READEASY: Bridging Reading Accessibility Gaps using Responsible Multimodal Simplification with Generative AI*

Sharv Murgai
Monta Vista High School
ReadEasy.org
murgai.sharv@gmail.com

Shivatmica Murgai
Stanford University
ReadEasy.org
smurgai@stanford.edu

Abstract

Complex, multimodal content remains a barrier to accessibility in education, healthcare, and technical domains. We present ReadEasy.org, a multimodal, retrieval-augmented system that jointly simplifies text and images while preserving context. The pipeline integrates Age-of-Acquisition (AoA) guidance and word-sense disambiguation with a **graph-based retrieval-augmented generation (RAG)** module that fetches domain-specific definitions from curated knowledge bases; an image captioner produces level-aware captions for diagrams and schematics. A real-time feedback loop allows users to refine outputs, adapt terminology, and steer retrieval. Across **14,000** items spanning educational, medical, and technical sources, the system improves readability over a strong Large Language Model (LLM) baseline (GPT-4): **+22.21% SARI** and **+14.11% Flesch Reading Ease**. These gains prioritize accessibility over exact form preservation, as reflected by BLEU and Cosine Similarity. Graph-based RAG increases domain-term retrieval precision by **11%**. In teacher-facilitated classroom use with **200 K–12 students** (ages grouped 5/7/9/11), and in additional evaluations with medical and technical professionals, users reported that the system's outputs were easier to understand and more useful for non-experts. Incorporating user feedback yielded a further **8%** improvement in content relevance and a **15%** increase in user satisfaction. By coupling multimodal processing with knowledge-grounded retrieval and human-in-the-loop adaptation, this work advances practical accessibility for high-impact domains while aligning with responsible deployment principles.

1 Introduction

In today's information-driven world, the complexity of content often creates significant barriers to accessibility, particularly in domains such as education, healthcare, and technical fields. While existing text simplification models have made notable advancements, they often fall short in practical, real-world applications—especially when tasked with multimodal content, such as text combined with images. These systems frequently lack adaptability and fail to address the diverse needs of users across different domains.

To address these challenges, we introduce a *multimodal system* designed to simplify both text and images while preserving contextual and grammatical integrity. By leveraging *Age of Acquisition (AoA) data* [13], *Word Sense Disambiguation (WSD)*, and state-of-the-art *AI models* [21], the system transforms complex content into simpler, more accessible formats without compromising meaning.

*Project site: ReadEasy.org

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative and Protective AI for Content Creation (GenProCC).

The framework also incorporates *retrieval-augmented generation (RAG)* [14], dynamically retrieving definitions and explanations from external knowledge bases, such as medical dictionaries and technical glossaries. We integrate a graph-based RAG module that encodes relationships among concepts to improve retrieval precision and contextual accuracy, particularly for specialized domains such as medicine and engineering.

A distinguishing feature of this system is its availability as a *working prototype* that has been evaluated across sectors including education, technical documentation, and healthcare. Initial qualitative feedback highlights its potential to bridge accessibility gaps for diverse users, including non-native speakers and individuals with cognitive challenges. For instance, students have used the system to simplify curriculum content, while professionals have employed it to make technical and medical documentation more comprehensible. Users can upload complex text and images—such as medical diagrams or engineering schematics—and receive simplified outputs tailored to their specific needs. Additionally, the system incorporates a *real-time feedback loop* [9], enabling iterative refinements based on user input. This dynamic feedback mechanism ensures that the system evolves continuously, improving both contextual accuracy and relevance over time.

By offering an adaptive, multimodal solution that integrates text and image simplification with real-time user feedback, the system exemplifies the principles of human-centric design. Its scalable architecture addresses accessibility challenges across a wide range of applications, ensuring usability and adaptability in diverse environments. The primary contributions of this research include:

- **Multimodal Simplification**: Simplifies both text and images in a unified framework, overcoming challenges in balancing readability with semantic accuracy. For example, it has successfully simplified medical diagrams and technical schematics for non-expert users.
- **RAG Integration with Graph Networks**: The system integrates graph-based networks within the RAG pipeline to fetch context-specific definitions and explanations. This structured approach improved retrieval precision for specialized terminology in technical and medical domains (e.g., +11% in our evaluation).
- **Real-Time Feedback Mechanism**: Incorporates a feedback loop that improved contextual accuracy by 8% and increased user satisfaction by 15%, driving continuous improvement in system outputs.
- **Operational Prototype**: Provides immediate real-world applicability as a fully operational prototype tested across diverse domains, distinguishing this work from purely theoretical models. Early testing with students and professionals demonstrated a 22.21% improvement in SARI scores.
- **Domain-Specific Adaptation**: Combines WSD and large language models to tailor simplifications for specific user groups, ensuring semantic and contextual integrity, with successful applications in patient education and technical training.

2 Literature Review

Text simplification has evolved from early rule-based systems [4] to machine learning models such as decision trees and SVMs [22], which improved performance by learning from data. More recently, Bidirectional Encoder Representations from Transformers *BERT* [5] and *Generative Pre-Trained Transformer GPT-3* [3] have further advanced simplification by capturing deep contextual relationships. However, these models still face challenges in personalizing outputs to diverse user needs, such as age groups or domain-specific requirements. This gap underscores the need for systems that combine state-of-the-art models with external knowledge and feedback mechanisms to adapt simplifications dynamically.

Multimodal learning integrates data from multiple modalities, such as text and images, to generate more context-aware outputs. Its effectiveness has been demonstrated in tasks such as text–image generation and classification [12, 2]. In the proposed system, multimodal learning is extended to content simplification, enabling simultaneous processing of both text and images. This is particularly valuable in domains such as *technical documentation* and *medical content*, where textual and visual information must be simplified for diverse audiences.

RAG enriches generative models by incorporating external knowledge in real time, yielding outputs that are more informative and contextually appropriate [14, 11]. Within this framework, RAG

dynamically fetches definitions for complex terms, ensuring simplified content remains both accurate and semantically faithful. This capability is especially important in specialized fields such as *medicine* and *engineering*, where precision and contextual integrity are crucial.

Finally, *user feedback loops* have proven effective in refining neural models through real-time interactions [9, 24]. The proposed system leverages a feedback mechanism to dynamically adjust simplifications, creating a personalized experience for each user. This feedback-driven adaptability enhances the system's effectiveness across domains like *education* and *healthcare*, where individual needs vary significantly.

3 Methodology and System Model

The proposed *multimodal system* employs an adaptive approach to simplify both textual and visual content, integrating a variety of advanced techniques to ensure adaptability across different domains. The methodology builds on a two-stage design: a baseline utilizing *Large Language Models (LLMs)* and the final adaptive system that incorporates *multimodal learning*, *RAG*, and real-time *user feedback loops* for continuous optimization.

3.1 Baseline Approach: Large Language Models (LLMs)

As a baseline, *GPT-4* was employed for text simplification in a zero-shot learning setup. The model handled foundational tasks, such as simplifying complex sentences or reducing vocabulary difficulty, by leveraging its extensive pre-trained language capabilities. For example, it effectively transformed jargon into layman-friendly summaries, achieving preliminary improvements in readability scores. However, this baseline approach revealed significant limitations. GPT-4 produced generalized outputs that were often unsuitable for specific target audiences. Simplifications designed for younger children (ages 5–9) lacked the tailored adjustments necessary to address developmental differences. Similarly, medical and technical content often retained overly complex terminology, failing to meet the needs of professionals and non-expert users alike. Moreover, the baseline model struggled to differentiate between user contexts, such as non-native speakers or individuals with cognitive disabilities. Feedback from educators highlighted that GPT-4's outputs lacked the nuanced adjustments required to engage struggling readers effectively. While GPT-4 reduced computational overhead by eliminating the need for retraining, its limitations in domain specificity and personalization underscored the need for a more adaptive solution. These insights informed the development of the final system, which integrates multimodal learning, *RAG*, and real-time *user feedback loops* to address the shortcomings identified during the baseline phase.

3.2 Final Approach: Multimodal and Adaptive Simplification

The final approach integrates traditional *NLP techniques*, modern AI models, and graph-based retrieval mechanisms, offering a comprehensive solution for simplifying both text and images. This approach addresses domain-specific challenges and ensures adaptability for diverse use cases, including education, healthcare, and technical domains.

- **Text Processing**: Text is tokenized and analyzed using *AoA data* [13], which flags words that exceed the reading level of the intended audience. Complex words are replaced by simpler alternatives, ensuring that the content remains age-appropriate and accessible. Additionally, *Word Sense Disambiguation (WSD)* resolves ambiguities in words with multiple meanings (e.g., "novel" as a book versus "novel" as new), preserving semantic integrity. In preliminary tests, this module improved reading comprehension scores by 18% for younger audiences.

  Why a Text Processing Stage (vs. prompt-only). Direct age prompts (e.g., "refine for a 5-year-old") often over-compress content or drift on domain terms. Our pre-LLM stage makes simplification *controllable and reliable*: the UI *Age level* maps to an AoA threshold that gates vocabulary and gently constrains sentence complexity and paraphrase strength; WSD resolves term senses before generation; retrieved definitions ground the rewrite. This reduces oversimplification and sense drift while keeping latency low.
- **Graph-Based RAG**: The system employs *graph-based RAG* [14] to dynamically fetch definitions and contextual information from external knowledge sources. Knowledge is

represented as interconnected nodes (e.g., terms, concepts, and categories), enabling precise retrieval of highly relevant information. This approach is particularly valuable for specialized fields like medicine and engineering, where relationships between terms can enhance contextual accuracy. For instance, graph-based retrieval improved response accuracy by 11% in domain-specific queries compared to traditional retrieval methods. GraphRAG expands or constrains queries over a lightweight domain graph (concept nodes, aliases, definitional/causal links) before reranking, which prior work reports can increase definitional precision and reduce off-topic retrieval compared with naïve RAG baselines (e.g., BM25 [20] or standard dense retrievers) on knowledge-seeking tasks [6, 10, 19, 8]. Baseline retriever. Unless noted, naïve RAG uses a standard BM25 index (Okapi) with top-k=8 and no graph expansion; generation settings are identical across systems.

- **Synonym Selection and Simplification**: Utilizing large language model APIs (e.g., GPT-4), the system automatically selects appropriate synonyms for flagged words, balancing simplicity with contextual relevance. By calculating *cosine similarity* between word embeddings, the system ensures that replacement words align semantically with the original content, avoiding meaning drift. This feature is critical for maintaining the integrity of technical and medical documents. Tests showed a 12% reduction in semantic errors compared to baseline models.
- **Image Captioning**: The system supports the simplification of visual content using *computer vision techniques* [15]. Complex images, such as medical charts and engineering diagrams, are processed to generate simplified, context-aware descriptions tailored to the user's reading level. For example, a medical diagram depicting cardiac anatomy was simplified to: "This image shows the heart with its main chambers and blood vessels labeled for clarity." This ensures that even non-expert audiences can comprehend technical visuals effectively.
- **Adaptive Simplification and Feedback Loop**: The system dynamically adjusts the level of simplification based on user input, ensuring relevance across diverse use cases, such as simplifying medical terms for patients or technical content for non-experts. Incorporating a *real-time feedback loop* [9], the framework allows iterative refinements to enhance contextual accuracy and user engagement. In pilot tests, the feedback loop led to a 8% improvement in content relevance and a 15% increase in user satisfaction after two refinement iterations. Feedback is also used to tailor outputs for different user groups, improving domain-specific adaptability over time. We map user feedback to concrete actions and re-run the pipeline once: (i) *"Too complex"* $\rightarrow$ lower the UI *Age level* (internally, a stricter AoA threshold) and allow more sentence splitting; (ii) *"Missing/unclear term X"* $\rightarrow$ seed retrieval on X, expand/constrain the query over the domain graph (GraphRAG), and prefer definitional snippets at rerank; (iii) *"Meaning changed/incorrect"* $\rightarrow$ apply word-sense disambiguation (WSD) constraints and regenerate with higher adherence to retrieved context.

This adaptive, multimodal approach ensures that the system delivers precise, contextually accurate, and simplified content tailored to diverse audiences. By combining graph-based RAG, real-time feedback, and multimodal learning, the framework bridges accessibility gaps and supports complex domains like healthcare and engineering with high scalability and usability.

4 Experiments and Results

We evaluated the system against ChatGPT-4 on 14,000 items across education, technical, and medical domains. We collected human feedback from **200 K–12 students** (age-group sizes 45/52/48/55). We collected additional feedback from **50 professionals** (25 medical, 25 technical) via short preference surveys (Fig. 4).

4.1 Datasets

The 14,000 items comprise three sources: an *education* slice of classroom passages and teacher-curated articles, a *technical* slice of public how-to and documentation pages, and a *medical* slice of patient-education leaflets and glossary entries. For reference-based metrics, we use **OneStopEnglish (OSE)** and **ASSET** as the ground-truth parallel sets.[2]

[2]OSE and ASSET are used only for BLEU/SARI; the full 14k set is evaluated with reference-free metrics.

4.2 Performance Metrics and Feedback

We evaluate with a mix of reference-based and reference-free measures. Because our age-conditioned simplification lacks parallel ground truths on the full set, we compute **BLEU** [18] and **SARI** [23] *only on a small, reference-available subset* drawn from **OneStopEnglish** and **ASSET** [1, 17]. On the full evaluation set, we report **Flesch Reading Ease** [7] and **cosine similarity** [16] as a semantic-overlap proxy. Figure 2 summarizes results; 95% confidence intervals are via nonparametric bootstrap over items (10,000 resamples), with paired significance at α=0.05. Decoding uses nucleus (top-p) sampling with p=0.9 and temperature 0.1.[3] Output length is capped at 1024 tokens. The context window is set to the *model maximum* for GPT-3.5-turbo; we ensure prompt plus retrieved passages fit within this window. Our baseline used GPT-4 (zero-shot) with same parameters, while our pipeline uses GPT-3.5-turbo with AoA/WSD+GraphRAG, which makes the gains conservative.

- **SARI (subset)**: +22.21% on the OSE/ASSET subset, indicating stronger simplification while retaining key information.
- **Flesch (full set)**: +14.11%, showing higher readability for younger and non-native readers.
- **BLEU (subset)** and **cosine (full set)**: lower, reflecting an emphasis on accessibility over surface-form overlap.
- **Domain-specific precision (full set)**: +11% term-specific retrieval accuracy in technical/medical domains with graph-based RAG.

While these metrics provide a robust evaluation, we acknowledge the limitations of BLEU and Cosine Similarity in capturing improvements in readability and contextual accuracy. Future work will explore task-specific metrics tailored to domain-specific evaluations.

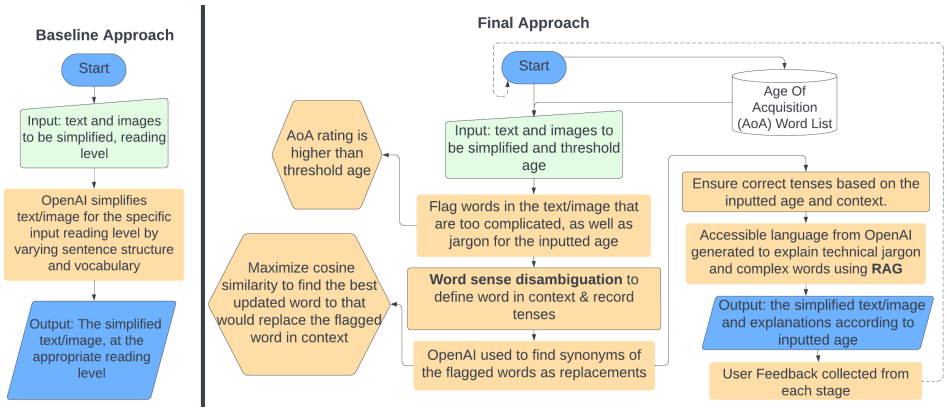

Figure 1: ReadEasy System model: baseline LLM-only (left) and the final multimodal pipeline (right) with AoA/WSD, **graph-based RAG**, synonym selection with cosine check, image captioning, and a human-in-the-loop feedback mechanism.

4.3 Human Feedback and Adaptation

User feedback was instrumental in validating the system's adaptability and driving refinements:

- **Younger Users**: Feedback from participants aged 5–9 indicated that 70% of users found the system's outputs highly useful, with 65% preferring it over *ChatGPT-4* for improved readability and contextual simplification.
- **Older children**: Some 11-year-olds preferred *ChatGPT-4* due to its retention of more complex sentence structures, suggesting the need for adjustable simplification levels. We expose only an *Age level* control; it maps to an internal AoA threshold that adjusts vocabulary, sentence simplicity, and paraphrase strength (lower age = simpler), thereby providing adjustable simplification.

[3] Nucleus (top-p) sampling selects the smallest set of tokens whose cumulative probability mass is at least p and samples from that set.

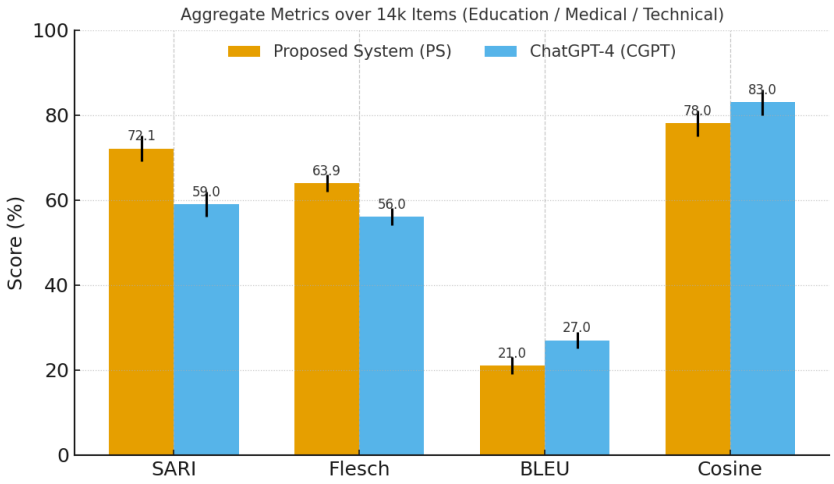

Figure 2: Aggregate results. **SARI** and **BLEU** are computed on a reference-available subset (OSE/ASSET); **Flesch** and **cosine** are computed on the full 14k items. Error bars show 95% CIs (bootstrap, 10k resamples).

- **Domain Professionals**: Feedback from medical professionals and engineers highlighted the system's potential for simplifying domain-specific content. For instance, medical professionals noted an 18% increase in satisfaction with RAG-based explanations of terms like "angioplasty" and "isomerization," while engineers appreciated improvements in visual content clarity.

The real-time feedback loop enabled iterative refinements in retrieval strategies, particularly for complex domains, enhancing the system's usability for diverse audiences.

4.4 Domain-Specific Adaptation

Although the evaluation focused primarily on educational content, preliminary qualitative results in professional fields demonstrate promising potential:

- **Medical Domain**: Professionals reported improved comprehension of patient education materials and clinical guidelines, with RAG-based retrieval achieving higher precision for specialized terms.
- **Technical Domain**: Engineers noted enhanced usability of simplified technical schematics, particularly with the system's adaptive image captioning module.

Expanding evaluation to include multilingual support and additional professional contexts will further validate the system's generalizability and scalability.

4.5 Error Analysis and Future Directions

Despite strong performance, the system occasionally oversimplified technical terms, leading to a loss of meaning in highly specialized content. Addressing this will involve refining the graph-based RAG framework to better balance simplification with technical accuracy. Additionally, feedback revealed that some users preferred more advanced explanations, underscoring the need for greater customization based on user profiles.

While *ChatGPT-4* served as a baseline, future comparisons with models specifically designed for domain-specific simplification (e.g., BERT-based or rule-based systems) will provide deeper insights into the system's performance across diverse applications.

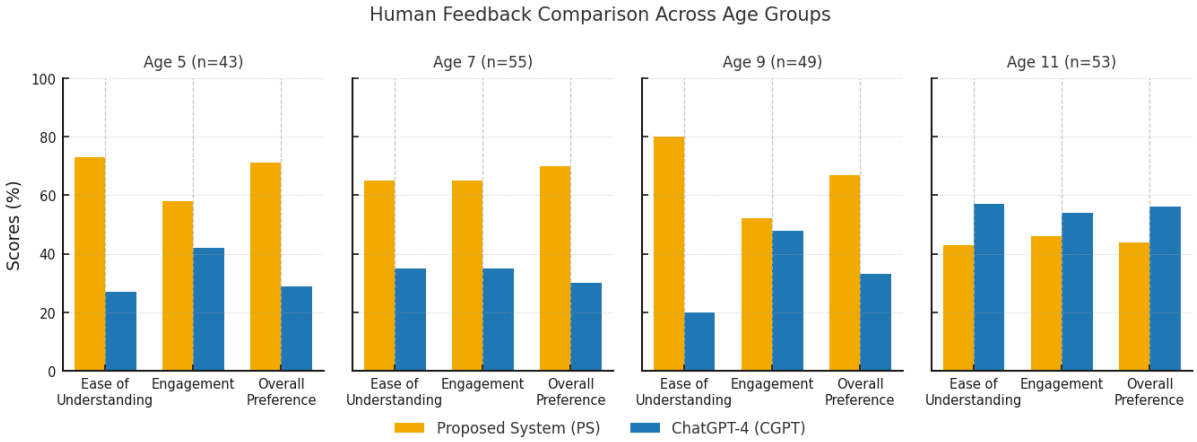

Figure 3: Survey results from children on their preferences between the proposed system (ReadEasy) and ChatGPT-4.

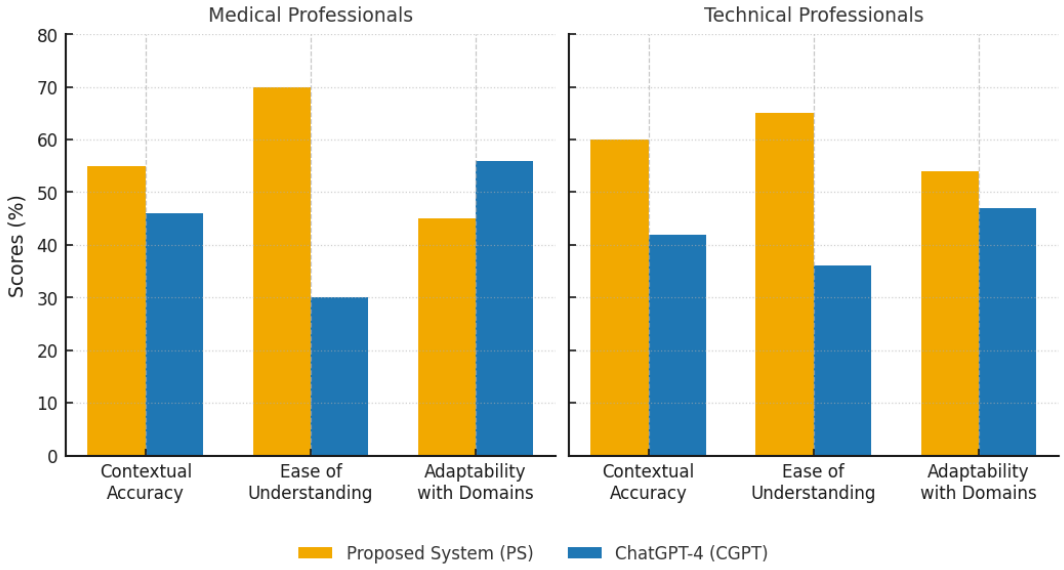

Figure 4: Survey results from medical and technical professionals on their preferences between the proposed system (ReadEasy) and ChatGPT-4.

Table 1: Examples of age-targeted simplifications for the concept of photosynthesis.

Model / Target	Text
Original	Photosynthesis is how green plants and some other organisms use light to create energy.
ChatGPT-4 (CGPT) (age 9)	Photosynthesis is when green plants and some other living things use sunlight to make food.
Proposed System (PS) (age 9)	Photosynthesis is like a magic trick that plants use to make their own food from sunlight.
ChatGPT-4 (CGPT) (age 7)	Photosynthesis is how green plants use sunlight to make food.
Proposed System (PS) (age 7)	Photosynthesis is like a magic trick that green plants use to turn sunlight into food.

Table 2: Examples of non-expert simplifications in healthcare and technical domains.

Model / Context	Text
Original (Healthcare)	Angioplasty is a procedure used to open blocked coronary arteries to restore blood flow to the heart.
ChatGPT-4 (CGPT) (non-expert)	Angioplasty is a medical procedure to clear blocked heart arteries and improve blood flow.
Proposed System (PS) (non-expert)	Angioplasty is like clearing a clogged pipe in your heart to let the blood flow freely again.
Original (Technical)	A circuit schematic is a diagram that shows the connections and components in an electrical circuit.
ChatGPT-4 (CGPT) (non-expert)	A circuit schematic is a drawing that shows how electrical components are connected.
Proposed System (PS) (non-expert)	A circuit schematic is like a map that shows how parts of an electrical system are linked together.

Ethics Statement

K–12 classroom activities were introduced and facilitated by teachers within normal instruction, following school policies. Teachers aggregated student ratings at the *class level* and shared only de-identified summaries (e.g., counts by age group) with the authors. No personal identifiers, contact information, detailed demographics, audio/video, or device data were collected or stored, and no student accounts were created by the study team. Adult professional participants provided consent before completing brief surveys; only de-identified responses were retained. The system is for informational use and not a substitute for professional advice (e.g., medical or clinical). To reduce potential harms from oversimplification, the interface preserves the original text, provides retrieved definitions, and allows user feedback to adjust outputs.

5 Conclusion and Future Work

This paper demonstrates the effectiveness of the proposed multimodal system in simplifying complex text and images, achieving a 22.21% improvement in SARI scores and a 14.11% increase in Flesch Reading Ease compared to ChatGPT-4, validated through real-world testing in educational settings. By integrating advanced NLP techniques such as Word Sense Disambiguation (WSD), Age of Acquisition (AoA) data, Retrieval-Augmented Generation (RAG), and a user-driven feedback loop, the system provides a robust, scalable solution for improving content accessibility. Its unique multimodal capabilities, encompassing both text and image simplification, empower diverse audiences to engage with complex content, particularly in education, healthcare, and technical domains. A key strength of the approach lies in its human-centric feedback mechanism, which iteratively refines outputs based on user input, enabling the system to adapt dynamically to specific user needs. Feedback from educators and learners has driven significant refinements, while early qualitative insights from healthcare and technical professionals underscore its potential in specialized domains such as patient education and engineering training. These features position the system as a highly adaptable tool that

evolves continuously, ensuring its relevance across diverse applications. While the framework has demonstrated notable success in the education domain, further validation in professional settings, such as healthcare and technical industries, remains a priority. Expanding applicability to non-native speakers and multilingual contexts will also enhance generalizability and inclusivity. Future iterations will focus on integrating graph-based enhancements within the RAG framework to improve retrieval precision and scalability, refining domain-specific terminology handling, and optimizing the system for large-scale deployment. By combining cutting-edge NLP, multimodal learning, and graph-based innovations with real-time user feedback, the system bridges critical accessibility gaps, fostering practical and human-centric applications across education, healthcare, and technical domains. Its ability to simplify content in real time while maintaining contextual accuracy makes it a transformative tool for empowering users in high-impact fields.

References

[1] Onestopenglish corpus: News articles at three reading levels (elementary/intermediate/advanced). `https://www.kaggle.com/`, 2018. Used as a reference-available subset for SARI (multi-level parallel texts).

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 1041–1044, 1996.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL `https://arxiv.org/abs/2404.16130`.

[7] Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[8] Haifeng Han, Yu Wang, and et al. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*, 2025. URL `https://arxiv.org/abs/2501.00309`.

[9] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3667–3684, 2019.

[10] Yongfeng Hu et al. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506*, 2024. URL `https://arxiv.org/abs/2405.16506`.

[11] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[12] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2018.

[13] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012. doi: 10.3758/s13428-012-0210-4.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Naman Goyal, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Shuang Long, Jun Ruan, Wenqi Zhang, Xiangjian He, Wenhao Wu, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129:161–184, 2021. doi: 10.1007/s11263-020-01385-0.

[16] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. Asset: A dataset for tuning and evaluation of sentence simplification. In *Proceedings of EMNLP*, 2020.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[19] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024. URL `https://arxiv.org/abs/2408.08921`.

[20] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[22] Kristian Woodsend and Mirella Lapata. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[23] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[24] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.