequacy. The agent implements the four-stage evaluation framework (PCAC/PCAI/PIAC/PIAI) to provide structured feedback categorizing both plan quality and execution effectiveness, enabling targeted corrections in subsequent iterations.

Memory Agent maintains persistent storage of simplification patterns to enable continuous learning. It stores successful strategies indexed by medical concept, text type, and complexity level for reuse in similar future tasks. The agent records

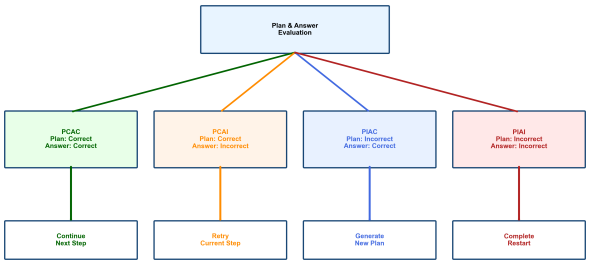

Figure 2: Four-stage self-correction evaluation mechanism showing decision paths for different plan and answer quality combinations.

failure patterns with contextual markers indicating why particular approaches failed, preventing repeated errors. It tracks correction histories showing which strategies (PCAI/PIAC/PIAI) proved effective for different failure modes, and records performance metrics including iteration counts, success rates, and convergence patterns across medical specialties. Memory retrieval prioritizes patterns matching current medical specialty, term complexity, and text structure, reducing iteration counts by approximately 30%.

The agent interaction follows a structured five-phase pipeline: (1) Term Agent extracts and scores medical terminology with UMLS enrichment; (2) Planning Agent develops simplification strategies with explicit success criteria; (3) Generation Agent executes plans using RAG-enhanced knowledge retrieval; (4) Evaluation Agent assesses outputs via the four-stage framework and triggers corrections; (5) Memory Agent stores patterns for future reuse. The process iterates for up to three cycles, selecting the highest-quality output based on comprehensive evaluation.

3.3 Iterative Self-Correction Protocol

ASIM implements a four-stage evaluation framework that assesses both planning quality and execution effectiveness, enabling fine-grained error analysis and targeted correction strategies. The framework evaluates four key aspects: Plan Correct/Answer Correct (PCAC), Plan Correct/Answer Incorrect (PCAI), Plan Incorrect/Answer Correct (PIAC), and Plan Incorrect/Answer Incorrect (PIAI).

PCAC (Plan Correct, Answer Correct): Both strategic planning and execution meet quality thresholds across all evaluation dimensions—medical accuracy, readability improvement, information preservation, and logical coherence. The system proceeds to complete the task, storing success patterns in memory for future reuse with similar medical concepts.

PCAI (Plan Correct, Answer Incorrect): The strategic approach is sound, but execution fails quality standards. This occurs when generated explanations are too technical, factually imprecise, or poorly structured despite correct planning. The system retains the successful planning strategy while regenerating the answer with enhanced prompting and Evaluation Agent feedback. This targeted approach reduces computational overhead by avoiding unnecessary plan regeneration.

PIAC (Plan Incorrect, Answer Correct): Execution produces acceptable results despite flawed strategic planning. This occurs when the Generation Agent successfully self-corrects during execution through effective RAG resource use. The system generates a new plan aligning with the successful execution approach to maintain consistency for future similar tasks.

PIAI (Plan Incorrect, Answer Incorrect): Both planning and execution require comprehensive revision. This indicates fundamental misunderstanding of medical content or systematic errors in the simplification approach. The system performs complete regeneration incorporating Memory Agent insights to avoid repeating similar failures.

This granular evaluation enables targeted corrections while minimizing computational waste. Each correction cycle updates memory with success/failure patterns, correction strategies, and context metadata including medical specialty, text complexity, and terminology density.

3.4 Knowledge Integration Architecture

ASIM integrates comprehensive medical knowledge through sophisticated retrieval-augmented generation mechanisms ensuring both accuracy and accessibility. The system accesses multiple authoritative medical knowledge bases: SMQ (Standardized MedDRA Queries) for standardized medical terminology in regulatory contexts, MeSH (Medical Subject Headings) for hierarchical medical concept organization, and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) for comprehensive clinical terminology coverage across diverse healthcare specialties.

The knowledge integration architecture employs a multi-layered approach combining structured medical ontologies with contextual retrieval mechanisms. UMLS (Unified Medical Language Sys-

tem) serves as the primary integration point, providing concept unique identifiers (CUIs) that enable consistent terminology mapping across different knowledge sources. This integration ensures simplifications maintain medical accuracy while providing appropriate explanations for complex concepts.

The system's knowledge base integration strategy operates through several key mechanisms: (1) **Term Normalization** maps medical terms to UMLS CUIs ensuring consistent representation across varying terminological forms; (2) **Hierarchical Relationship Extraction** uses MeSH hierarchies to identify broader and narrower concepts for generating appropriate explanations at various detail levels; (3) **Cross-Reference Validation** employs SNOMED CT to validate terminology accuracy and identify potential synonyms; (4) **Context-Aware Retrieval** selects knowledge sources based on medical specialty context, prioritizing SMQ for pharmacological content and SNOMED CT for clinical procedures.

Vector database implementation using FAISS v1.7.4 enables efficient similarity-based retrieval of relevant medical concepts and simplified explanations. Documents are processed using RecursiveCharacterTextSplitter with 300-character chunks and 100-character overlap, then vectorized using OpenAI's text-embedding-3-large model. The system maintains separate embedding spaces for medical terms and authoritative definitions, previously successful simplifications indexed by complexity and target audience, and analogies that proved effective in past tasks. Retrieval strategies incorporate both semantic similarity via cosine distance and domain-specific relevance scoring based on medical specialty alignment and historical success rates.

4 Experimental Setup

4.1 Implementation Details

ASIM is implemented in Python using Langchain v0.1.0 with GPT-4-turbo-preview as the backbone LLM for all agents, configured with temperature=0.3 and max_tokens=2000 for balanced creativity and consistency. The system integrates multiple medical knowledge bases through UMLS API providing comprehensive terminology coverage. Vector databases are implemented using FAISS v1.7.4 for efficient similarity search across medical concepts, with ChromaDB serving as the persistent storage layer.

The iterative correction mechanism operates with a maximum of three cycles per input, balancing thoroughness with computational efficiency. Each cycle involves full plan-answer evaluation, with the system automatically terminating when quality thresholds are met or maximum iterations reached. Text embeddings use OpenAI's text-embedding-3-large model for semantic similarity computation in the RAG retrieval process. Agent interaction protocols are implemented through structured message passing with JSON-formatted communications to ensure reliable inter-agent coordination.

4.2 Dataset and Evaluation

Our evaluation consists of two components: **Automated readability assessment** uses the publicly available MeDAL dataset (Wen et al., 2020) from HuggingFace (McGill-NLP/medal), a large medical corpus of approximately 14 million PubMed abstracts originally designed for medical abbreviation disambiguation. We randomly sampled 500 abstracts for automated readability metric evaluation using eight standard readability formulas. **Human evaluation** employs medical documents spanning diverse clinical specialties including cardiology, endocrinology, and psychiatry, with 68 participants (mean age 58.33, 65.2% female, 69.6% non-White) providing 205 preference judgments across multiple simplification approaches. All participants provided informed consent and received $10 gift cards; data was anonymized following institutional ethics protocols.

Human evaluation employs pairwise comparison methodology where participants rate simplifications based on comprehensibility, accuracy preservation, and overall preference. We employ eight standard readability metrics (detailed in Table 1) to assess text complexity from multiple linguistic perspectives.

4.3 Baseline Methods

We compare against two state-of-the-art approaches representing different paradigms in medical text simplification:

ConBART (Context-Aware Simplification) (Cripwell et al., 2023): A BART-based architecture with context attention layers for document-level simplification, representing state-of-the-art context-aware approaches. **UL+Decoder (Optimization-Based Simplification)** (Flores et al., 2023): Em-

ploys unlikelihood training with beam search reranking to optimize readability metrics, designed specifically for medical text simplification. Both baselines were evaluated under identical experimental conditions.

5 Results and Analysis

5.1 Readability Metrics Analysis

Table 1 presents comprehensive readability analysis across eight standard metrics. ASIM achieves competitive performance with FKGL of 8.36 and GFI of 10.35, comparable to ConBART's 7.99 and 10.78 respectively, indicating similar grade-level complexity suitable for general audiences. These scores place ASIM's outputs at approximately 8th-9th grade reading level, appropriate for health literacy applications targeting general populations.

Notably, ASIM excels in metrics emphasizing balanced simplification. The GFI score of 10.35 (best among all methods) and DCR score of 8.37 demonstrate effective vocabulary selection and sentence structure management. ASIM maintains readability improvements while preserving content through contextual explanations rather than simple term substitution.

UL+Decoder's extreme scores (FKGL: 39.36, FRE: -28.87) indicate problematic oversimplification. The negative Flesch Reading Ease score and grade level exceeding college education suggest the metric-driven optimization compromises text quality. ConBART maintains better performance across most metrics but shows limitations in certain dimensions, reflecting its conservative approach to medical term modification.

However, traditional metrics focus on surface features (sentence length, syllable count) and cannot capture semantic complexity or genuine comprehension. This motivates our comprehensive human evaluation with actual target demographic participants.

5.2 Human Evaluation: The Ultimate Test of Readability

We recruited 68 participants from four diverse community locations (Canopy of Neighbors, Conventus/UBMD Outpatient, Hopewell Baptist Church, and UB On The Green), ensuring representation of the target demographic who encounter medical texts but lack specialized training. Each participant completed 3 randomly assigned surveys, each containing an original medical text paired with one simplified version. The evaluation protocol used 21 half-page medical texts (7 informed consent documents, 6 clinical documents, 4 research abstracts, 4 health websites), each simplified by three methods, generating 63 versions total.

Table 2 presents human evaluation results from 68 participants (69.6% non-White, 65.2% female, mean age 58.33, SD 21.5, range 18-90) evaluating medical texts on 5-point Likert scales (1=poor, 5=excellent).

ASIM achieves 79.70% preference rate compared to 46.70% for ConBART and 43.50% for UL+Decoder, representing a 70% improvement over the strongest baseline. This substantial preference gap highlights the effectiveness of our agent-based approach in producing genuinely accessible medical texts. ASIM demonstrates superior performance across all evaluation dimensions, with particularly strong advantages in completeness (4.07 vs 2.89/2.66) and clarity (4.34 vs 3.83/3.89).

The results reveal critical weaknesses in baseline approaches. ConBART's context-aware sentence-level simplification struggles to capture global document coherence and strategic simplification priorities, resulting in poor completeness scores (2.89/5) as essential medical information is lost during localized optimization. The model preserves technical terminology that limits accessibility, achieving moderate clarity scores (3.83/5) reflecting its conservative approach.

UL+Decoder's optimization-based approach with unlikelihood training achieves slightly better readability scores (4.20/5) through aggressive simplification targeting Flesch-Kincaid metrics. However, this metric-driven optimization performs worse on completeness (2.66/5), indicating systematic information loss. The unlikelihood training successfully reduces complex terminology but fails to replace removed concepts with adequate explanations, creating oversimplified texts that sacrifice essential medical details.

In contrast, ASIM's multi-agent approach with iterative refinement maintains high completeness (4.07/5) while achieving superior readability (4.34/5) and clarity (4.34/5). This demonstrates the framework's ability to balance accessibility with medical accuracy through strategic planning, contextual explanation generation, and multi-dimensional quality assessment. The evaluation-correction loop enables detection and correction of both incompleteness and complexity issues. F Statistical significance analysis confirms robust results

Method	FKGL	FRE	GFI	DCR	ARI	SPACHE	LW	CLI
ConBART	**7.99**	**64.56**	10.78	8.38	**7.67**	**5.92**	8.69	**8.61**
UL+Decoder	39.36	-28.87	44.03	14.37	47.97	17.77	**65.89**	12.47
ASIM	8.36	60.88	**10.35**	**8.37**	8.79	6.08	9.01	10.52

Table 1: Readability metrics comparison across methods. FKGL: Flesch-Kincaid Grade Level, FRE: Flesch Reading Ease, GFI: Gunning Fog Index, DCR: Dale-Chall Readability, ARI: Automated Readability Index, SPACHE: Spache Readability Formula, LW: Linsear Write, CLI: Coleman-Liau Index. Bold values indicate best performance per metric.

Evaluation Dimension	ConBART	UL+Decoder	ASIM
Readability	4.11	4.20	**4.34**
Clarity	3.83	3.89	**4.34**
Completeness	2.89	2.66	**4.07**
Overall Score	3.41	3.27	**4.28**
Preference (%)	46.70	43.50	**79.70**
Sample Size	65	72	68

Table 2: Human evaluation results across assessment dimensions (5-point scale except preference rate). ASIM significantly outperforms baseline methods.

with $p < 0.001$ for all pairwise comparisons across all evaluation dimensions. The large effect size (Cohen's $d > 0.8$) indicates practical significance beyond statistical significance. Subgroup analysis by text type reveals consistent ASIM superiority across informed consent documents, clinical notes, research abstracts, and health information materials, demonstrating generalization across diverse medical text genres.

5.2.1 Demographic Analysis

ASIM's preference advantage (79.70% vs. 46.70%) represents a 33 percentage point improvement (95% CI: [28.1%, 37.9%], $p<0.001$, Cohen's $h = 0.72$). Likert scale improvements: readability +0.23 ($p<0.01$), clarity +0.51 ($p<0.001$), completeness +1.18 ($p<0.001$). Performance remained strong across participants with high school education or less (n=13, 19.4%), age groups (18-90 years), and racial/ethnic groups (69.6% non-White), validating broad applicability for health equity.

5.3 Self-Correction Mechanism Performance

Analysis of ASIM's iterative correction behavior reveals efficient convergence patterns contributing to superior outcomes. The four-stage evaluation protocol (PCAC, PCAI, PIAC, PIAI) enables targeted improvements without complete regeneration, providing substantial efficiency advantages.

The most common correction pattern involves Plan_correct_Answer_incorrect scenarios (PCAI), where sound planning enables focused answer improvement without plan regeneration. This targeted approach reduces computational overhead while maximizing quality improvements. The memory component demonstrates increasing effectiveness over processing volume, with pattern recognition improving simplification consistency. Analysis reveals that successful simplification strategies stored in memory reduce iteration counts by approximately 30% for similar medical concepts encountered subsequently.

Most corrections converge within 2-3 iterations, demonstrating the evaluation agent's effective feedback mechanisms. The system achieves PCAC (both plan and answer correct) status in 67% of cases after first iteration, rising to 91% by third iteration. This rapid convergence contributes to ASIM's practical feasibility for real-time medical communication applications, with average processing time of 45 seconds per medical text paragraph.

5.4 Qualitative Analysis

Manual inspection of ASIM outputs reveals several key strengths distinguishing the system from baseline approaches.

Analogical Reasoning: ASIM consistently generates effective analogies connecting abstract medical concepts to familiar everyday experiences. For example, antioxidants are described as "the body's cleanup crew that removes harmful waste," while blood circulation is explained through city traffic flow metaphors. This analogical approach makes complex biomedical concepts genuinely understandable for high school-educated readers, contrasting with ConBART's preserved technical terminology and UL+Decoder's simple term deletion.

Structured Explanations: The step-based planning produces logically organized explanations progressing from basic concepts to specific details. Medical processes are broken down into sequential steps with clear causal relationships, helping non-expert readers follow complex physiological mechanisms systematically.

Contextual Terminology Management: Rather

than simple lexical substitution, ASIM provides contextual explanations. For instance, "myocardial infarction" is explained as "heart attack - when blood flow to part of the heart muscle is blocked, causing tissue damage," preserving the technical term while ensuring comprehension. This enables readers to understand professional medical reports.

Content Preservation with Accessibility: ASIM maintains essential medical information while transforming presentation. Critical details about dosages, side effects, contraindications, and treatment protocols are retained with clarifying explanations rather than deletion, explaining the superior completeness scores (4.07 vs. 2.89/2.66).

6 Discussion

The multi-agent architecture provides several advantages over traditional end-to-end methods for medical text simplification. The modular design enables specialized optimization for each component while maintaining system coherence through structured interaction protocols. Each agent contributes specialized capabilities: the Term Agent identifies complex terminology requiring explanation through UMLS integration, the Planning Agent structures information for systematic understanding through strategic decomposition, the Generation Agent creates analogies and contextual explanations via RAG-enhanced synthesis, the Evaluation Agent ensures practical comprehension through multi-dimensional assessment, and the Memory Agent enables continuous improvement through pattern learning.

This decomposition contrasts with monolithic end-to-end approaches where all functions are entangled within a single model. The separation of concerns enables targeted debugging, component-level optimization, and transparent error localization. When simplification quality issues arise, the agent framework supports precise diagnosis of whether problems stem from terminology identification, strategic planning, explanation generation, or quality assessment failures.

The memory-based learning component represents a paradigm shift from static to adaptive simplification systems. This capability addresses a fundamental limitation of existing approaches that cannot improve from experience or adapt to specific comprehension patterns observed across multiple texts. As the system processes more medical documents, the memory accumulates successful simplification strategies for recurring medical concepts, reducing computational overhead while improving consistency.

6.1 Theoretical Contributions

ASIM's agent-based architecture contributes three key theoretical advances to text simplification research. First, the decomposition of simplification into specialized agent roles enables targeted optimization and error localization, contrasting with monolithic approaches where all functions are entangled. This separation facilitates systematic improvement through component-level refinement.

Second, the four-stage correction protocol (PCAC/PCAI/PIAC/PIAI) provides a formal framework for iterative refinement that balances exploration (trying new strategies when plans fail) with exploitation (reusing successful patterns from memory). This framework extends beyond traditional readability metric optimization by explicitly modeling the relationship between strategic planning quality and execution effectiveness.

Third, the integration of memory-based learning introduces temporal dynamics to simplification systems, transforming them from static translation models to adaptive communication systems that improve continuously over time. This learning capability addresses a fundamental gap in existing approaches that cannot adapt to recurring medical concepts or accumulate expertise from processing multiple documents.

6.2 Practical Impact and Broader Implications

The human evaluation results (79.70% preference rate, 4.07/5 completeness score) demonstrate ASIM's potential for real-world healthcare applications. The significant improvement in information completeness (4.07 vs. 2.89/2.66 for baselines) suggests that agent-based approaches can maintain medical accuracy while improving accessibility—a critical requirement for patient education materials, informed consent documents, and health literacy initiatives.

Healthcare institutions could deploy ASIM to automatically generate patient-friendly versions of clinical documents, reducing the communication burden on medical professionals while empowering patients with comprehensible health information. The system's strong performance across diverse text types (informed consent, clinical notes, research abstracts, health websites) indicates broad

applicability across medical communication contexts.

The framework's effectiveness across diverse demographic groups, particularly among participants with high school education or less, validates its utility for health equity initiatives targeting underserved populations. The consistent preference patterns across racial/ethnic groups (69.6% non-White representation) support ASIM's potential for addressing health disparities through improved medical text accessibility.

Beyond medical simplification, the methodological contributions—specialized agent decomposition, systematic correction protocols, and memory-based learning—provide a blueprint for developing adaptive text processing systems in other specialized domains. Any field requiring careful balance between accessibility and accuracy while handling complex technical concepts could benefit from similar agent-based architectures, including legal document simplification, technical writing adaptation, and scientific communication for public understanding.

7 Conclusion

We present ASIM, the first agent-based framework for medical text simplification that employs iterative self-correction mechanisms to bridge the gap between professional medical language and high school-level comprehension. Our approach addresses fundamental limitations of existing methods by treating simplification as a collaborative problem-solving task requiring specialized expertise, strategic planning, knowledge integration, and continuous refinement.

Experimental results demonstrate ASIM's significant superiority in human evaluation, achieving a 79.70% preference rate compared to 46.70% for ConBART and 43.50% for UL+Decoder. The system excels particularly in information completeness (4.07 vs. 2.89/2.66) and clarity (4.34 vs. 3.83/3.89), demonstrating successful balance between accessibility and medical accuracy. These results, validated across diverse demographic groups including 69.6% non-White participants and individuals with varying educational backgrounds, confirm practical utility for health equity applications.

The key insight underlying ASIM's success is that medical simplification requires not just lexical substitution or syntactic restructuring, but coordinated expertise across terminology identification, strategic planning, knowledge-enhanced generation, multi-dimensional evaluation, and iterative refinement. The four-stage correction protocol (PCAC/PCAI/PIAC/PIAI) enables targeted improvements while the memory-based learning transforms simplification from a static translation problem to an adaptive communication challenge.

The multi-agent architecture with memory-driven learning represents a paradigm shift toward adaptive simplification systems that improve continuously over time. Our work opens new research directions in agent-based text processing, domain-specific simplification methodologies, and learning-enhanced communication systems.

Limitations

ASIM's reliance on GPT-4 introduces computational costs limiting accessibility for resource-constrained settings. Evaluation focuses on English medical texts, limiting multilingual generalizability. The 68-participant evaluation may not fully represent all demographic groups and health literacy levels. Performance on highly specialized subspecialty content (e.g., advanced oncology protocols, complex surgical procedures) requires additional investigation.

Future work will expand participant diversity for more robust validation across age groups, educational backgrounds, and health literacy levels. We plan to extend evaluation to specialized medical domains, investigate integration with smaller open-source LLMs to address computational efficiency, and explore cross-domain applications to legal and technical document simplification. Additionally, we will develop user-customizable simplification levels to accommodate varying reader expertise and preferences.

References

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873.

Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Pew Research Center*. Online; accessed March 28, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692.

Mark Kutner, Elizabeth Greenberg, Ying Jin, and Christine Paulsen. 2006. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. *NCES 2006-483. National Center for Education Statistics*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems*, volume 33, pages 9459–9474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Lee Rainie and Susannah Fox. 2000. Health care information and the internet. *Pew Research Center*. Online; accessed March 28, 2024.

Dean Schillinger, Kevin Grumbach, John Piette, Frances Wang, Dennis Osmond, Carolyn Daher, Jorge Palacios, Gabriela Diaz Sullivan, and Andrew B Bindman. 2002. Association of health literacy with diabetes outcomes. *JAMA*, 288(4):475–482.

Xiaoyu Wang, Can Xu, Qian Wang, Lilun Zhou, Hongyan Chen, and Zhiwei Zhang. 2023. Self-correct and refine: A general framework for self-correcting large language models. *arXiv preprint arXiv:2305.14327*.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. Medal: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135. Association for Computational Linguistics.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wang He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

A Agent Interaction Protocols

A.1 Term Agent Protocol

The Term Agent operates through a two-stage process: (1) Medical term identification using domain-specific prompts that target technical vocabulary, anatomical references, and procedural descriptions; (2) UMLS API integration for standardized definition retrieval with CUI mapping for consistency across simplifications.

A.2 Evaluation Agent Criteria

The Evaluation Agent assesses outputs using structured criteria: Plan evaluation considers logical sequence, concept breakdown, and completeness; Answer evaluation examines factual accuracy, clarity, accessibility, and connection to previous explanations. The four-stage classification enables targeted improvement strategies.

B Example Simplifications

This section provides concrete examples of ASIM's medical text simplification capabilities compared to baseline approaches, demonstrating our system's strengths in making medical content accessible to non-specialists.

B.1 Example 1: Medical Report Simplification

Original Medical Text: *Type 2 DIABETES and PRE-DIABETES*

Diabetes is a chronic disease that affects over 30 million Americans, or nearly 10% of the country's population. More than 1 in 4 people with diabetes do not know that they have the disease. People with diabetes have high blood sugar (aka glucose) levels due to the fact that insulin, a hormone that helps keeping blood sugars in check, is at low levels or does not work well. This is called Insulin Resistance. A common sign of insulin resistance is a darkening of the skin over the neck and other skin folds. Over time, high blood sugars can cause serious health issues, such as eye, kidney and heart disease.

Pre-diabetes affects 1 in 3 adults in the U.S. As the word implies, pre-diabetes precedes diabetes. Blood sugars are higher than normal but not as high as in diabetes. More than 8 out of 10 adults who have pre-diabetes do not know they have pre-diabetes. This is too bad because the progression from pre-diabetes to diabetes can be stopped with life-style changes. People with overweight and obesity are at higher risk for pre-diabetes and diabetes.

The following blood tests, easily done with a finger stick and readily available to all, are done to check if a person has normal sugar levels or levels that are diagnostic of pre-diabetes or diabetes: HbA1c: Hemoglobin (Hb) is the protein in the red blood cells that carries oxygen. Besides oxygen, sugar can also "stick" to the hemoglobin. Hemoglobin A1c measures the percentage of hemoglobin that has sugar "stuck" to it. The percentage is higher when the sugar has been higher. Based on the life span of red blood cells that carry Hemoglobin, HbA1c reflects the "average" of sugar levels over a 3 month-period. HbA1c should be less than 5.7%. Fasting sugar: Done after the person has not had anything to eat or drink EXCEPT water for at least 8 hours prior to the check (usually first thing in the morning). It should be $<$ 100 mg/dl.

ASIM Output: *Let's think of your body as a bustling city, and the sugar (or glucose) in your blood as the citizens moving around. Insulin is like the city's transportation system, helping sugar get to where it needs to go. In conditions like Type 2 Diabetes and Pre-diabetes, it's as if the city's transportation system is broken, and too many citizens (sugar) are stuck wandering around, not reaching their homes (cells).*

Type 2 Diabetes is a long-term health condition where the body has too much sugar in the blood because it doesn't use insulin properly. It's like the key isn't working properly. The fuel (sugar) can't get into the engine (cells) to be used for energy, so it stays in the blood. This means there's too much sugar floating around in your blood, which isn't good for your body over time.

Pre-diabetes is like a warning light on your car's dashboard. It means the sugar levels in your blood are higher than they should be, but not high enough to say you have Type 2 Diabetes. It's a heads-up that if you don't make some changes (like eating healthier and getting more exercise), you might end up with Type 2 Diabetes...

ConBART (Cripwell et al., 2023) Output: *Diabetes is a disease that affects more than 30 million Americans. More than 1 in 4 people with diabetes do not know that they have the disease. People with diabetes have high levels of sugar in their blood. The hormone that helps keep blood sugars in check is at low levels. It is known as Insulin Resistance. This is called Insulin Resistance. A common sign of insulin resistance is a darkening of the skin over the neck and other skin folds. Over time; high blood sugars can cause serious health problems; such as eye, kidney and heart disease...*

UL+Decoder (Flores et al., 2023) Output: *Diabetes happens because the pancreas does not work well with the help of insulin. Over time, high blood sugars can cause serious health issues, such as eye, kidney and heart disease.*

B.2 Example 2: Clinical Document Simplification

Original Medical Text: *Memory Loss and Nicotine*

What is Nicotine? Nicotine is most commonly known as an ingredient in cigarettes, but it may also have medicinal benefits. When not associated with smoking, researchers believe nicotine can stimulate brain systems for memory and attention and be used to treat conditions like memory loss, mental illness, and Parkinson's disease.

Doesn't Nicotine Cause Cancer? Nicotine does not cause cancer. Fifty years of public health messaging has been instrumental in helping people understand that smoking is detrimental to your health and may cause cancer, however nicotine is not to blame. It's the tar and thousands of other chemicals found in cigarettes that make smokers susceptible to cancer, heart disease, and respiratory issues.

Isn't Nicotine Addictive? Researchers have used nicotine in memory studies for more than 30 years, and there appears to be no risk of dependency or addiction when using the nicotine patch to treat mild memory loss or mild cognitive impairment (MCI). In an earlier study, researchers used the nicotine patch to treat MCI in non-smokers, and there were no signs of withdrawal symptoms or cravings after stopping use of the patch. In this same study, participants using nicotine patches showed more improvement in attention and memory than those using the placebo patches containing no nicotine.

Does the Nicotine Patch Have Side Effects? All medications can cause side effects, but most people have no, or minimal, side effects when using the nicotine patch in memory studies. If they occur at all, the most common side effects are nausea, dizziness, and headaches, which typically go away very quickly.

Why Use Nicotine to Treat Memory Loss? Nicotine stimulates systems in the brain related to attention and memory. Researchers believe nicotine can imitate these naturally occurring chemicals in the brain and be used to treat memory loss, among other medical conditions. In fact, researchers have used nicotine in memory studies for more than 30 years. The MIND Study is by far the largest and longest-running study of its kind testing whether the safe use of nicotine, delivered via a patch, can improve memory and functioning in people who have been diagnosed with mild memory loss, or MCI.

ASIM Output: *Nicotine, a key ingredient in cigarettes, is like a helper for our brain's memory and attention "muscles." It's not the nicotine in cigarettes that causes cancer, but the tar and thousands of other harmful chemicals. Nico-*

tine itself, when used in a controlled way, can potentially aid in treating memory problems, certain mental health issues, and Parkinson's disease, a condition that affects movement.

While nicotine can be addictive, especially in cigarettes, it doesn't seem to be addictive when used in a controlled way, like through a patch on your skin for memory problems. People who have used nicotine patches for their memory didn't crave it or have withdrawal symptoms when they stopped using it...

ConBART (Cripwell et al., 2023) Output: *What is Nicotine? Nicotine is most commonly known as an ingredient in cigarettes. But it may also be used as medicine. Nicotine acts like a substance in cigarettes. Researchers believe it works to treat brain loss, mental illness, and even cancer. Doesn't Nicotine cause Cancer? Nicotine does not cause cancer. But the biggest study of its kind is the MIND Study. It shows that smoking is harmful to your health. It may cause cancer, but nicotine is not to blame...*

UL+Decoder (Flores et al., 2023) Output: *Nicotine is most commonly known as an ingredient in cigarettes, but it may also have other medical benefits. When not associated with smoking, researchers believe nicotine can stimulate brain systems for memory and attention and be used to treat conditions like mental illness, and Parkinson's disease. Isn't Nicotine Addictive? Researchers have used nicotine in memory studies for more than 30 years.*