Divergent Readability-Accuracy Strategies of Two Mistral and QWen Models in Biomedical Text Simplification

P. Bilha Githinji[1 [0009−0000−2080−4979]], Aikaterini Melliou[1 [0009−0001−3472−4678]], Zeming Liang[1], Lian Zhang[3], and Peiwu Qin[1 [0000−0001−7336−7848]] ✉

[1] Guangdong Provincial Laboratory of Traditional Chinese Medicine, Hengqin, Guangdong, China qinpeiwu@hqcmlab.cn
[2] Medical Artificial Intelligence Lab, The First Hospital of Hebei Medical University, Hebei Medical University, Shijiazhuang, China.

Abstract. The growing public demand for accessible biomedical information calls for scalable text simplification. While large language models (LLMs) offer solutions, they too struggle with balancing improved readability against preservation of meaning. This report empirically compares how two LLMs - instruction-tuned Mistral-Small 3 24B and the reasoning-augmented Qwen 2.5 32B - navigate this trade-off in biomedical text simplification, benchmarked against human performance. Our analysis highlights how each model applies distinct operational strategies when simplifying biomedical text. Mistral exhibits a tempered lexical simplification approach that consistently enhances readability across multiple metrics while preserving discourse fidelity (BERTScore: 0.91, statistically comparable to that of humans). In comparison, QWen also attains enhanced readability performance and a reasonable BERTScore of 0.89, but presents a disconnect in balancing between readability and accuracy. Additionally, a comprehensive correlation analysis of a suite of 21 metrics confirms strong functional redundancies in metrics and informs adaptation requirements.

Keywords: Text simplification · Large language models · Readability-accuracy trade-off · Health accessibility.

1 Introduction

Access to understandable health information is fundamental to informed decision-making and public health [29,25]. Yet, patient-facing materials frequently exceed the recommended reading levels [29,22,25], and the digital proliferation of biomedical content poses significant risks, including misinformation, oversimplification, and lack of necessary clinical context [30,11]. Reliable automated text simplification offers a scalable solution to render complex clinical and scientific texts into plain language at scale, bridging the gap between expert knowledge and public understanding.

Using biomedical abstracts, which are inherently compressed summaries with concentrated technical jargon, complex sentence structures, and high informational density, this study isolates and comprehensively investigates the readability-accuracy trade-off of two medium-sized general-purpose LLMs that are a practical sweet-spot for research and practice. We comparatively investigate the instruction-tuned Mistral-Small 3 24B [32] and the reasoning-augmented Qwen 2.5 32B [39] under two temperature configurations each, and benchmark their performance against that of human experts. Specific contributions include

- Empirical assessment with only prompting, updating practical baselines and reiterating lexical control as focal for domain adaptation.
- Identification of superior performance by the instruction-tuned Mistral model, which attains higher readability with discourse fidelity that is comparable to human experts, and is robust to temperature adjustment.
- A rigorous assessment of the associations within and across readability and accuracy metrics, revealing how the readability-accuracy tension presents.

2 Related Works

Large language models (LLMs), trained on vast and varied corpora, possess inherent linguistic capabilities that may extend to text simplification out of the box. Text simplification entails reducing specialized vocabulary and complex sentence structures while strictly maintaining semantic equivalence between source and output text [38]. It is distinct from lay summarization, which introduces an additional content distillation step that requires balancing semantic equivalence with conciseness [14].

2.1 Readability and accuracy tension

Empirical evidence consistently demonstrates a persistent and critical performance trade-off between readability gains and content accuracy, with solutions, including LLM-based methods, often achieving high readability at the cost of factual inaccuracies, semantic drift, and undesirable omissions [21,23,37,2]. Some studies emphasize that domain adaptation for biomedical text is necessary during simplification since LLMs do not reason over biomedical semantics to appropriately translate lexical and syntactic changes [35,20]. Domain adaptation strategies for text simplification, however, report conflicting results, and some fail to outperform their general-purpose counterparts [10,27,5,40,31,3,23,17,36,3]. As for general-purpose LLMs, large architectural models such as GPT-4 and Llama3 70B exhibit superior performance or minimal numerical differences from domain-adapted alternatives, while small models underperform considerably [10,9,6].

As LLMs become integrated into everyday information-seeking practices, it is increasingly important to understand their capacity to consistently navigate the tension between maximizing content readability and ensuring discourse fidelity and safety without fine-tuning or technical adaptations typically inaccessible

to lay users. Moreover, rigorous assessment of automated text simplification requires consideration of both readability and preservation of semantic fidelity. While evidence establishes various readability formulas and cautions against traditional accuracy metrics [38,2,17], a comprehensive view of the associations within and between these two functional groups of metrics needs clarification.

2.2 Metrics employed

Readability and coherence metrics. This functional group of metrics comprises of established readability formulas that quantify lexical and syntactic complexity, differing primarily in their treatment of word difficulty. The metrics include Automated Readability Index (ARI) [28], Dale-Chall score [8], Gunning Fog Index [12], Flesch-Kincaid Grade Level (FKGL) [15], Flesch Reading Ease score [16], and SMOG Index [18]. Additionally, SARI [38], a metric specifically formulated for the evaluation of automated text simplification processes, assesses the overall goodness of a simplification system against human reference text.

Word difficulty is typically determined by polysyllabic counts, with notable exceptions being ARI, which relies on character count, and the Dale-Chall formula, which employs a predefined lexicon of 3000 words familiar to a U.S. fourth-grader. Dale-Chall score, Gunning Fog Index, FKGL, SMOG Index, and ARI report performance as an estimated U.S. school grade level, where a lower score indicates improved simplification.

Accuracy and discourse preservation metrics. This functional group assesses semantic integrity, completeness, and subject matter relevance. Traditional semantic congruence metrics include SacreBLEU [24] and ROUGE-L [19], while more robust meaning preservation is evaluated using BERTScore [41], and semantic similarity of document-level embeddings [9]. Thematic consistency and subject matter relevance employ vocabulary matching to trace handling of specialized jargon, and topic modeling techniques such as Latent Dirichlet Allocation (LDA) [13]. While LDA is not ideal for short documents, abstracts have high information density and reasonable length for meaningful word co-occurrence.

Content safety metrics. We additionally consider safety checks using established toxicity classifiers [34], and recognize that, as general-purpose tools, these classifiers may not capture safety risks such as oversimplification or misplaced emphasis. This inclusion hopes to foster continual monitoring necessary for the responsible integration of LLMs.

3 Method

3.1 Data

The primary benchmark is a public dataset of 750 biomedical abstracts that are paired with human-simplified texts [4]. We refer to the process of text simplification by human experts (the gold standard) as the human model, while the

original scientific abstracts of this dataset serve as a curated control cohort (the control set) spanning 75 biomedical topics. An uncurated custom dataset is derived from a random sample of domain-specific abstracts covering Traditional Chinese Medicine (TCM) and Oncology. The selection of TCM and Oncology is partly motivated by the density of unique terminology in this subdomain and partly by the translational relevance given growing public interest in TCM for cancer symptom management [26,33].

3.2 Text simplification systems

We consider two common models for research, Mistral-Small 3 24B, an instruction-tuned model optimized for task fidelity, and Qwen 2.5 32B, a reasoning-augmented model designed for complex problem solving. To assess robustness, we configure each with two temperature settings, namely: a *strict* configuration with temperature $T = 0.2$, and a relatively higher stochasticity state (tagged with $flexi$ for flexible) with $T = 0.4$. The four resulting LLM simplification processes and the human-expert simplification process constitute the five plain-text adaptation systems in our study.

3.3 Prompt design

We specify the simplification process to the LLMs via a standardized, empirically developed, zero-shot prompt aimed at consistent reception of the task across models. The prompt defines the task as a direct sentence-by-sentence adaptation of the input text, explicitly prohibiting summarisation, and addressing both linguistic complexity (e.g., jargon replacement, splitting complex sentences) and discourse complexity (e.g., adding explanations, abstracting esoteric details). The prompt design follows domain-agnostic established principles for plain language communication [7,4], and aligns with the human-benchmark process.

Moreover, we incorporate a self-reporting mechanism where LLMs tag each output sentence with the simplification transform they have applied and an associated rationale. Overall, the prompt design results in a controlled instrument for the assessment. Due to space constraints, the full prompt is presented in a Github page containing the evaluation tools, and comprehensive analysis outputs and reference tables (https://tinyurl.com/text-simplification-eval).

3.4 Evaluation metrics and analysis

We assemble a suite of commonly employed metrics for readability, discourse fidelity, and content safety. Foundational distributional metrics such as average sentence length and proportion of difficult words, inherent in readability formulas, are also tracked to provide fine-grained visibility into drivers of score variation. The metrics are introduced under related works and their operating properties are detailed in the paper's Github page.

For each metric, performance estimators are calculated as the document-level mean scores, and statistically compared using Welch's t-test. Evaluation

metrics calculated from the gold-standard human-simplified text serve as the benchmarking thresholds. All statistical results are reported at a significance level of $\alpha = 0.05$.

4 Results

4.1 System goodness

Overall system goodness is assessed via SARI [38], which compares LLM simplified text against human simplified text. As shown in Table 1, both models advance over previous baseline of 34 (best-reported baseline for encoder-decoder models T5 and BART [4]). Mistral outperforms QWen, achieving SARI scores of 42.46 (flexible, 95% CI: 41.86 - 43.05) and 42.37 (strict, 95% CI: 41.77 - 42.96), while QWen scores 38.38 (flexi, 95% CI: 38.28 - 38.47) and 37.84 (strict, 95% CI: 37.16 - 38.52), indicating Mistral's superior simplification across both temperature settings. Moreover, Mistral's performance approaches that of GPT-4.1-mini, which has a best of five runs SARI score of 43.83 [23]. GPT-4.1-mini (more capable and proprietary model [1]) represents the state of the art at time of analysis.

Table 1: SARI simplification results

Model name	mean	ci (.95)	n
Mistral - flexi	42.46	41.86 - 43.05	606
Mistral - strict	42.37	41.77 - 42.96	606
QWen - flexi	38.38	38.28 - 38.47	569
QWen - strict	37.84	37.16 - 38.52	443
GPT-4.1-mini [23]	43.83	best of 5 runs	-

4.2 Readability and coherence

Standard readability formulas are employed, and their mean results across the five simplification systems are presented in Figure 1, with human benchmark indicated as a threshold and associated statistical comparisons presented in Figure 3.

The LLMs achieve statistically superior readability on four of six metrics but underperform on Dale-Chall and Gunning Fog. This divergence reflects formulaic differences, where these two metrics particularly employ a predefined lexicon of familiar words. Dale-Chall index identifies the human benchmark as having the best readability with a score of 9.40 compared to 12.30 and 12.39 for Mistral and QWen, respectively. Conversely, Flesch-Kincaid Grade Level (QWen 11.45,

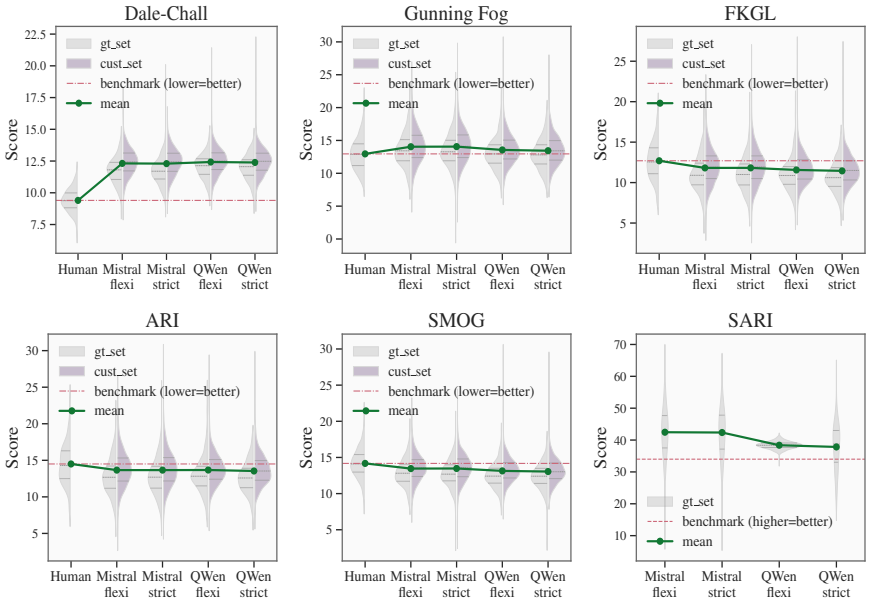

Fig. 1: Readability performance: Average scores and underlying data distributions

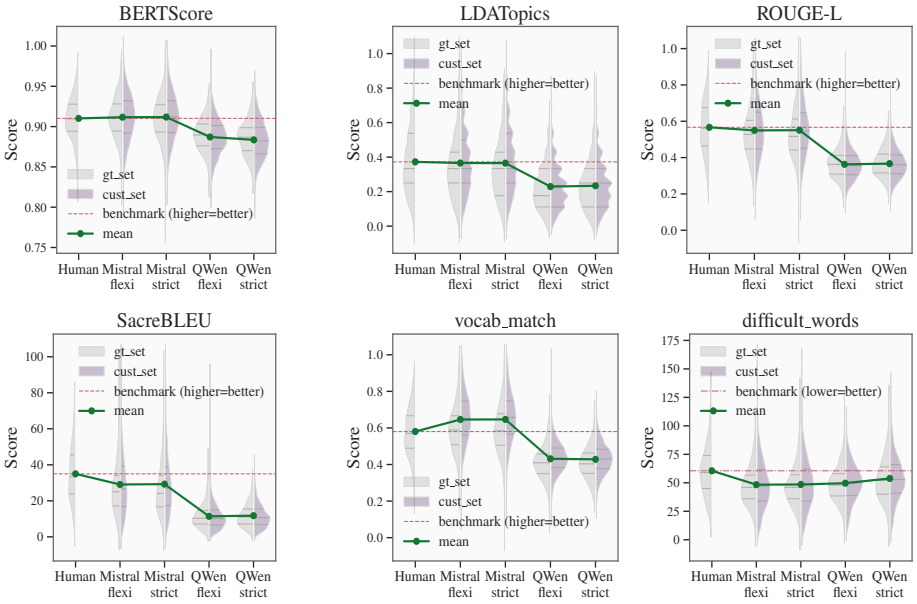

Fig. 2: Accuracy performance: Average scores and underlying data distributions

		n-gram-based			term-based		USA School Grade					scaled					
	Semantic Similarity	BERTScore	ROUGE-L	SacreBLEU	LDATopics	vocab_match	Dale-Chall	Gunning Fog	FKGL	SMOG	ARI	Flesch Ease	difficult_words	Toxicity	avg_words_per_sent	sentences_comp_ratio	words_comp_ratio	SARI
Human benchmark	0.90	0.91	0.57	34.99	0.37	0.58	9.40	12.95	12.70	14.18	14.50	39.77	60.55	0.00	23.39	1.18	1.09	
Mistral means	0.88	0.91	0.55	29.18	0.37	0.65	12.31	14.05	11.82	13.47	13.66	42.09	48.44	0.00	20.48	1.05	0.83	42.41
QWen means	0.84	0.89	0.36	11.55	0.23	0.43	12.41	13.50	11.49	13.10	13.61	43.58	51.48	0.00	20.23	1.07	0.85	38.14
μ_{LLM} Vs μ_{human}																		
All LLMs	0.86	0.90	0.46	21.00	0.30	0.55	12.35	13.80	11.69	13.30	13.64	42.67	49.86	0.00	20.36	1.06	0.84	
Mistral flexi	0.88	0.91	0.55	29.07	0.37	0.65	12.32	14.04	11.81	13.46	13.66	42.08	48.35	0.00	20.50	1.05	0.83	
Mistral strict	0.88	0.91	0.55	29.28	0.37	0.65	12.30	14.06	11.82	13.47	13.66	42.11	48.53	0.00	20.47	1.06	0.83	
QWen flexi	0.84	0.89	0.36	11.39	0.23	0.43	12.42	13.56	11.56	13.14	13.67	43.20	49.66	0.00	20.26	1.07	0.85	
QWen strict	0.84	0.88	0.37	11.76	0.23	0.43	12.39	13.44	11.45	13.05	13.54	43.83	53.85	0.00	20.19	1.08	0.84	
$\mu_1 - \mu_2$																		
QWen vs Mistral	-0.05	-0.03	-0.19	-17.63	-0.13	-0.22	0.10	-0.55	-0.32	-0.37	-0.05	1.49	3.04	-0.00	-0.25	0.02	0.02	-4.27
Mistral flexi vs strict	-0.00	-0.00	-0.00	-0.20	0.00	-0.00	0.02	-0.02	-0.01	-0.01	0.00	-0.03	-0.17	0.00	0.03	-0.01	-0.00	0.09
QWen flexi vs strict	-0.01	0.00	-0.00	-0.37	-0.00	0.00	0.04	0.12	0.11	0.09	0.13	-0.64	-4.20	-0.00	0.07	-0.01	0.00	0.54

Fig. 3: Hypotheses test results. For 'μ_{LLM} Vs μ_{human}' rows, mean values are presented and then shaded with a darker hue if not statistically different from human experts (pvalue > 0.05). For 'μ_1 - μ_2' rows, the difference between means is presented and a darker shading highlights results with pvalue > 0.05 (no significant difference between the models).

Mistral 11.81, human 12.70) and Flesch Reading Ease (QWen 43.83, Mistral 42.11, human 39.77, higher is better) rank QWen best.

While there are statistical differences between model scores, the numerical differences are small. In particular, collectively, the models have readability U.S. school grade scores in the range $12 - 14$ compared to the human benchmark's grade $9 - 15$, with the Dale-Chall index appearing most difficult for the LLMs. Differences due to temperature settings are similarly minimal, though Mistral demonstrates statistically consistent performance across both configurations.

4.3 Accuracy

We assess accuracy through semantic congruence (BERTScore, semantic similarity score), topical relevance (LDA-topics score), traditional content preservation metrics (ROUGE-L, SacreBLUE), and underlying measures (vocabulary matching, difficult words proportion). Full distributions and statistical results appear in Figures 2 and 3.

On BERTScore and LDA-topics score, Mistral achieves human-level performance (0.91 and 0.37, respectively) across temperature settings, with no statistical difference from the human benchmark. QWen attains lower scores that vary with temperature, peaking under flexible configuration (0.89 BERTScore, 0.23 LDA-topics score). Moreover, while both models significantly reduce difficult words compared to human experts (Mistral flexi: 48.35%; QWen flexi: 49.66%; human: 60.55%), Mistral has the highest vocabulary retention and QWen the lowest (Mistral 0.65, QWen 0.43, human 0.58), suggesting conservative treatment of specialized vocabulary by Mistral and exploration by QWen.

Mistral appears to have a conservative strategy that selectively balances lexical simplification with semantic fidelity. Conversely, QWen's approach, while achieving readability gains (presented previously), risks greater semantic displacement. We revisit this trade-off under correlation analysis.

4.4 Content safety

The toxicity scores are virtually zero for all five plain-text adaptation processes and the LLM outputs are not statistically distinguishable from human-simplified text with regards to toxicity. This outcome, while not surprising given the benign nature of the source biomedical abstracts, provides necessary empirical validation of safety constraints and does not obviate the need for continual monitoring.

4.5 Associations between metrics

Here, we consider the tension between readability and accuracy by analysing inter-metric relationships. Pairwise correlation matrices ($\alpha = 0.05$) are visualized in Figure 4, with individual distributional pair-plots provided in the project's Github page

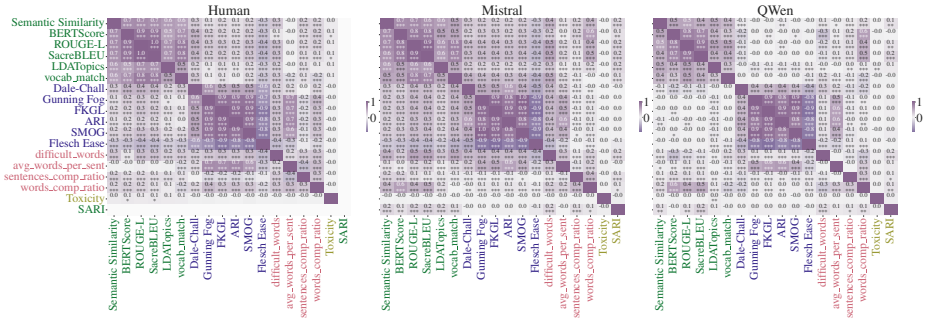

Fig. 4: Correlations between metrics

Readability and coherence. Readability metrics exhibit strong functional congruence and high redundancy within the metric set. Five of six indicators show statistically significant correlations ≥ 0.7 across human and the LLMs (accounting for the directionality of the Flesch Reading Ease score). Dale-Chall correlates more weakly (coefficients in the range $[0.4, 0.6]$), reiterating its relative difficulty for the LLMs.

All systems (human and LLMs) appear to prioritise syntactic over lexical simplification. Correlations for sentence length (average words per sentence) are in the range $[0.4, 0.7]$, while difficult words correlate weakly (coefficients range

$[0.1, 0.4]$). This suggests that lexical control is more challenging or less integrated compared to syntactic restructuring.

Moreover, correlations between readability formulas and difficult words are highest for Mistral compared to Qwen (human $0.2 - 0.3$, Mistral 0.4, Qwen 0.1). The lexical scores for QWen present as less connected to readability formulas despite competitive overall readability performance by QWen.

Additionally, the LLMs appear to achieve readability gains more efficiently, since their compression ratios are below 1 (document length decreases), while humans expand the text during simplification. Mistral and QWen address readability with greater conciseness, a practical advantage for scalable simplification.

Accuracy. Similarly, discourse preservation metrics exhibit internal coherence, with BERTScore appearing to encompass the signal embedded in traditional lexical metrics (ROUGE-L and SacreBLEU), as correlation coefficients are ≥ 0.9. Furthermore, the correlation between BERTScore and words compression ratio is substantially stronger for LLM outputs (~ 0.6) than for human simplification (~ 0.2). While humans use text expansion as a readability-enhancing tactic, the LLMs appear to primarily expand text for the benefit of maintaining semantic integrity.

Cross-function associations. Cross-functional analysis reveals divergent algorithmic strategies between instruction-tuned Mistral and reasoning-augmented QWen. Associations between readability and accuracy, while weakly correlated, have coefficient magnitudes that are larger for Mistral (coefficients range $[0.2, 0.4]$) than for QWen ($[-0.2, 0.1]$), suggesting Mistral yields output that better aligns accuracy and readability optimisation. Similarly, Mistral has larger correlation coefficients for accuracy and word difficulty (coefficients in the range $[0.2, 0.5]$ for Mistral, $[-0.3, 0.0]$ for QWen, and $[0.1, 0.3]$ for human experts). This association is relatively stronger (0.5) for traditional lexical alignment accuracy metrics (ROUGE-L, SacreBLEU), pointing to Mistral's conservative retention-oriented treatment of specialised vocabulary. As for BERTScore, however, correlation with word difficulty is weaker and positive for Mistral (0.2) and human experts (0.1) but weakly negative for QWen (-0.3), reiterating QWen's treatment of lexical complexity that risks semantic integrity.

Regression on principal components of the metrics reinforces these findings. Human simplification yields the highest adjusted R^2 for both readability (Dale-Chall: 0.39; Flesch: 0.45) and accuracy (BERTScore: 0.49), followed by Mistral (Dale-Chall: 0.34; Flesch: 0.37; BERTScore: 0.28), and then QWen (Dale-Chall: 0.26; Flesch: 0.23; BERTScore: 0.29). Humans appear to jointly optimise readability and accuracy objectives, while LLMs show unidirectional responsiveness, where readability improves with accuracy, but accuracy does not correspondingly respond to measured readability aspects. Additionally, Mistral's tempered lexical control is further reflected in higher R^2 for difficult words and vocabulary matching.

It is noteworthy, though, that by itself, QWen is still a capable model for the simplification task; its aggressive lexical substitution yields superior readability

(by most of the formulas) and a reasonable BERTScore (0.89). Its weaker associations may reflect exploration of its broader lexical search space, warranting further investigation. These results may also signal limitations in capturing LLM simplification processes with the current suite of metrics.

4.6 Self-reported rationale on changes made

Analysis of the self-reported rationales offers additional insights into each model's simplification style, quantifying choices made against standardised transforms detailed under prompt design and summarised in Figure 5.

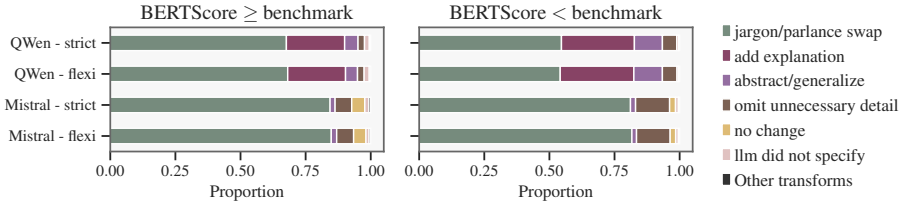

Fig. 5: Self-reported rationale for changes made

Both models rely on jargon/parlance swapping as their dominant simplification strategy but exhibit divergent secondary tactics. Mistral favours omitting superfluous details or retaining original text, treating complex phrases or vocabulary cautiously. Conversely, QWen engages in conceptual expansion, adding explanatory context or abstracting complex concepts. Underperforming samples (observations with below the average benchmark scores) receive more of these secondary tactics in both models. These patterns reflect underlying design philosophies of the models; Mistral is conservative and operates primarily within the bounds of the input, while QWen engages in conceptual exploration. For Mistral, this constrained behaviour seems to favour tempered simplification that preserves discourse fidelity.

5 Discussion

Overall performance. The two general-purpose LLMs in this study demonstrate foundational readiness for biomedical text simplification with zero-shot prompting alone. Both models advance over prior SARI baselines, and Mistral's performance is competitive with GPT-4.1-mini, a proprietary model. In addition, the models have statistically superior readability on four of six readability indices (SMOG Index, FKGL, ARI, and Flesch Reading Ease) and attain BERTScore values of 0.91 and 0.89 for Mistral-Small 3 24B and Qwen 2.5 32B, respectively.

Model differences. The instruction-tuned Mistral exhibits consistent performance across temperature settings, attaining accuracy scores that are statistically indistinguishable from human experts, and balancing readability and accuracy more effectively than QWen. Correlation analysis and self-reported rationales reveal Mistral has a conservative strategy that favors selective retention of specialized language, preserving semantic integrity.

Conversely, the reasoning-augmented QWen adopts a more exploratory approach to lexical complexity, achieving superior readability on four indices but risking semantic degradation (BERTScore 0.89, statistically below human score of 0.91). Its weaker coupling between readability and accuracy suggests disconnected optimization strategies by the model or a metric suite that does not adequately capture QWen's approach. Whether this observation points to a limitation (in model or metric suite) or a potential advantage from exploration of a wider lexical search space warrants further investigation, particularly as LLMs incorporate more domain-specific training data.

Operational insights. The Dale-Chall and Gunning Fog indices, however, indicate statistically inferior results compared to the human benchmark. This discrepancy is traced to formulaic differences concerning the inherent definition (static word list) and weighting of lexical importance. Collectively, results point to apparent syntactic mastery and lexical control as the potential central challenge. As such, domain adaptation for text simplification may benefit from integrating specialized lexical support rather than generalized knowledge expansion.

Strong correlations among SMOG, FKGL, ARI, Flesh Reading Ease, and Gunning Fog (≥ 0.7) confirm redundancies within readability metrics, enabling principled metric reduction. Dale-Chall is distinctively different with coefficients in the range $[0.4 - 0.6]$. In addition, while research recommends text expansion as a tactic for enhancing readability [7,4], the LLMs attain superior readability with more concise output.

6 Conclusion

This study assesses the zero-shot text simplification capabilities of two medium-sized (24B and 32B) general-purpose LLMs that are widely employed in research and practice. Mistral-Small 3 24B exhibits superior operational robustness, attaining readability with a discourse fidelity comparable to human experts. Mistral displays a conservative algorithmic strategy that selectively balances lexical transformation and semantic preservation. In contrast, Qwen 2.5 32B exhibits characteristic conceptual expansion that attains higher readability scores but appears disconnected from semantic resolution or inadequately accounted for by existing metrics. Additionally, the within- and cross-functional evaluation of various readability and accuracy metrics provides a basis for objective comparison and metric selection heuristics, and points to lexical control as the apparent primary hurdle during plain-text adaptation.

Limitations

While we setup a controlled assessment and observe differences between the models in the study, further evaluation across a wider range of LLMs would comprehensively characterize model-size and architectural readiness for plain-text adaptation out of the box. In addition, despite employing established and transferable plain-text adaptation tactics, which should offer directional evidence or an informed starting point for broader application areas, the study explicitly addresses only one domain (biomedicine). Moreover, the inherent limitations of conventional readability formulas must be recognized, and quantitative metrics should be complemented by human validation strategies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. GPT-4.1 Mini - API Pricing & Benchmarks, https://openrouter.ai/openai/gpt-4.1-mini
2. Agrawal, M., Chen, I.Y., Gulamali, F., Joshi, S.: The evaluation illusion of large language models in medicine. npj Digital Medicine **8**(1), 1–4 (Oct 2025). https://doi.org/10.1038/s41746-025-01963-x, https://www.nature.com/articles/s41746-025-01963-x
3. Alamleh, S., Mavedatnia, D., Francis, G., Le, T., Davies, J., Lin, V., Lee, J.J.: Readability, Reliability, and Quality Analysis of Internet-Based Patient Education Materials and Large Language Models on Meniere's Disease. Journal of Otolaryngology - Head & Neck Surgery **54**, 19160216251360651 (Jul 2025). https://doi.org/10.1177/19160216251360651, https://journals.sagepub.com/doi/10.1177/19160216251360651
4. Attal, K., Ondov, B., Demner-Fushman, D.: A dataset for plain language adaptation of biomedical abstracts. Scientific Data **10**(1), 8 (Jan 2023). https://doi.org/10.1038/s41597-022-01920-3, https://www.nature.com/articles/s41597-022-01920-3
5. Balde, G., Roy, S., Mondal, M., Ganguly, N.: MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization. vol. 7, pp. 6180–6188 (Aug 2024). https://doi.org/10.24963/ijcai.2024/683, https://www.ijcai.org/proceedings/2024/683
6. Chen, Q., Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M.B., Ai, X., Lai, P.T., Wang, Z., Keloth, V.K., Raja, K., Huang, J., He, H., Lin, F., Du, J., Zhang, R., Zheng, W.J., Adelman, R.A., Lu, Z., Xu, H.: Benchmarking large language models for biomedical natural language processing applications and recommendations. Nature Communications **16**(1), 3280 (Apr 2025). https://doi.org/10.1038/s41467-025-56989-2, https://www.nature.com/articles/s41467-025-56989-2
7. Cramm, H., Breimer, J., Lee, L., Burch, J., Ashford, V., Schaub, M.: Best practices for writing effective lay summaries. Journal of Military, Veteran and Family Health **3**(1), 7–20 (Apr 2017). https://doi.org/10.3138/jmvfh.3.1.004, https://utppublishing.com/doi/10.3138/jmvfh.3.1.004

8. Dale, E., Chall, J.S.: The Concept of Readability. Elementary English **26**(1), 19–26 (1949), http://www.jstor.org/stable/41383594
9. Dorfner, F.J., Dada, A., Busch, F., Makowski, M.R., Han, T., Truhn, D., Kleesiek, J., Sushil, M., Lammert, J., Adams, L.C., Bressem, K.K.: Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data (Aug 2024). https://doi.org/10.48550/arXiv.2408.13833, http://arxiv.org/abs/2408.13833, arXiv:2408.13833
10. Feng, H., Ronzano, F., LaFleur, J., Garber, M., De Oliveira, R., Rough, K., Roth, K., Nanavati, J., El Abidine, K.Z., Mack, C.: Evaluation of Large Language Model Performance on the Biomedical Language Understanding and Reasoning Benchmark: Comparative Study (May 2024). https://doi.org/10.1101/2024.05.17.24307411, http://medrxiv.org/lookup/doi/10.1101/2024.05.17.24307411
11. Gallardo, M.O., Ebardo, R.: Online Health Information Seeking in Social Media. In: Patel, K.K., Santosh, K., Patel, A., Ghosh, A. (eds.) Soft Computing and Its Engineering Applications, vol. 2030, pp. 168–179. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-53731-8_14, https://link.springer.com/10.1007/978-3-031-53731-8_14
12. Gunning, R.: The technique of clear writing. N.Y., rev ed. edn., oCLC: 1260373335
13. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (Jun 2019). https://doi.org/10.1007/s11042-018-6894-4, http://link.springer.com/10.1007/s11042-018-6894-4
14. Khan, B., Shah, Z.A., Usman, M., Khan, I., Niazi, B.: Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey. IEEE Access **11**, 109819–109840 (2023). https://doi.org/10.1109/ACCESS.2023.3322188, https://ieeexplore.ieee.org/document/10272614/
15. Kincaid, J.P., Braby, R., Mears, J.E.: Electronic authoring and delivery of technical information. Journal of Instructional Development **11**(2), 8–13 (Jun 1988). https://doi.org/10.1007/BF02904998, http://link.springer.com/10.1007/BF02904998
16. Klare, G.R., Rowe, P.P., John, M.G.S., Stolurow, L.M.: Automation of the Flesch Reading Ease Readability Formula, with Various Options. Reading Research Quarterly **4**(4), 550 (1969). https://doi.org/10.2307/747070, https://www.jstor.org/stable/747070?origin=crossref
17. Kocbek, P., Kopitar, L., Stiglic, G.: Plain Language Adaptations of Biomedical Text Using LLMs: Comparision of Evaluation Metrics. In: Househ, M.S., Tariq, Z.U.A., Al-Zubaidi, M., Shah, U., Huesing, E. (eds.) Studies in Health Technology and Informatics. IOS Press (Aug 2025). https://doi.org/10.3233/SHTI250946, https://ebooks.iospress.nl/doi/10.3233/SHTI250946
18. Laughlin, G.H.M.: Smog grading-a new readability formula. Journal of Reading **12**(8), 639–646 (1969), http://www.jstor.org/stable/40011226
19. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
20. Mai, H.T., Chu, C.X., Paulheim, H.: Do LLMs Really Adapt to Domains? An Ontology Learning Perspective. In: Demartini, G., Hose, K., Acosta, M., Palmonari, M., Cheng, G., Skaf-Molli, H., Ferranti, N., Hernández, D., Hogan, A. (eds.) The Semantic Web – ISWC 2024, vol. 15231, pp. 126–143. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-77844-5_7, https://link.springer.com/10.1007/978-3-031-77844-5_7

21. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On Faithfulness and Factuality in Abstractive Summarization. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1906–1919. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.173, https://aclanthology.org/2020.acl-main.173/
22. Mishra, V., Dexter, J.P.: Comparison of Readability of Official Public Health Information About COVID-19 on Websites of International Agencies and the Governments of 15 Countries. JAMA Network Open **3**(8), e2018033 (Aug 2020). https://doi.org/10.1001/jamanetworkopen.2020.18033, https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2769382
23. Ondov, B., Xia, W., Attal, K., Unde, I., He, J., Demner-Fushman, D.: Lessons from the TREC Plain Language Adaptation of Biomedical Abstracts (PLABA) track (2025). https://doi.org/10.48550/ARXIV.2507.14096, https://arxiv.org/abs/2507.14096
24. Post, M.: A Call for Clarity in Reporting BLEU Scores. In: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., Verspoor, K. (eds.) Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6319, https://aclanthology.org/W18-6319/
25. Schlacher, A.: A guide for policy and decision makers on health literacy policies. European Journal of Public Health **34**(Supplement_3), ckae144.787 (Nov 2024). https://doi.org/10.1093/eurpub/ckae144.787, https://academic.oup.com/eurpub/article/doi/10.1093/eurpub/ckae144.787/7844567
26. Schuerger, N., Klein, E., Hapfelmeier, A., Kiechle, M., Brambs, C., Paepke, D.: Evaluating the Demand for Integrative Medicine Practices in Breast and Gynecological Cancer Patients. Breast Care **14**(1), 35–40 (2019). https://doi.org/10.1159/000492235, https://karger.com/article/doi/10.1159/000492235
27. Shao, Y., Yan, M., Liu, Y., Chen, S., Chen, W., Long, X., Yan, Z., Li, L., Zhang, C., Sebe, N., Tang, H., Wang, Y., Zhao, H., Wang, M., Guo, J.: In-Context Meta LoRA Generation. In: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. pp. 6138–6146. International Joint Conferences on Artificial Intelligence Organization, Jeju, South Korea (Aug 2024). https://doi.org/10.24963/ijcai.2025/683, https://www.ijcai.org/proceedings/2025/683
28. Smith, E.A., Senter, R.J.: Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories (U.S.) pp. 1–14 (May 1967)
29. Stacey Dawn, Hill Sophie, McCaffery Kirsten, Boland Laura, Lewis Krystina B., Horvat Lidia: Shared Decision Making Interventions: Theoretical and Empirical Evidence with Implications for Health Literacy. In: Studies in Health Technology and Informatics. IOS Press (2017). https://doi.org/10.3233/978-1-61499-790-0-263, https://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-789-4&spage=263&doi=10.3233/978-1-61499-790-0-263
30. Suarez-Lledo, V., Alvarez-Galvez, J.: Prevalence of Health Misinformation on Social Media: Systematic Review. Journal of Medical Internet Research **23**(1), e17187 (Jan 2021). https://doi.org/10.2196/17187, http://www.jmir.org/2021/1/e17187/

31. Swanson, K., He, S., Calvano, J., Chen, D., Telvizian, T., Jiang, L., Chong, P., Schwell, J., Mak, G., Lee, J.: Biomedical text readability after hypernym substitution with fine-tuned large language models. PLOS Digital Health **3**(4), e0000489 (Apr 2024). https://doi.org/10.1371/journal.pdig.0000489, https://dx.plos.org/10.1371/journal.pdig.0000489
32. Team, M.A.: Mistral Small 3 | Mistral AI, https://mistral.ai/news/mistral-small-3
33. Trübner, M., Patzina, A., Lehmann, J., Brinkhaus, B., Kessler, C.S., Hoffmann, R.: Health information-seeking behavior among users of traditional, complementary and integrative medicine (TCIM). BMC Complementary Medicine and Therapies **25**(1), 111 (Mar 2025). https://doi.org/10.1186/s12906-025-04843-9, https://doi.org/10.1186/s12906-025-04843-9
34. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D.: Learning from the worst: Dynamically generated datasets to improve online hate detection. In: ACL (2021)
35. Wu, J., Wu, X., Qiu, Z., Li, M., Lin, S., Zhang, Y., Zheng, Y., Yuan, C., Yang, J.: Large language models leverage external knowledge to extend clinical insight beyond language boundaries. Journal of the American Medical Informatics Association **31**(9), 2054–2064 (Sep 2024). https://doi.org/10.1093/jamia/ocae079, https://academic.oup.com/jamia/article/31/9/2054/7659846
36. Wu, T.C., Shih, H., Nattam, A., Chintalapalli, H., Hanauer, D.A., Zheng, K., Wu, D.T.: Readability Assessment and Comparison of Large Language Model-Generated Summaries of Trial Descriptions on ClinicalTrials.gov. In: Househ, M.S., Tariq, Z.U.A., Al-Zubaidi, M., Shah, U., Huesing, E. (eds.) Studies in Health Technology and Informatics. IOS Press (Aug 2025). https://doi.org/10.3233/SHTI250982, https://ebooks.iospress.nl/doi/10.3233/SHTI250982
37. Wu, X., Arase, Y.: An In-depth Evaluation of Large Language Models in Sentence Simplification with Error-based Human Assessment. ACM Transactions on Intelligent Systems and Technology p. 3744744 (Jun 2025). https://doi.org/10.1145/3744744, https://dl.acm.org/doi/10.1145/3744744
38. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics **4**, 401–415 (Dec 2016). https://doi.org/10.1162/tacl_a_00107, https://direct.mit.edu/tacl/article/43364
39. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, year = 2024
40. Yang, H., Zhang, Y., Xu, J., Lu, H., Heng, P.A., Lam, W.: Unveiling the Generalization Power of Fine-Tuned Large Language Models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 884–899. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.51, https://aclanthology.org/2024.naacl-long.51
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr