MedReadCtrl: Readability-Controlled Medical Text Generation to Support Patient Education and Health Equity

Anonymous ACL submission

Abstract

Generative AI has demonstrated strong potential in healthcare, from clinical decision support to patient-facing chatbots that improve outcomes. A critical challenge for deployment is effective human–AI communication, where content must be both personalized and understandable. We introduce MedReadCtrl, a readability-controlled instruction tuning framework that enables LLMs to adjust output complexity without compromising meaning. Evaluations of nine datasets and three tasks in medical and general domains show that MedReadCtrl achieves lower instruction-following errors than GPT-4 (e.g., 1.39 vs. 1.59 on ReadMe, p<0.001) and delivers substantial gains on unseen clinical tasks (e.g., +14.7 ROUGE-L, +6.18 SARI on MTSamples). Experts consistently preferred MedReadCtrl (71.7% vs. 23.3%), especially at low literacy levels. These gains reflect MedReadCtrl's ability to restructure clinical content into accessible, readability-aligned language while preserving medical intent, offering a scalable solution to support patient education and expand equitable access to AI-enabled care.

1 Introduction

The rapid evolution of generative artificial intelligence (AI) is introducing a new era of highly personalized and seamless human-machine communication (Golan et al., 2024). In healthcare, this transformation has the potential to fundamentally redefine clinical care by tailoring interactions and information delivery to meet the unique needs, preferences, and comprehension levels of individual patients (Rajpurkar et al., 2022; Silcox et al., 2024; Yao and Yu, 2025). At its core, a key challenge for AI in healthcare is to maximize its communicative effectiveness with humans—empowering patients, improving engagement, and ultimately advancing patient-centered care (Sauerbrei et al., 2023; Robinson et al., 2024). Achieving this vision, however, requires overcoming a critical and persistent barrier: the vast disparities in domain-specific literacy between users, ranging from patients with limited health literacy to professionals in specialized fields such as law, finance, or medicine. Policies such as Open Medical Records and initiatives such as Blue Button (Delbanco et al., 2010; HealthIT.gov, 2024) exemplify ongoing efforts to democratize access to medical information and promote transparency, engagement, and shared decision-making (Nutbeam, 2023). However, the benefits of these policies are often limited by the patient's ability to comprehend complex medical records and health-related information (Khasawneh et al., 2022). Without ensuring that the information is understandable, accessibility alone cannot translate into meaningful empowerment.

The emergence of large language models (LLMs) introduces unprecedented opportunities to bridge this communication gap (Meng et al., 2024). LLMs excel at complex medical reasoning and content generation, and their growing integration into healthcare systems offers the promise of scalable, personalized patient education and support (Achiam et al., 2023; Kung et al., 2023; Yang et al., 2025). However, for these systems to realize their full potential, it is not enough for AI to generate clinically accurate information; the content must also be tailored to individual patients' literacy levels, backgrounds, and preferences.

In this work, we focus specifically on patient education, a critical component of patient–provider communication that centers on delivering health-related information in a manner patients can understand and act upon. While patient communication encompasses a broader range of interactions, including dialogue, emotional support, and shared decision-making, patient education refers more narrowly to the informational aspect—providing explanations about diagnoses, treatments, and health behaviors. Effective patient education is essen-

tial for improving comprehension, adherence, and health outcomes, especially when tailored to individuals' health literacy levels and backgrounds.

In this context, readability-controlled text generation emerges as a key enabler of personalization, which not only helps providers educate patients more efficiently, but also provides important support for communication and mutual assistance among patients, especially within groups with similar diseases or support groups. This truly enables information to flow among patients, promoting shared decision-making and self-management.

However, existing controllable text generation efforts are limited to binary simplification, complication, or style-transfer tasks (Yao et al., 2024; Guo et al., 2021; Luo et al., 2022; Yao and Yu, 2021; Guo et al., 2023). These approaches fall short of meeting the nuanced and diverse personalization needs of real-world healthcare users, largely due to the limitations of available data and model adaptability. To address these challenges, the field is shifting towards more dynamic, user-adaptive systems capable of fine-grained output control (Liang et al., 2024). This adaptability is especially crucial in healthcare, where demographic and socioeconomic factors significantly influence patient health literacy and comprehension (Vo et al., 2023). Static, one-size-fits-all approaches to readability are insufficient. Instead, there is a growing demand for AI systems that can dynamically adjust communication complexity to fit each patient's unique context, thereby enhancing engagement, decision-making, and self-management (Shi et al., 2024).

In response, we propose Medical Readability-Controlled Instruction Learning (MedReadCtrl), a novel framework designed to endow LLMs with fine-grained readability control capabilities specifically tailored for healthcare contexts. Our approach integrates explicit instruction tuning based on targeted readability levels, enabling LLMs to transform input text into outputs that align with patients' comprehension abilities. We systematically evaluate our system, LlaMA3-MedReadCtrl, across various tasks—text simplification, paraphrase generation, and semantic entailment generation—each crafted to assess the model's ability to adjust output complexity while preserving semantic fidelity. Our extensive automatic and human evaluations across nine datasets in both medical and general domains demonstrate that MedReadCtrl significantly improves LLMs' performance in readability-controlled generation, outperforming existing models and laying a strong foundation for more personalized, accessible healthcare communication.

Our contributions are summarized as follows:

- We propose MedReadCtrl, a novel medical readability-controlled instruction learning framework that enhances LLMs' ability to generate content with controlled readability, making it more suitable for patient education and engagement.
- To the best of our knowledge, LlaMA3-MedReadCtrl is the first system specifically designed to address readability-controlled text generation in the medical domain, enabling fine-grained control over text complexity across different readability levels.
- Extensive automatic and human evaluations validate LlaMA3-MedReadCtrl's effectiveness in both the medical and general NLP domains, demonstrating its ability to improve LLM's performance across three key tasks (text simplification, paraphrase generation, and semantic entailment generation) on nine datasets. This highlights its potential for enhancing patient education by ensuring readability-controlled medical text generation.

2 Methods

Task Overview Addressing health literacy disparities in patient education requires not only factual accuracy but also communication that is accessible to individuals with diverse reading abilities. Direct evaluation with the ultimate target populations, such as patients with limited health literacy or marginalized groups, presents significant logistical, ethical, and cost barriers at scale (Berkman et al., 2011; Nutbeam, 2008). As a pragmatic first step, we focus on a suite of controlled text generation tasks—text simplification, paraphrase generation, and semantic entailment—that are widely recognized as meaningful proxies for assessing a system's capacity to adapt health information for varied literacy levels (Paasche-Orlow and Wolf, 2010; Yao et al., 2024). These tasks collectively reflect key real-world demands in healthcare, such as restating discharge instructions in lay language, generating accessible patient education materials, and supporting self-management across different populations. While not a substitute for direct patient evaluation, this approach offers a scalable and reproducible foundation for benchmarking and tech-

nology development, with clear pathways for future extension to more complex, end-user-centered studies. [1]

Instruction Design for Readability-Controlled Generation Our approach utilizes explicit natural language instructions to guide LLMs in generating outputs at specified readability levels, employing a FLAN-style instruction tuning paradigm. Prompts specify both the generation objective and target reading grade (e.g., 'Generate a paraphrase with grade 5 readability"), supporting transparent and flexible adaptation across communication scenarios. Compared to prior approaches that rely on implicit control or post-hoc selection, this explicit strategy enables better alignment with transparency, regulatory, and user-adaptive requirements. The resulting framework is well suited to rapid customization for a variety of user groups and healthcare settings. [2]

Implementation and Readability Scoring To quantitatively assess output complexity, we employ a comprehensive suite of established readability metrics widely used in both academic research and public health practice: Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Automated Readability Index (ARI), and Coleman-Liau Index (CLI). Their average, the Reading Grade Level (RGL), serves as a practical benchmark for evaluating whether generated content aligns with target audiences. [3]

Datasets To ensure both practical relevance and robust generalization, we employ a diverse set of public datasets spanning general and biomedical domains, with both seen and unseen task splits. Datasets include: ASSET and WikiSmall for general simplification; PAWS and MRPC for paraphrase; SNLI, MultiNLI, and MedNLI for semantic entailment; and ReadME and MTSamples for medical simplification. This selection reflects realistic scenarios in both patient-facing and professional communication, supporting evaluation of models' adaptability across contexts. [4]

[1] Details are in Appendix Task Overview.

[2] Further details and prompt examples are given in AppendixInstruction Design for Readability Control Instruction Learning.

[3] Exact formulas and implementation details appear in Appendix Implementation and Readability Scoring.

[4] Full dataset statistics and splits are described in AppendixDataset.

Evaluation Metrics We apply a multi-metric automatic evaluation framework: ROUGE for content coverage and fluency, BLEU for surface similarity, and SARI for simplification quality. These metrics, widely adopted in NLP and health literacy research, allow us to comprehensively assess content accuracy, coherence, and effective adaptation to intended readability levels—factors essential for practical impact in patient education. [5]

Evaluated Models Our evaluation includes both proprietary (GPT-3.5, GPT-4, Claude-3) and open-source (LLaMA3 8B MedReadCtrl) LLMs. The inclusion of commercial baselines represents the current state-of-the-art in real-world AI deployment, while our instruction-tuned open-source model is explicitly designed for reproducibility and accessibility in low-resource or socially focused applications. All code, data, and model weights are released to facilitate community-driven research, replication, and further extension. [6]

Human and AI-Based Evaluation To complement automatic scoring, we conduct expert human evaluation by clinicians and medical communication specialists, who rate output quality, readability adherence, and suitability for patient-facing use. We also implement LLM-as-a-judge protocols for scalable comparative assessment, using randomization and bias-mitigation procedures. This evaluation framework supports trustworthy and reproducible conclusions about model performance in socially impactful health communication tasks. [7]

3 Results

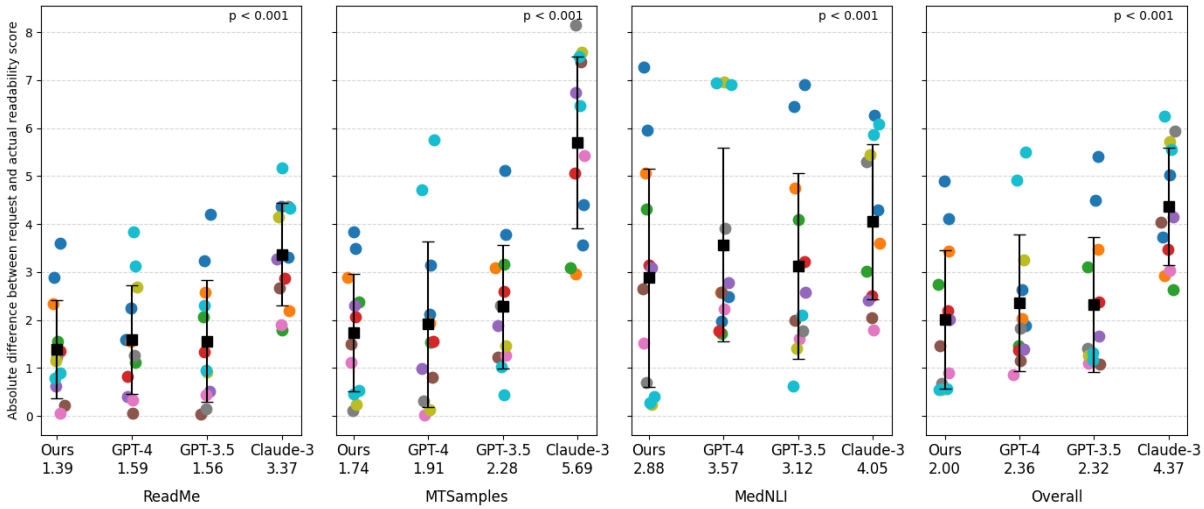
3.1 Automatic evaluation results

Figure 1 presents a comprehensive evaluation of the readability control precision of our LLAMA3-MedReadCtrl model compared with GPT-4, GPT-3.5, Claude-3, and a LLAMA3 baseline across both general and medical domains. In Figure 1a, our LLAMA3-MedReadCtrl model demonstrates significantly improved precision in readability control relative to GPT-4 across multiple medical datasets. Specifically, the model achieves substantially lower average absolute readability errors (computed across four standard readability metrics:

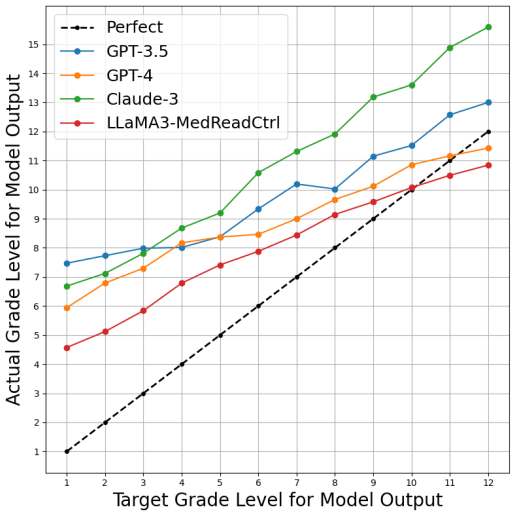
[5] See AppendixEvaluation Metrics for details.

[6] Implementation and model details are provided in AppendixEvaluated Models.

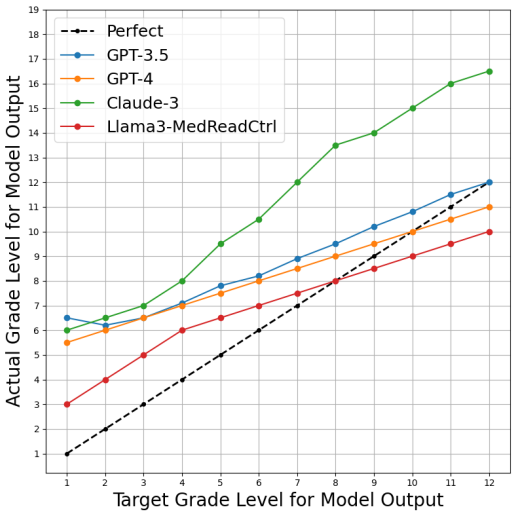
[7] Detailed protocols and guidelines are included in AppendixHuman Evaluation and Appendix AI evaluation.



(a) Comparative analysis of adability control capabilities between different models. The results are measured using the absolute mean difference across four common readability metrics. Each point represents the difference between the request grade and the actual-generated readability (12 points for Grades 1-12). The median and interquartile range for each variable are plotted. The reported P-values correspond to results from nonparametric Mann-Whitney tests. Across three medical-related datasets (ReadMe, MTSamples, MedNLI), Ours consistently achieves lower readability control error (Delta) compared to GPT-4: ReadMe (1.39 vs. 1.59, $p < 0.001$), MTSamples (1.74 vs. 1.91, $p < 0.001$), MedNLI (2.88 vs. 3.57, $p < 0.001$), and Overall (2.00 vs. 2.36, $p < 0.001$). Similar trends are observed when comparing our system to GPT-3.5 and Claude 3 across all datasets, with consistently lower Delta scores (all $p < 0.001$).



(b) MedReadCtrl instruction following on the general domain



(c) MedReadCtrl instruction following on the medical domain

Figure 1: Comparison of readability control precision between our LLAMA3-MedReadCtrl model and GPT-4. (a) LLAMA3-MedReadCtrl demonstrates significantly lower errors in readability control across medical datasets. (b) presents the ability of different LLMs to follow readability-control instructions in general-domain texts. The y-axis represents the model's actual readability output, while the x-axis denotes the target readability level. GPT-4, GPT-3.5, and Claude-3 exhibit an overall upward trend, indicating partial adherence to readability-control instructions, but their curves deviate significantly from the ideal (Perfect) curve, with errors increasing at higher readability levels. In contrast, LLAMA3-8B (without MedReadCtrl training) shows a nearly flat trend, suggesting minimal readability-control capability. LLAMA3-MedReadCtrl closely follows the ideal curve, demonstrating superior readability-control instruction-following ability and precise readability adjustment. (c) evaluates different LLMs on readability-control instruction-following ability in medical texts. GPT and Claude models exhibit patterns similar to their general-domain performance, showing moderate readability-control capability but substantial deviations from the ideal curve, especially at higher target readability levels where errors increase. LLAMA3-MedReadCtrl significantly outperforms other models, closely aligning with the ideal curve, highlighting its strong and stable readability-control ability in the medical domain.

Gunning Fog, Flesch-Kincaid, Automated Readability Index, Coleman-Liau):

- **ReadMe dataset:** LLAMA3-MedReadCtrl (1.39) significantly outperformed GPT-4 (1.59, $p<0.001$).
- **MTSamples dataset:** LLAMA3-MedReadCtrl (1.74) significantly outperformed GPT-4 (1.91, $p<0.001$).
- **MedNLI dataset:** LLAMA3-MedReadCtrl (2.88) significantly outperformed GPT-4 (3.57, $p<0.001$).

These results demonstrate that our system provides superior readability calibration tailored for medical applications. In general-domain readability tasks (Figure 1b), LLAMA3-MedReadCtrl consistently follows the ideal readability level (Perfect curve). In contrast, GPT-4, GPT-3.5, and Claude-3 show increasing deviations as the requested readability level rises, indicating their limited capability to precisely adhere to readability-level instructions. Notably, the baseline LLAMA3 model without MedReadCtrl training fails to demonstrate meaningful readability-level adherence, highlighting the essential role of our readability control training. Similar patterns appear within medical-domain instruction-following evaluations (Figure 1c). LLAMA3-MedReadCtrl closely aligns with targeted readability levels, whereas GPT-4, GPT-3.5, and Claude-3 deviate increasingly at higher readability targets. This disparity underscores the suitability of LLAMA3-MedReadCtrl for precise readability control within specialized medical contexts, significantly enhancing its utility for clinical communication tasks where accurate readability adjustments are crucial.

Table 1 and 6 summarizes the automatic evaluation results of LLAMA3-MedReadCtrl compared to Claude-3, GPT-3.5, and GPT-4 on medical and general domain text generation tasks, using standard NLP metrics (ROUGE, BLEU, and SARI). Across medical domain datasets, LLAMA3-MedReadCtrl consistently achieved the highest performance, clearly outperforming all other evaluated models, both in seen and unseen scenarios—where seen refers to datasets (e.g., MedNLI) used during the readability-controlled instruction tuning phase, and unseen refers to datasets (e.g., MTSamples) that were not included in our training pipeline and are used to assess the model's generalization ability. Note that since base models like LLaMA3 are pretrained on large-scale web corpora, we cannot rule out potential exposure to publicly available datasets during pretraining. Specifically, on the seen MedNLI Semantic Entailment Generation dataset, LLAMA3-MedReadCtrl's performance substantially exceeded GPT-4 in terms of ROUGE-1 (40.27 vs. 24.68), ROUGE-L (39.49 vs. 23.90), BLEU (10.35 vs. 1.78), demonstrating its remarkable capability to generate medically accurate and semantically aligned outputs. On the unseen MTSamples Medical Text Simplification task, LLAMA3-MedReadCtrl notably surpassed GPT-4 by wide margins across all metrics, including ROUGE-1 (53.99 vs. 39.40), ROUGE-L (51.14 vs. 36.44), BLEU (29.73 vs. 11.63), and SARI (23.21 vs. 17.03). This consistently superior performance on unseen medical data demonstrates LLAMA3-MedReadCtrl's robustness and its effectiveness in generalizing readability simplification within unseen medical contexts, which is essential for real-world clinical utility. For general domain tasks, LLAMA3-MedReadCtrl maintained similarly robust performance. In the ASSET Text Simplification dataset, LLAMA3-MedReadCtrl again led the evaluation with clear margins on ROUGE-1 (66.09), ROUGE-L (59.16), BLEU (38.56), and SARI (45.60), highlighting its strong capacity for readability control in general language scenarios. Additionally, LLAMA3-MedReadCtrl demonstrated excellent paraphrasing skills on the PAWS dataset, achieving remarkably superior performance over GPT-4, Claude-3, and GPT-3.5 on all key metrics including ROUGE-1 (93.09 vs. GPT-4's 75.37), ROUGE-L (84.34 vs. 70.64), BLEU (66.49 significantly higher than GPT-4's 31.22), and SARI (60.53 vs. GPT-4's 34.35), emphasizing LLAMA3-MedReadCtrl's outstanding ability to generate precise paraphrased texts that effectively balance lexical novelty with content preservation.

3.2 Human evaluation results

Table 7 presents comprehensive human evaluation results comparing LLAMA3-MedReadCtrl with GPT-4 across three biomedical datasets (MedNLI, MTSamples, ReadMe) and four readability levels (Grades 2, 5, 8, and 11), evaluated along four dimensions: Accuracy, Clarity, Consistency, and Fluency. Overall, LLAMA3-MedReadCtrl achieves superior performance, with an average score of 4.56 compared to GPT-4's 4.17. No-

Models	ROUGE-1	ROUGE-L	BLEU	SARI
ReadMe (seen) – Medical Text Simplification				
Claude-3	34.77 (34.00, 35.54)	30.66 (29.90, 31.42)	07.25 (06.50, 08.00)	51.39 (50.50, 52.28)
GPT-3.5	35.67 (35.10, 36.44)	30.69 (29.90, 31.48)	08.32 (07.55, 09.09)	50.27 (49.40, 51.14)
GPT-4	38.48 (37.70, 39.26)	27.46 (26.70, 28.22)	06.36 (05.60, 07.12)	51.39 (50.50, 52.28)
Llama3-MedReadCtrl	**39.93 (39.16, 40.70)**	**30.79 (30.00, 31.58)**	**10.24 (09.45, 11.03)**	**53.52 (52.65, 54.39)**
MedNLI (seen) – Semantic Entailment Generation				
Claude-3	26.23 (25.46, 26.99)	25.28 (24.52, 26.04)	02.76 (02.00, 03.52)	53.92 (53.08, 54.76)
GPT-3.5	29.91 (29.14, 30.68)	29.08 (28.32, 29.84)	04.59 (03.83, 05.35)	56.57 (55.70, 57.44)
GPT-4	24.68 (23.92, 25.44)	23.90 (23.14, 24.66)	01.78 (01.02, 02.54)	56.12 (55.26, 56.98)
Llama3-MedReadCtrl	**40.27 (39.50, 41.04)**	**39.49 (38.72, 40.26)**	**10.35 (09.59, 11.11)**	**63.97 (63.10, 64.84)**
MTSamples (unseen) – Medical Text Simplification				
Claude-3	37.93 (37.16, 38.70)	34.75 (33.98, 35.52)	10.77 (10.00, 11.54)	18.97 (18.20, 19.74)
GPT-3.5	46.85 (46.08, 47.62)	42.28 (41.51, 43.05)	16.11 (15.34, 16.88)	19.22 (18.45, 19.99)
GPT-4	39.40 (38.63, 40.17)	36.44 (35.67, 37.21)	11.63 (10.86, 12.40)	17.03 (16.26, 17.80)
Llama3-MedReadCtrl	**53.99 (53.22, 54.76)**	**51.14 (50.37, 51.91)**	**29.73 (28.96, 30.50)**	**23.21 (22.44, 23.98)**

Table 1: Main results for MedReadCtrl on medical domain tasks. All values are shown as mean (lower bound, upper bound), representing 95% confidence intervals.

tably, LLAMA3-MedReadCtrl demonstrates particularly strong gains at lower readability levels, where simplifying biomedical content is most challenging. For Grade 2 outputs, our model significantly outperforms GPT-4 in Clarity across all datasets—ReadMe (4.05 vs. 2.60), MTSamples (4.75 vs. 3.10), and MedNLI (4.00 vs. 3.05)—while simultaneously improving Accuracy and Consistency, indicating that readability control does not compromise content fidelity. Similar trends are observed at Grade 5, where LLAMA3-MedReadCtrl consistently achieves higher scores, with improvements of up to +0.8 points in Accuracy and Fluency. Even at higher reading levels (Grades 8 and 11), LLAMA3-MedReadCtrl maintains parity or achieves marginal gains over GPT-4, highlighting its ability to generalize across complexity levels. When broken down by evaluation criteria, LLAMA3-MedReadCtrl consistently surpasses GPT-4: its average Clarity score improves markedly from 3.87 to 4.52, while Accuracy (4.61 vs. 4.33), Consistency (4.57 vs. 4.32), and Fluency (4.56 vs. 4.17) also exhibit consistent gains. Examining individual datasets further confirms this trend; for example, on ReadMe at Grade 11, LLAMA3-MedReadCtrl achieves higher Clarity (4.65 vs. 4.55) and Fluency (4.80 vs. 4.65), and similar improvements are evident on MTSamples and MedNLI. These results collectively underscore LLAMA3-MedReadCtrl's effectiveness in generating readability-controlled biomedical text that not only enhances simplicity but also preserves factual accuracy, logical consistency, and linguistic fluency, outperforming GPT-4 across diverse datasets and readability levels.

Figure 2 and 3 present human preference evaluations comparing LLAMA3-MedReadCtrl and GPT-4 across both general and medical domains, assessed by AI judges and human experts. Overall, LLAMA3-MedReadCtrl demonstrates a clear advantage in both domains. In the general domain, AI evaluations favored LLAMA3-MedReadCtrl in 58.3% of cases, significantly higher than GPT-4's 38.4% preference rate. Human evaluators exhibit a consistent trend, preferring LLAMA3-MedReadCtrl in 52.1% of cases compared to GPT-4's 35.7%. Notably, datasets such as WikiSmall and SNLI show particularly strong human preference for LLAMA3-MedReadCtrl, reaching 65.6% and 66.7% respectively, underscoring its robustness in tasks like text simplification and natural language inference. In the medical domain, LLAMA3-MedReadCtrl's superiority is even more pronounced. AI judges preferred our model in 60.0% of cases versus GPT-4's 37.1%. More importantly, human expert evaluations further amplify this difference: LLAMA3-MedReadCtrl achieves a striking overall preference rate of 71.7%, compared to just 23.3% for GPT-4. Expert agreement is remarkably high, with a Cohen's Kappa of 0.807, indicating strong annotation reliability. Examining individual datasets reveals consistent advantages—LLAMA3-MedReadCtrl achieves over 65% expert preference on ReadMe, MTSamples, and MedNLI, with both MTSamples and MedNLI reaching 75.0%. These results not only reaffirm LLAMA3-MedReadCtrl's effectiveness in medical domains but also highlight its ability to generate text that aligns closely with expert expectations in terms of readability, factual accu-

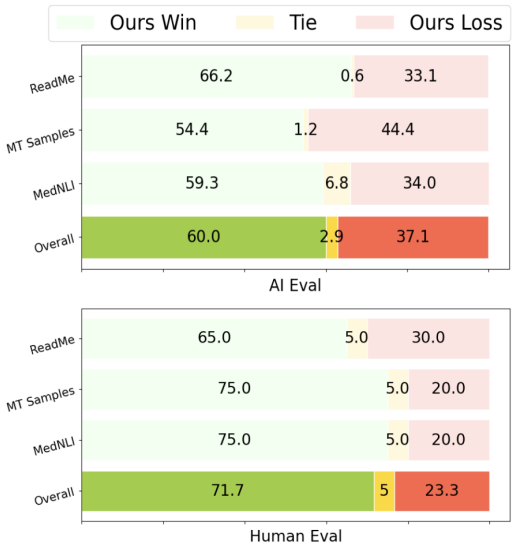


Figure 2: Human preference. Cohen's Kappa for two medical experts: 0.807. Ours vs GPT4 on the medical domain.

racy, and overall quality, making it particularly well-suited for high-stakes applications such as patient education and clinical communication.

4 Discussion

Effective communication in healthcare is widely recognized as a cornerstone of health equity, patient empowerment, and improved health outcomes (Sauerbrei et al., 2023; Robinson et al., 2024). Low health literacy remains a pervasive barrier, disproportionately affecting older adults, minority groups, and individuals with limited education, and has been directly linked to higher rates of hospitalization, increased healthcare costs, and poor medication adherence (Berkman et al., 2011; Paasche-Orlow and Wolf, 2010). Despite years of policy efforts and technological innovation, a substantial gap persists in making health information accessible and actionable for diverse populations (Koh et al., 2012; Berkman et al., 2011). Within the AI community, most language generation systems remain focused on clinical accuracy or general communication, overlooking the need for adaptive readability to address real-world disparities (Nutbeam, 2023; Rajpurkar et al., 2022). Our results demonstrate that LLAMA3-MedReadCtrl represents a significant step toward closing this gap: by enabling large language models to dynamically control text readability while preserving medical quality, our approach directly addresses a long-standing, socially impactful challenge. Notably, compared to state-of-the-art models such as GPT-4 and Claude-3, MedReadCtrl achieves markedly better readability control across both medical and general domains, especially in complex clinical datasets where balancing simplification and precision is critical (Figure 1). These advances point not just to technical progress, but to a new capability for supporting more equitable, patient-centered communication in health systems.

Crucially, human expert evaluations confirm that MedReadCtrl's gains are most pronounced in scenarios with acute health literacy challenges—such as content tailored for early grade reading levels (Grades 2 and 5)—which are precisely the contexts where real-world misunderstandings most frequently lead to adverse outcomes (Vo et al., 2023; Morrison et al., 2019; Suen et al., 2025). Our improvements go beyond stylistic preference: by increasing clarity, fluency, and clinical accuracy for low-literacy audiences, MedReadCtrl enables clearer comprehension of essential instructions, from medication regimens to post-discharge plans—an area where misunderstanding can contribute to preventable hospitalizations and medical errors (Stanceski et al., 2024). These findings closely echo the priorities outlined in health communication and patient empowerment literature, which call for language that is not only technically correct but also culturally and cognitively accessible (Speros, 2005; Nutbeam, 2008). Importantly, our analyses highlight that the most significant gains are observed for groups historically underserved by both healthcare systems and technology, including older adults, non-native speakers, and individuals from lower socioeconomic backgrounds. By grounding our evaluation in both expert judgment and established public health criteria, we demonstrate that MedReadCtrl is well positioned to reduce the risks associated with low health literacy and to advance the field's understanding of effective, equitable AI-driven health communication.

At lower readability levels, accessible language is vital for closing the health comprehension gap among vulnerable populations, such as children, older adults, and non-native speakers—groups widely recognized as most at risk for health misunderstanding and negative outcomes (Vo et al., 2023; Morrison et al., 2019). Our analysis (Table 8) shows that MedReadCtrl consistently produces clear, jargon-free explanations that align with "plain language" guidelines emphasized in public

health policy (Brach, 2024; Nutbeam, 2023). For instance, when describing an X-ray procedure at a Grade 2 level, MedReadCtrl rephrases technical content into simple, intuitive language ("A special picture of the spine that helps doctors see inside"), enabling understanding even for readers with minimal background. This stands in contrast to standard LLMs, which often retain unfamiliar terms or complex structures. Similar improvements are observed for other key scenarios, including symptoms ("yucky stuff from her lungs" in place of "phlegm") and anatomical descriptions ("the area where her legs and hips meet" for "groin"), reflecting strategies used by professional health educators. These results highlight how fine-grained readability control can directly address widely reported health communication breakdowns and suggest that MedReadCtrl could serve as a foundational tool for designing more inclusive, patient-centered digital health interventions—particularly for populations routinely marginalized by conventional medical communication (Berkman et al., 2011; Organization et al., 2014).

At higher readability levels (Grades 8 and 11), MedReadCtrl continues to add value by generating context-sensitive, empathetic, and non-stigmatizing explanations that adapt to both the complexity and sensitivity of medical topics (Table 9). For example, MedReadCtrl reframes "depression treatment drug" as "a medicine that helps people feel better when they are sad," and avoids stigmatizing language in describing sensitive conditions. Notably, the system adapts follow-up care statements to the audience, scaling from conversational phrasing for children ("I will see you again in three months") to detailed, formal plans for adults ("I will schedule a follow-up appointment...to assess the status of your condition..."). These examples demonstrate MedReadCtrl's ability to support tailored health education across diverse user profiles—a requirement repeatedly emphasized in health equity and digital health strategy reports (Koh et al., 2012; Nutbeam, 2023).

Taken together, our results underscore MedReadCtrl's unique value as an enabling technology for equitable, patient-centered health communication. The model's ability to generate semantically faithful, audience-tailored content—across a range of literacy levels—goes beyond surface simplification and directly supports critical applications such as discharge instruction portals, caregiver education, and AI-powered health chatbots, all of which require adaptation to diverse real-world needs (Sauerbrei et al., 2023; Nutbeam, 2023). At the same time, deploying such systems in practice introduces new challenges. Like other large language models, MedReadCtrl remains susceptible to factual hallucinations, at times introducing subtle but clinically meaningful inaccuracies (Table 10). These errors can undermine patient trust and even contribute to adverse outcomes, highlighting the urgent need for robust factuality mechanisms, such as retrieval-augmented generation and pre-generation fact-checking (Lewis et al., 2020; Chung et al., 2025). We also observe that higher-grade outputs may become verbose or overly technical, risking cognitive overload for intended audiences; future work should explore multidimensional control—including tone and brevity—and reinforcement learning from human feedback to better align with user expectations. Additionally, occasional misinterpretation of instructions reveals limitations in zero-shot generalization, calling for improved prompt engineering and more diverse instruction-following training. Yet these challenges do not diminish the transformative potential of fine-grained readability control in LLMs. By closing persistent gaps in health information accessibility and engagement—especially for underserved populations—our work points to a new paradigm for socially responsible, patient-empowering AI in medicine. Importantly, our open-source framework and public release of resources lower barriers for adoption and co-creation, inviting the broader health, technology, and policy communities to participate in collaborative innovation. Realizing the full promise of equitable, high-impact medical AI will require broad, interdisciplinary engagement, and we encourage the AI for Social Impact community to join us in moving from research artifacts to real-world, user-centered solutions.

5 Conclusions

MedReadCtrl establishes a foundation for adaptive, audience-aware medical text generation, directly addressing the health literacy gap, a barrier to equity and empowerment. By combining technical innovation with an open, collaborative ethos, we aim to catalyze further research and responsible deployment of AI for social good. We invite continued partnership across disciplines and communities to realize trustworthy, impactful, inclusive health AI.

6 Limitations

While MedReadCtrl makes substantial strides toward closing the health literacy gap, several limitations and ethical challenges remain. First, our current evaluations are restricted to English-language datasets, which limits immediate generalizability to non-English-speaking populations and cross-cultural contexts. Future work should prioritize multilingual adaptation and culturally sensitive content generation to avoid reinforcing existing inequities. Second, our focus on text simplification, paraphrase, and entailment does not capture the full range of communication formats and real-world clinical scenarios, such as interactive patient-provider dialogues or multi-modal communication. Additional research is needed to extend readability control into these domains and validate effectiveness with diverse user groups, including those most at risk of health miscommunication.

7 Ethical Considerations

From an ethical perspective, deploying LLM-based systems in health contexts raises important questions of responsibility, transparency, and potential for harm. Notably, MedReadCtrl, like other generative models, remains susceptible to factual errors and hallucinations, which can undermine trust, cause anxiety, or even lead to harmful decisions if used without adequate oversight. As AI-generated content is increasingly integrated into clinical and patient-facing applications, it is crucial to establish robust safeguards, including automated fact-checking, human-in-the-loop review, and clear communication about model limitations to users. Additionally, privacy, informed consent, and bias mitigation must be proactively addressed, especially when deploying models in sensitive or marginalized communities. Our future work will focus on incorporating real user feedback from target populations, developing explainable and auditable outputs, and collaborating with ethicists, clinicians, and patient advocates to shape responsible deployment guidelines and regulatory frameworks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. *arXiv preprint arXiv:2005.00481*.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Cindy Brach, editor. 2024. *AHRQ Health Literacy Universal Precautions Toolkit*, 3rd edition. Agency for Healthcare Research and Quality, Rockville, MD. AHRQ Publication No. 15-0023-EF.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Tom Delbanco, Jan Walker, Jonathan D Darer, Joann G Elmore, Henry J Feldman, Suzanne G Leveille, James D Ralston, Stephen E Ross, Elisabeth Vodicka, and Valerie D Weber. 2010. Open notes: doctors and patients signing on.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Roei Golan, Rohit Reddy, and Ranjith Ramasamy. 2024. The rise of artificial intelligence-driven health communication. *Translational Andrology and Urology*, 13(2):356.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2023. Appls: A meta-evaluation testbed for plain language summarization. *arXiv preprint arXiv:2305.14341*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

HealthIT.gov. 2024. About the blue button movement. https://www.healthit.gov/patients-families/about-blue-button-movement. [accessed 2024-10-29].

Amro Khasawneh, Ian Kratzke, Karthik Adapa, Lawrence Marks, and Lukasz Mazur. 2022. Effect of notes' access and complexity on opennotes' utility. *Applied Clinical Informatics*, 13(05):1015–1023.

Howard K Koh, Donald M Berwick, Carolyn M Clancy, Cynthia Baur, Cindy Brach, Linda M Harris, and Eileen G Zerhusen. 2012. New federal policy initiatives to boost health literacy can help the nation move beyond the cycle of costly 'crisis care'. *Health Affairs*, 31(2):434–443.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751.

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. *Advances in neural information processing systems*, 33:9459–9474.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, and 1 others. 2024. The application of large language models in medicine: A scoping review. *Iscience*, 27(5).

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, and 1 others. 2022. Towards more patient friendly clinical notes through language models and ontologies. In *AMIA Annual Symposium Proceedings*, volume 2021, page 881.

Andrea K Morrison, Alexander Glick, and H Shonna Yin. 2019. Health literacy: implications for child health. *Pediatrics in review*, 40(6):263–277.

Don Nutbeam. 2008. The evolving concept of health literacy. *Social science & medicine*, 67(12):2072–2078.

Don Nutbeam. 2023. Artificial intelligence and health literacy—proceed with caution. *Health Literacy and Communication Open*, 1(1):2263355.

World Health Organization and 1 others. 2014. World health report 2013: research for universal health coverage. geneva, switzerland; 2013.

Michael K Paasche-Orlow and Michael S Wolf. 2010. Promoting health literacy research to reduce health disparities. *Journal of health communication*, 15(S2):34–41.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. Ai in health and medicine. *Nature medicine*, 28(1):31–38.

Christopher L Robinson, Ryan S D'Souza, Cyrus Yazdi, Efemena M Diejomaoh, Michael E Schatman, Trent Emerick, and Vwaire Orhurhu. 2024. Reviewing the potential role of artificial intelligence in delivering personalized and interactive pain medicine education for chronic pain patients. *Journal of Pain Research*, pages 923–929.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Aurelia Sauerbrei, Angeliki Kerasidou, Federica Lucivero, and Nina Hallowell. 2023. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Medical Informatics and Decision Making*, 23(1):73.

Ruize Shi, Hong Huang, Wei Zhou, Kehan Yin, Kai Zhao, and Yun Zhao. 2024. From general to specific: Tailoring large language models for personalized healthcare. *arXiv preprint arXiv:2412.15957*.

Christina Silcox, Eyal Zimlichmann, Katie Huber, Neil Rowen, Robert Saunders, Mark McClellan, Charles N Kahn III, Claudia A Salzberg, and David W Bates. 2024. The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. *NPJ Digital Medicine*, 7(1):88.

Carolyn Speros. 2005. Health literacy: concept analysis. *Journal of advanced nursing*, 50(6):633–640.

Kristian Stanceski, Sharleen Zhong, Xumou Zhang, Sam Khadra, Marguerite Tracy, Linda Koria, Sarita Lo, Vasi Naganathan, Jinman Kim, Adam G Dunn, and 1 others. 2024. The quality and safety of using

generative ai to produce patient-centred discharge instructions. *npj Digital Medicine*, 7(1):329.

Kelly Suen, Sunil Shrestha, Samira Osman, and Vibhu Paudyal. 2025. Association between patient race/ethnicity, health literacy, socio-economic status, and incidence of medication errors: A systematic review. *Journal of Racial and Ethnic Health Disparities*, pages 1–12.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Ace Vo, Youyou Tao, Yan Li, and Abdulaziz Albarrak. 2023. The association between social determinants of health and population health outcomes: ecological analysis. *JMIR Public Health and Surveillance*, 9:e44070.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, David McManus, Dan Berlowitz, and Hong Yu. 2025. Unveiling gpt-4v's hidden challenges behind high accuracy on usmle questions: Observational study. *Journal of Medical Internet Research*, 27:e65146.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.

Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1561–1570.

Zonghai Yao and Hong Yu. 2025. A survey on llm-based multi-agent ai hospital. *OSF*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

A Methods

A.1 Task Overview

Our methodology is designed to evaluate the effectiveness of instruction tuning conditional on readability across a suite of tasks, specifically focusing on text simplifications, paraphrase generation, and semantic entailment generation. These tasks are strategically chosen to test the model's capability in adjusting the complexity of its output to match specified readability levels. They serve a broad spectrum of applications, from enhancing educational material accessibility to refining technical documentation for diverse audiences.

- **Text Simplifications:** Here, the aim is to reduce the readability level of the given input text, making it more accessible to a wider audience or readers with varying comprehension skills. This task challenges the model to simplify complex text while preserving its essential content and meaning, demonstrating the ability to decrease textual complexity upon demand.

- **Paraphrase Generation:** In this task, the model is tasked with rewording the given text to produce a paraphrase that maintains the original's readability level. This requires a nuanced understanding of language to ensure the output remains true to the input's complexity and style, facilitating content reformulation without altering its accessibility.

- **Semantic Entailment Generation:** This involves creating text that semantically follows from the given input, with the flexibility to increase or decrease the readability level. The model must grasp the underlying meaning of the input text and generate output that logically entails the input, demonstrating versatility in producing content with adjustable complexity levels.

We employ the instruction tuning approach conditional on readability for all these tasks. This method provides explicit instructions to the model to control the output text's readability score, ensuring that the generated content aligns with the intended complexity level for the target audience. This approach underlines our belief that these tasks can all contribute to readability control generation, where, depending on the task—be it text simplification, paraphrase generation, or semantic entailment generation—the model is calibrated to generate output with the desired readability level. In text simplification, the goal is to lower the readability of the output relative to the input, while in paraphrase generation, the output's readability should mirror the input's. For the semantic entailment generation task, the output's readability may vary, being either higher or lower than the input's, thereby offering a versatile tool for adjusting text complexity across a wide range of contexts.

A.2 Instruction Design for Readability Control Instruction Learning

Instruction Tuning enhances LLMs by training them on NLP tasks formatted as natural language instructions, enabling models to generate structured responses (Wei et al., 2021). This technique has been used to transform English-centric LLMs into open-ended chat models with GPT-like performance by leveraging distilled data from GPT itself (Wang et al., 2022). Our work follows a similar approach, applying instruction tuning to LLM using our created dataset in Section A.3. Unlike traditional supervised fine-tuning (SFT), which relies on manual annotation or synthetic data for readability control, instruction tuning provides a more generalizable approach (Zhang et al., 2023). It enables models to leverage their pre-trained knowledge while learning how to follow structured instructions effectively. Recent advances (Wei et al., 2021; Wang et al., 2022; Tran et al., 2024) highlight the benefits of instruction learning for adaptability across tasks. In our study, we adopt a FLAN-style instruction fine-tuning method (Wei et al., 2021) to train models on task-specific instructions for MedReadCtrl. This approach enhances LLMs' ability to follow readability-controlled instructions, improving usability while reducing the resource-intensive demands of domain-specific fine-tuning.

To achieve the desired readability level across various tasks, we employ straightforward and singular instruction. This approach emphasizes the model's ability to tailor its output to meet specific readability goals, demonstrating its versatility and effectiveness in readability control. The instruction is as follows: "*Given an input text, please output an entailment with a readability score around target readability score.* This concise instruction mandates the model to generate content that not only semantically follows from the given input but also aligns with a specified readability level, showcasing the model's capacity to produce targeted outputs that cater to diverse comprehension needs and preferences.

Implementation and Readability Scoring

The readability of the generated text is quantitatively evaluated using a suite of established readability metrics. We calculate the following readability scores:

- **Flesch-Kincaid Grade Level (FKGL)**: Estimates the U.S. grade level required for comprehension, xQ as:

$$\text{FKGL} = 206.835 - 1.015\left(\frac{\text{totalWords}}{\text{totalSentences}}\right) - 84.6\left(\frac{\text{totalSyllables}}{\text{totalWords}}\right). \quad (1)$$

- **Gunning Fog Index (GFI)**: Measures the years of formal education needed for first-read comprehension, given by:

$$\text{GFI} = 0.4\left(\frac{\text{totalWords}}{\text{totalSentences}} + 100\frac{\text{longWords}}{\text{totalWords}}\right), \quad (2)$$

  where *longWords* are words with more than seven characters.

- **Automated Readability Index (ARI)**: Correlates readability with U.S. school grade level:

$$\text{ARI} = 4.71\left(\frac{\text{totalCharacters}}{\text{totalWords}}\right) + 0.5\left(\frac{\text{totalWords}}{\text{totalSentences}}\right) - 21.43. \quad (3)$$

- **Coleman-Liau Index (CLI)**: Uses character-based analysis for readability estimation:

$$CLI = 0.0588L - 0.296S - 15.8, \quad (4)$$

where L is the average number of letters per 100 words, and S is the average number of sentences per 100 words.

These metrics are selected for their diverse approaches to assessing text complexity, offering a comprehensive understanding of the text's readability. Subsequently, an average Reading Grade Level (RGL) is derived from these scores to represent the text's overall readability. The integration of these readability assessments into our methodology allows a nuanced approach to generating text that meets the specified readability criteria. By adjusting the instruction based on the target RGL, we can fine-tune the complexity of the output, making our approach adaptable to a wide range of applications, from educational content to technical documentation. This process underscores the importance of readability in tailoring content to specific audience needs, a critical factor in communication effectiveness across various domains.

A.3 Dataset

Our experimental framework is designed to assess the model's performance across various tasks, specifically focusing on text simplification, paraphrase generation, and semantic entailment generation. To facilitate a comprehensive evaluation, we utilize six distinct datasets, two for each task, which enables us to explore the model's capabilities in both seen and unseen settings. The datasets employed in our experiments are outlined as follows:

- **Text Simplification:** For this task, we use the ASSET (Alva-Manchego et al., 2020) and WikiSmall (Zhu et al., 2010) datasets. ASSET is a diverse corpus for automatic sentence simplification, providing high-quality simplifications with multiple references per source sentence, making it ideal for instruction tuning and evaluation in seen settings. WikiSmall serves as an additional dataset for evaluating performance in an unseen setting, offering a different collection of simplified sentences derived from Wikipedia articles.

- **Medical Text Simplification:** For this task, we utilize the ReadME and MTSamples datasets. The README dataset (Yao et al., 2024) contains extensive pairs of medical jargon (Kwon et al., 2022) and their corresponding lay definitions, making it ideal for training models to convert complex medical jargon into patient-friendly language in seen settings. MTSamples (Moramarco et al., 2022) provides a rich collection of medical transcription reports across numerous specialties, allowing models to further generalize and evaluate simplification performance in an unseen, clinically varied setting. Together, these datasets support the development of systems that simplify medical language, enhancing patient comprehension and accessibility in healthcare communication.

- **Paraphrase Generation:** We utilize the PAWS (Zhang et al., 2019) (Paraphrase Adversaries from Word Scrambling) and MRPC (Microsoft Research Paraphrase Corpus) (Dolan and Brockett, 2005) datasets. PAWS contains pairs of sentences paraphrasing each other, including those constructed through controlled word scrambling, making it suitable for training and the seen setting evaluations. MRPC offers a collection of sentence pairs labeled as paraphrases or not, sourced from online news sources, to test the model's paraphrasing ability in unseen settings.

- **General Semantic Entailment Generation:** For this task, the SNLI (Stanford Natural Language Inference) (Bowman et al., 2015) and MultiNLI (Multi-Genre Natural Language Inference) (Williams et al., 2017) datasets are employed. SNLI is a large collection of sentence pairs annotated with textual entailment information, used for instruction tuning and seen setting evaluation. MultiNLI extends this to a broader range of genres and contexts, providing a robust challenge for the model in unseen settings. Together, these datasets enable the model to learn and generalize entailment patterns across a wide range of non-specialized texts.

- **Medical Semantic Entailment Generation:** In the biomedical domain, we employ the MedNLI (Medical Natural Language Inference) dataset (Romanov and Shivade, 2018) for entailment tasks specific to clinical set-

tings. Unlike SNLI and MultiNLI, MedNLI consists of sentence pairs derived from electronic health records (EHRs) where medical terminology, context, and clinical reasoning are critical for accurate entailment. This dataset allows the model to focus on domain-specific entailment challenges, making it especially useful for evaluating inference abilities in unseen, medically-focused settings.

In our experimental setup, instruction tuning is performed on the training sets of ASSET, PAWS, and SNLI to align the model's output with specific readability goals. The effectiveness of this approach is then evaluated in two distinct settings: a *seen setting*, using the test sets of ASSET, PAWS, and SNLI, and an *unseen setting*, using the test sets of WikiSmall, MRPC, and MultiNLI. This methodology allows us to not only measure the model's immediate response to the instruction tuning but also its generalizability and adaptability to different textual contexts and tasks.

A.4 Evaluation Metrics

To comprehensively evaluate the model's performance across the different tasks, we employ a multifaceted set of metrics that assess various aspects of the generated texts. These metrics enable us to gauge the model's effectiveness in adjusting readability, maintaining factual accuracy, and ensuring textual coherence and consistency. The following metrics are used:

- **ROUGE:** The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) metric is used to evaluate the quality of text simplification and summarization tasks. Specifically, we employ ROUGE-1 and ROUGE-L, which measure the overlap of unigrams and the longest common subsequence between the generated texts and reference texts, respectively. ROUGE-1 captures basic content similarity, while ROUGE-L assesses the fluency and coherence of the generated text, providing a comprehensive evaluation of the model's ability to retain key information while simplifying or summarizing content.

- **BLEU:** The BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) metric is applied to evaluate paraphrase generation and semantic entailment tasks. It quantifies the linguistic similarity between the generated texts and reference texts, indicating the model's capability to produce coherent and contextually appropriate content.

- **SARI:** The SARI (System output Against References and the Input sentence) (Xu et al., 2016) metric is utilized to assess the quality of text simplification. It measures the model's ability to produce simplified text that is both accurate and helpful, comparing the generated output against both the original text and reference simplifications.

These metrics collectively offer a robust framework for assessing the nuanced performance of the model across various dimensions of text generation, readability adjustment, and content quality.

A.5 Evaluated Models

In our study, we evaluate a diverse set of models to understand their efficacy in handling tasks related to term definition generation, text simplification, and text complication, particularly focusing on adjusting text complexity according to specified readability levels. The models include:

- **GPT-3.5:** As a precursor to GPT-4, GPT-3.5 has demonstrated substantial capabilities in generating human-like text across various tasks. It serves as a baseline to understand the incremental improvements brought about by its successors and other models.

- **GPT-4:** The latest iteration from OpenAI's GPT series at the time of our study, GPT-4, represents a significant leap in language model performance, offering improved comprehension and generation capabilities over its predecessors.

- **Claude-3:** As a model known for its understanding and generation abilities, Claude-3 has been included as a baseline for its efficiency in handling various NLP tasks and its purported adaptability to instruction-based prompts, making it a relevant comparison for our instruction-tuned model.

- **LLaMA3 8B MedReadCtrl:** Our proposed model has been instruction-tuned to adjust the readability level of generated texts based on explicit instructions. LLaMA3 8B is designed to excel in the specific tasks of text simplification, paraphrase generation, and semantic

entailment generation, leveraging instruction tuning to achieve precise control over the readability of its outputs.

Each of these models brings unique strengths and capabilities to the table, allowing us to conduct a comprehensive comparison that not only highlights LLaMA3 8B's advancements in controlling readability but also situates these achievements within the broader context of current NLP technologies. By evaluating LLaMA3 8B against these established models, we aim to demonstrate its efficacy and potential applications in enhancing readability control in automatic text generation.

A.6 Human Evaluation

Our human evaluation was conducted by 2 medical expert evaluators. For 3 medical datasets, We randomly sampled 10 data from the test datasets of 3 data sets, and a total of 30 data appeared in the human evaluation. For 6 general domain datasets, We randomly sampled 6 data from the test datasets of 6 data sets, and a total of 36 data appeared in the human evaluation. For human preference collection, we give detailed instructions to the annotators: "*You are evaluating two systems, both of which are trying to convert inputs to specific readability requirements to produce output suitable for the user. I will show you the input and output of the two systems on grade 2/5/8/11, respectively. Tell me which system's output you prefer by specifying system 1 or system 2 or tie if the quality is the same. Please explain the reason for your preference.*". For human rating evaluation, we asked them to follow annotation guideline in Table 5. Each time, we randomly shuffle the outputs of two systems (LLaMA3-MedReadCtrl and GPT-4), and they can choose the one that better meets the readability requirements and has higher output quality. If they think the outputs of the two systems are tied, they can choose both. After we get judgments from two people per instance, we do not aggregate their labels before calculating the win rate but count them individually.

A.7 AI evaluation

To reduce the heavy human evaluation and make the evaluation easier to reproduce, we use a similar setting of our human preference evaluation for AI evaluation. Comparison-based feedback evaluation assesses the accuracy of LLM in deciding preferences between two responses. However, it is widely acknowledged that current LLMs exhibit significant **positional bias**, i.e., LLMs tend to prefer responses based on their specific position in the prompt. We implement a rigorous verification process to mitigate the effects of positional bias to evaluate the real capability. Specifically, given responses R_a and R_b to be compared, we obtain the comparison based on two orders, noted as $F_a^c = F_c(R_a, R_b)$ and $F_b^c = F_c(R_b, R_a)$. The objective scores are computed by:

$$s = \frac{1}{N}\sum_{i=1}^{N} 1(L(F_{a,i}^c, F_{b,i}^c))$$

where $L(F_a^c, F_b^c)$ is true if and only if $F_a^c \neq F_b^c$ and F_a^c, F_b^c align with ground-truth preference label. N is the number of test samples. All of our experiments were conducted on the version of GPT3.5, GPT4 and Claude 3 between 25 March 2023 and 13 March 2025 by using the OpenAI's API.10 We set temperature = 1, top_p=1, frequency penalty = 0, and presence penalty = 0. The prompts we used for LLM-as-a-judge (claude-3-opus-20240229 and gpt-3.5-turbo-0125) evaluation are in Table 4.

A.8 Hyper-parameter Settings

The experiments were executed using the version 4.37.1 of the transformers library released by Hugging Face. In Table 2, we report the hyperparameters used to train the models on our combined dataset. We use the Adam optimizer and employ a linearly decreasing learning rate schedule with warm-up step is 200. In this section, we detail our experimental setup, the datasets employed, and the evaluation strategy adopted for assessing the performance of our instruction-tuned LLMs in various BioNLP tasks. Furthermore, all experiments were conducted using two Nvidia A100 GPUs, each with 40 GB of memory. The CPU used was an Intel Xeon Gold 6230 processor, and the system was equipped with 192 GB of RAM.

Parameter	Value
Computing Infrastructure	40GB NVIDIA A100 GPU
Optimizer	Adam
Optimizer Params	$\beta = (0.9, 0.999), \epsilon = 10^{-8}$
Learning rate	3×10^{-4}
Learning Rate Decay	Linear
Weight Decay	0
Warmup Steps	200
Batch size	128
Epoch	5

Table 2: Hyperparameter settings for LLaMA3 8B MedReadCtrl.

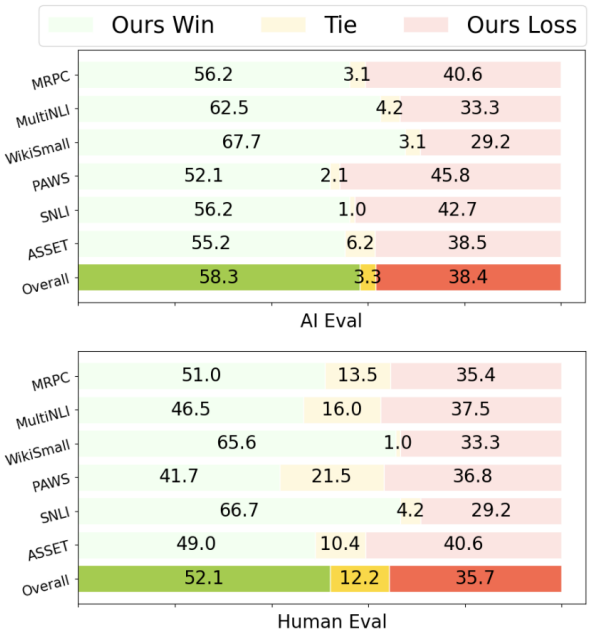


Figure 3: Human preference. Ours vs GPT4 on the general domain.

Dataset (Split)		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12	Total
ReadMe	Train	150(6.10%)	180(7.32%)	170(6.91%)	160(6.50%)	190(7.72%)	200(8.13%)	210(8.54%)	220(8.94%)	230(9.35%)	240(9.76%)	250(10.16%)	260(10.57%)	2460
	Dev	130(6.22%)	140(6.69%)	150(7.17%)	160(7.65%)	170(8.13%)	180(8.61%)	190(9.09%)	200(9.57%)	210(10.05%)	220(10.53%)	230(11.01%)	240(11.49%)	2090
	Test	140(6.22%)	150(6.67%)	160(7.11%)	170(7.56%)	180(8.00%)	190(8.44%)	200(8.89%)	210(9.33%)	220(9.78%)	230(10.22%)	240(10.67%)	250(11.11%)	2250
MedNLI	Train	160(6.36%)	170(6.75%)	180(7.14%)	190(7.54%)	200(7.94%)	210(8.33%)	220(8.73%)	230(9.13%)	240(9.52%)	250(9.92%)	260(10.32%)	270(10.71%)	2520
	Dev	140(6.42%)	150(6.88%)	160(7.34%)	170(7.80%)	180(8.26%)	190(8.72%)	200(9.17%)	210(9.63%)	220(10.09%)	230(10.55%)	240(11.01%)	250(11.47%)	2180
	Test	150(6.28%)	160(6.70%)	170(7.11%)	180(7.53%)	190(7.94%)	200(8.36%)	210(8.77%)	220(9.19%)	230(9.61%)	240(10.03%)	250(10.45%)	260(10.86%)	2390
Asset	Train	170(6.50%)	180(6.88%)	190(7.26%)	200(7.64%)	210(8.02%)	220(8.40%)	230(8.78%)	240(9.16%)	250(9.54%)	260(9.92%)	270(10.30%)	280(10.68%)	2620
	Dev	150(6.12%)	160(6.53%)	170(6.94%)	180(7.35%)	190(7.76%)	200(8.16%)	210(8.57%)	220(8.98%)	230(9.39%)	240(9.80%)	250(10.20%)	260(10.61%)	2450
	Test	160(6.32%)	170(6.71%)	180(7.09%)	190(7.47%)	200(7.85%)	210(8.23%)	220(8.61%)	230(8.99%)	240(9.38%)	250(9.76%)	260(10.14%)	270(10.52%)	2550
SNLI	Train	110(5.00%)	120(5.50%)	130(6.00%)	140(6.50%)	150(7.00%)	160(7.50%)	200(9.50%)	210(10.00%)	220(10.50%)	230(11.00%)	240(11.50%)	250(12.00%)	2100
	Dev	100(5.26%)	110(5.79%)	120(6.32%)	130(6.84%)	140(7.37%)	150(7.89%)	180(9.47%)	190(10.00%)	200(10.53%)	210(11.05%)	220(11.58%)	230(12.11%)	1900
	Test	110(5.08%)	120(5.56%)	130(6.05%)	140(6.53%)	150(7.02%)	160(7.51%)	190(8.91%)	200(9.40%)	210(9.89%)	220(10.38%)	230(10.87%)	240(11.36%)	2100
PAWS	Train	180(6.72%)	190(7.10%)	200(7.46%)	210(7.84%)	220(8.20%)	230(8.56%)	240(8.92%)	250(9.28%)	260(9.64%)	270(10.00%)	280(10.36%)	290(10.72%)	2700
	Dev	160(6.35%)	170(6.75%)	180(7.14%)	190(7.54%)	200(7.94%)	210(8.33%)	220(8.73%)	230(9.13%)	240(9.52%)	250(9.92%)	260(10.32%)	270(10.71%)	2520
	Test	170(6.46%)	180(6.84%)	190(7.22%)	200(7.60%)	210(7.98%)	220(8.36%)	230(8.74%)	240(9.12%)	250(9.50%)	260(9.88%)	270(10.26%)	280(10.64%)	2630
MTSamples	Test	88(10.11%)	78(8.97%)	64(7.36%)	92(10.57%)	57(6.55%)	70(8.05%)	88(10.11%)	68(7.82%)	72(8.28%)	60(6.90%)	60(6.90%)	73(8.39%)	870
WikiSmall	Test	85(9.73%)	89(10.18%)	73(8.35%)	52(5.95%)	71(8.12%)	51(5.84%)	73(8.35%)	93(10.64%)	79(9.04%)	87(9.95%)	51(5.84%)	70(8.01%)	874
MultiNLI	Test	78(8.49%)	59(6.42%)	67(7.29%)	90(9.80%)	72(7.84%)	97(10.57%)	75(8.17%)	89(9.71%)	76(8.29%)	62(6.76%)	61(6.65%)	94(10.25%)	920
MRPC	Test	82(8.65%)	61(6.43%)	71(7.49%)	93(9.81%)	74(7.81%)	98(10.34%)	76(8.02%)	91(9.60%)	77(8.12%)	65(6.86%)	64(6.75%)	96(10.13%)	948

Table 3: Distribution of examples readability scores from instruction tuning datasets

You are evaluating two systems, both of which are trying to convert inputs to specific readability requirements to produce output suitable for the user. I will show you the input and output of the two systems on grade 2/5/8/11, respectively. Tell me which system's output you prefer by specifying system 1 or system 2 or tie if the quality is the same. Please explain the reason for your preference.
Input: [*input*]
System 1 output:
Grade 2: [*system1_2*]
Grade 5: [*system1_5*]
Grade 8: [*system1_8*]
Grade 11: [*system1_11*]
System 2 output:
Grade 2: [*system2_2*]
Grade 5: [*system2_5*]
Grade 8: [*system2_8*]
Grade 11: [*system2_11*]
Please use the following JSON format for your output:
`{'grade 2 preference': xxxx, 'grade 2 preference reasons': xxxx, 'grade 5 preference': xxxx, 'grade 5 preference reasons': xxxx, 'grade 8 preference': xxxx, 'grade 8 preference reasons': xxxx, 'grade 11 preference': xxxx, 'grade 11 preference reasons': xxxx}`
Please only output your response following the required format, and do not output any other content. Now tell me your preference and reasons:

Table 4: Prompt of AI-evaluating readability-controlled text generation outputs.

Table 5: Annotation Guideline for Human Evaluation

Clarity and Simplicity
Definition: Assesses whether the explanation aligns with the reading level in terms of vocabulary, sentence structure, and complexity. **Instructions:** – Grade 2: Use simple, short sentences with basic vocabulary suitable for early readers. – Grade 5: Introduce slightly complex sentence structures but remain easily understandable. – Grade 8: Include moderate complexity in language and concepts appropriate for middle school readers. – Grade 11: Demonstrate advanced vocabulary and technical accuracy suitable for high school students. **Scoring:** – 1: Not suitable for the grade level. – 3: Somewhat matches the grade level but could be improved. – 5: Perfectly matches the grade level.
Accuracy of Content
Definition: Evaluates if the generated text correctly explains the concept without factual errors or misleading simplifications. **Instructions:** Check if the explanation maintains factual consistency while being appropriately detailed for the grade level. **Scoring:** –1: Contains factual errors or misrepresentation. –3: Partially accurate, with minor inaccuracies. –5: Fully accurate and aligned with the input content.
Consistency with Input Meaning
Definition: Ensures that the simplified version retains the core meaning and intent of the input text. **Instructions:** Check for omissions or changes in the intended message or meaning. **Scoring:** –1: Major deviation from the input meaning. –3: Minor deviations or slight loss of meaning. –5: Fully consistent with the original input.
Overall Readability and Fluency
Definition: Assesses the ease of reading the text as a whole, considering grammar, punctuation, and formatting. **Instructions:** –Ensure no major grammatical issues or formatting inconsistencies. –Grade 2 and 5: Prioritize simplicity and readability. –Grade 8 and 11: Balance complexity with natural flow. **Scoring:** –1: Hard to read or poorly formatted. –3: Readable with some flaws. Partially fluent. –5: Easy to read and well-presented.
Annotation Steps 1. Read the original input text and familiarize yourself with the concept. 2. Review the generated outputs for each reading level. 3. Annotate each output across the four dimensions (Clarity, Accuracy, Consistency, and Fluency & Readability). 4. Assign scores for each dimension and write a brief justification (1-2 sentences) for your score. 5. Compare Model 1 and Model 2 based on the aggregated scores across all dimensions and provide justification.

Table 8: Qualitative Analysis for lower readability (e.g., Grade 2 & 5), our model in general outperformed GPT4 in clarity, accuracy, consistency, and fluency.

Grade2	**Input:** X-ray visualization of the spinal cord following injection of contrast medium into the spinal arachnoid space. (README) **GPT4 output:** Taking an X-ray picture of the spinal cord after putting special dye into a space around it. – Rating: Clarity:3, Accuracy:5, Consistency:5, Fluency:4 – Comments: I wouldn't use the term "x-ray" or even "dye" with a 2nd grader) **Ours output:** A special picture of the spine that helps doctors see inside. – Rating: Clarity:5, Accuracy:5, Consistency:5, Fluency:5 – Comments: Great. Appropriate wording.
	Input: She has cough with sputum, occasional blood streaks but no gross blood. (mednli) **GPT4 output:** She has a cough with some phlegm and a little bit of blood sometimes, but no big blood. – Rating: Clarity:3, Accuracy:5, Consistency:5, Fluency:4 – Comments: **Phlegm may not be known to a 2nd grader; I like the effort to use terms like "little bit of" and "big" to quantify, but would have been more fluent to say "lot of" instead of "big"** **Ours output:** She has a cough and sometimes brings up yucky stuff from her lungs. Sometimes it might have a little bit of red in it, but it's not too much. – Rating: Clarity:4, Accuracy:5, Consistency:5, Fluency:5 – Comments: Good.
Grade5	**Input:** She does, however, have some tenderness in the groin bilaterally . (mtsamples) **GPT4 output:** She does have some soreness in her groin on both sides. – Rating: Clarity:3, Accuracy:5, Consistency:5, Fluency:5 – Comments: Not sure a 5th grader knows "groin" either. **Ours output:** She does have some pain in the area where her legs and hips meet, on both sides. – Rating: Clarity:5, Accuracy:5, Consistency:5, Fluency:5 – Comments: I really like this; it doesn't use an anatomical term but describes the location. Perfect!
	Input: Chronic inflammatory and autoimmune disease in which the salivary and lacrimal glands undergo progressive destruction by lymphocytes and plasma cells resulting in decreased production of saliva and tears. The primary form, often called sicca syndrome, involves both KERATOCONJUNCTIVITIS SICCA and XEROSTOMIA. The secondary form includes, in addition, the presence of a connective tissue disease, usually rheumatoid arthritis. (README) **GPT4 output:** This is a long-lasting illness that causes swelling and problems with the body's defense system. It mostly affects the glands that make saliva and tears, which makes it harder to produce them. The main type, known as sicca syndrome, includes two issues: dry eyes and a dry mouth. The second type happens along with another illness, often rheumatoid arthritis. – Rating: Clarity:3, Accuracy:3, Consistency:3, Fluency:3 – Comments: **Worse than 2nd grade version with factual deviation; "swelling" is not really appropriate. Also, mentioning "rheumatoid arthritis" not grade appropriate.** **Ours output:** This is a long-term illness where the body's immune system mistakenly attacks healthy glands, making it hard for people to make saliva and tears. There are two main types of this condition. One type mainly affects the eyes and mouth. The other type is linked to another condition called rheumatoid arthritis. – Rating: Clarity:4, Accuracy:5, Consistency:5, Fluency:5 – Comments: Good; would prefer term other than "rheumatoid arthritis."

Table 9: Qualitative Analysis for higher readability (e.g., Grade 8 & 11). GPT4's performance matches ours. However, our model tends to generate more diversified yet accurate output, while GPT4's output is less diversified, demonstrating the effect of integrated generation leveraging both readability and language modeling.

1	**Input:** The inability in the male to have a PENILE ERECTION due to psychological or organ dysfunction. (README)

GPT4 output for grade 11: The inability of a man to get an erection because of mental issues or problems with his body.
Ours output for grade 11: A condition in men where they are unable to get or maintain an erection, often due to a physical or mental health issue.

Input: depression treatment drug (README)
GPT4 output for grade 8: a medicine for treating depression
Ours output for grade 8: A medicine that helps people feel better when they are sad.

Justification: GPT4 output was essentially just rephrasing the input (replacing the input with synonyms).

---

2 **Input:** I will see in followup in three months . (mtsamples)

GPT4 output (comments):

- Grade 2: I will check back in three months. **– Comments: See in followup is not the same as "check back" (which could imply a phone call, portal message, etc).**
- Grade 5: I will check back in three months. **– Comments: the same**
- Grade 8: I will check back in three months. **– Comments: the same**
- Grade 11: I will check back in three months. **– Comments: the same**

Ours output (comments):

- Grade 2: I will see you again in three months. **– Comments: Good.**
- Grade 5: I will see you again in three months to check on your progress. **– Comments: Would avoid word "progress".**
- Grade 8: I will see you again in three months to check on your progress. **– Comments: Good.**
- Grade 11: I will schedule a follow-up appointment with you in approximately three months to assess the status of your condition and determine any necessary adjustments to your treatment plan. **– Comments: Adds additional text (presumably accurate).**

Justification: GPT4 output was similar across all levels.

Table 10: Error Analysis of our system for hallucinations, excessive length control, and occasional instruction misinterpretation.

---

1 **Input:** She developed a frequent dry cough about 2 weeks ago . (mtsamples)
Output (comments): She got a cough a couple of weeks ago and it makes her throat feel dry.
– Comments: Suggests throat was dry, when in fact the cough was dry.

Input: He was initially reluctant to undergo surgery but has had increasing amounts of chest pain episodes. (mednli)
Output (comments): He didn't want to get surgery at first, but now he gets hurt in his chest a lot.
– Comments: Content not technically accurate; chest pain is not "getting hurt in his chest." But not bad.

---

2 **Input:** Chronic inflammatory and autoimmune disease in which the salivary and lacrimal glands undergo progressive destruction by lymphocytes and plasma cells resulting in decreased production of saliva and tears. The primary form, often called sicca syndrome, involves both KERATOCONJUNCTIVITIS SICCA and XEROSTOMIA. The secondary form includes, in addition, the presence of a connective tissue disease, usually rheumatoid arthritis. (README)

	Output for grade11 (comments): Sjögren's syndrome is a long-standing autoimmune disorder characterized by the gradual destruction of the salivary and lacrimal glands by immune cells, leading to diminished tear and saliva production. This condition can manifest in two primary forms: the first, known as sicca syndrome, involves the simultaneous occurrence of dry eye (keratoconjunctivitis sicca) and dry mouth (xerostomia). The second form is secondary, and in addition to the primary symptoms, it is typically accompanied by the presence of a connective tissue disease, such as rheumatoid arthritis. **– Comments: Very wordy, and vocabulary unnecessarily complicated ("diminished"). Also, awkward phrasing ("the second form is secondary") which would only confuse the reader.**
3	**Input:** Description : A sample note on Rheumatoid Arthritis (mtsamples) **Output:** Rheumatoid Arthritis is a chronic inflammatory disorder characterized by a systemic autoimmune response, resulting in the inflammation and destruction of joints and other tissues, and is often accompanied by symptoms such as joint pain, swelling, and stiffness. **– Comments: Model misinterprets input (maybe perceiving it as instructions to create a sample note?) and generates a definition of rheumatoid arthritis.**

Models	ROUGE-1	ROUGE-L	BLEU	SARI
General Domain Tasks				
ASSET (seen) – Text Simplification				
Claude-3	53.84 (53.30, 54.38)	47.29 (45.96, 48.62)	18.74 (17.64, 19.84)	40.70 (39.80, 41.60)
GPT-3.5	59.76 (58.50, 61.02)	52.42 (51.26, 53.58)	27.39 (26.81, 27.97)	41.01 (40.10, 41.92)
GPT-4	54.87 (53.97, 55.77)	48.26 (47.54, 48.98)	20.61 (19.40, 21.82)	39.73 (38.74, 40.72)
Llama3-MedReadCtrl	**66.09** (**65.38**, **66.80**)	**59.16** (**57.80**, **60.52**)	**38.56** (**37.27**, **39.85**)	**45.60** (**44.44**, **46.76**)
SNLI (seen) – Semantic Entailment Generation				
Claude-3	33.96 (32.97, 34.95)	31.81 (31.04, 32.58)	4.46 (3.13, 5.79)	48.33 (47.51, 49.15)
GPT-3.5	38.08 (36.84, 39.32)	34.81 (33.56, 36.06)	8.75 (8.21, 9.29)	51.02 (49.88, 52.16)
GPT-4	39.32 (37.90, 40.74)	36.18 (35.56, 36.80)	10.50 (9.33, 11.67)	**52.12** (**50.87**, **53.37**)
Llama3-MedReadCtrl	**41.77** (**40.55**, **42.99**)	**37.53** (**36.37**, **38.69**)	**11.22** (**10.05**, **12.39**)	49.93 (48.74, 51.12)
PAWS (seen) – Paraphrase Generation				
Claude-3	57.21 (56.11, 58.31)	50.83 (49.82, 51.84)	23.93 (22.89, 24.97)	38.35 (37.38, 39.32)
GPT-3.5	62.93 (61.51, 64.35)	52.87 (51.41, 54.33)	38.73 (37.87, 39.59)	37.98 (37.41, 38.55)
GPT-4	75.37 (74.44, 76.30)	70.64 (69.38, 71.90)	31.22 (30.37, 32.07)	34.35 (33.04, 35.66)
LLaMA3-MedReadCtrl	**93.09** (**92.25**, **93.93**)	**84.34** (**82.99**, **85.69**)	**66.49** (**65.55**, **67.43**)	**60.53** (**59.05**, **62.01**)
WikiSmall (unseen) – Text Simplification				
Claude-3	48.74 (48.00, 49.48)	40.43 (39.70, 41.16)	16.06 (15.30, 16.82)	33.00 (32.30, 33.70)
GPT-3.5	50.18 (49.42, 50.94)	44.69 (43.94, 45.44)	19.40 (18.64, 20.16)	33.98 (33.24, 34.72)
GPT-4	46.66 (45.92, 47.40)	40.23 (39.50, 40.96)	16.66 (15.90, 17.42)	31.47 (30.72, 32.22)
Llama3-MedReadCtrl	**61.35** (**60.60**, **62.10**)	**55.86** (**55.12**, **56.60**)	**37.28** (**36.54**, **38.02**)	**40.55** (**39.80**, **41.30**)
MultiNLI (unseen) – Semantic Entailment Generation				
Claude-3	30.13 (29.38, 30.88)	27.23 (26.50, 27.96)	03.03 (02.30, 03.76)	44.43 (43.68, 45.18)
GPT-3.5	31.81 (31.06, 32.56)	26.19 (25.44, 26.94)	03.78 (03.03, 04.53)	44.06 (43.32, 44.80)
GPT-4	33.48 (32.73, 34.23)	29.91 (29.16, 30.66)	05.62 (04.87, 06.37)	**46.42** (**45.68**, **47.16**)
Llama3-MedReadCtrl	**40.27** (**39.52**, **41.02**)	**33.49** (**32.75**, **34.23**)	**11.37** (**10.63**, **12.11**)	43.83 (43.08, 44.58)
MRPC (unseen) – Paraphrase Generation				
Claude-3	41.54 (40.80, 42.28)	37.12 (36.38, 37.86)	16.79 (16.05, 17.53)	36.78 (36.04, 37.52)
GPT-3.5	44.93 (44.18, 45.68)	38.23 (37.48, 38.98)	20.59 (19.84, 21.34)	37.34 (36.60, 38.08)
GPT-4	**66.82** (**66.06**, **67.58**)	**61.39** (**60.64**, **62.14**)	15.30 (14.56, 16.04)	34.85 (34.10, 35.60)
LLaMA3-MedReadCtrl	64.68 (63.93, 65.43)	58.47 (57.72, 59.22)	**37.98** (**37.23**, **38.73**)	**44.43** (**43.68**, **45.18**)

Table 6: Main results for MedReadCtrl on general and medical domain tasks. All values are shown as mean (lower bound, upper bound), representing 95% confidence intervals.

		GPT-4				Ours (LLAMA3-MedReadCtrl)			
Dataset	Grade	Accuracy	Clarity	Consistency	Fluency	Accuracy	Clarity	Consistency	Fluency
readme	grade 2	4.10	2.60	4.05	3.95	**4.25**	**4.05**	4.05	**4.40**
readme	grade 5	4.20	3.95	4.20	4.20	**4.80**	**4.25**	**4.80**	**4.45**
readme	grade 8	4.60	4.40	4.60	4.60	**4.90**	**4.75**	**4.90**	**4.65**
readme	grade 11	4.90	4.55	4.90	4.65	4.90	**4.65**	*4.85*	*4.50*
mtsamples	grade 2	3.90	3.10	4.05	3.95	**4.20**	**4.75**	**4.45**	**4.80**
mtsamples	grade 5	3.80	3.80	3.95	4.00	**4.20**	**4.75**	**4.45**	**4.85**
mtsamples	grade 8	4.25	4.25	4.25	4.20	**4.60**	**4.80**	**4.60**	**4.80**
mtsamples	grade 11	4.35	4.25	4.35	4.30	4.35	**4.80**	*4.30*	**4.40**
mednli	grade 2	4.40	3.05	4.30	3.75	**4.55**	**4.00**	**4.40**	**4.30**
mednli	grade 5	4.50	4.40	4.15	4.00	**4.70**	**4.55**	**4.55**	**4.30**
mednli	grade 8	4.40	4.40	4.30	4.30	**4.90**	**4.65**	**4.70**	**4.70**
mednli	grade 11	4.50	4.10	4.50	3.95	**4.85**	**4.70**	**4.75**	**4.55**
Average by Dataset		**GPT-4: mean (95% CI)**				**Ours: mean (95% CI)**			
MedNLI		4.19 (4.06, 4.32)				**4.57 (4.49, 4.66)**			
MTSamples		4.05 (3.94, 4.16)				**4.55 (4.44, 4.66)**			
ReadMe		4.28 (4.16, 4.40)				**4.57 (4.48, 4.67)**			
Average by Grade		**GPT-4**				**Ours**			
Grade 2		3.77 (3.60, 3.93)				**4.33 (4.19, 4.46)**			
Grade 5		4.10 (3.97, 4.22)				**4.55 (4.44, 4.66)**			
Grade 8		4.38 (4.26, 4.50)				**4.75 (4.66, 4.84)**			
Grade 11		4.44 (4.32, 4.56)				**4.63 (4.53, 4.74)**			
Average by Criterion		**GPT-4**				**Ours**			
Accuracy		4.33 (4.20, 4.45)				**4.61 (4.49, 4.73)**			
Clarity		3.87 (3.70, 4.04)				**4.52 (4.41, 4.63)**			
Consistency		4.32 (4.20, 4.44)				**4.57 (4.45, 4.69)**			
Fluency		4.17 (4.03, 4.30)				**4.56 (4.46, 4.66)**			
Overall Average		4.17 (4.10, 4.24)				**4.56 (4.51, 4.62)**			

Table 7: Human evaluation results comparing our **LLAMA3-MedReadCtrl** model and GPT-4 across biomedical datasets (MedNLI, MTSamples, ReadMe) and reading levels (Grades 2, 5, 8, 11). Evaluation dimensions include Accuracy, Clarity, Consistency, and Fluency. **Bold** indicates "Ours better", *Italic* indicates "GPT-4 better", and Underline indicates "Same".