Evaluating Health Question Answering Under Readability-Controlled Style Perturbations

Md Mushfiqur Rahman and **Kevin Lybarger**
George Mason University
Fairfax, VA
{mrahma45,klybarge}@gmu.edu

Abstract

Patients often ask semantically similar medical questions in linguistically diverse ways that vary in readability, tone, and background knowledge. A robust question answering (QA) system should both provide semantically consistent answers across stylistic differences and adapt its response style to match the user's input; however, existing QA evaluations rarely test this capability, creating critical gaps in QA evaluation that undermine accessibility and health literacy. We introduce SPQA, an evaluation framework and benchmark that applies controlled stylistic perturbations to consumer health questions while preserving semantic intent, then measures how model answers change across correctness, completeness, coherence, fluency, and linguistic adaptability using a human-validated LLM-based judge. The style axes include reading level, formality, and patient background knowledge; all perturbations are grounded in human annotations to ensure fidelity and alignment with human judgments. Our contributions include a readability-aware evaluation methodology, a style-diverse benchmark with human-grounded perturbations, and an automated evaluation pipeline validated against expert judgments. Evaluation results across multiple health QA models indicate that stylistic perturbations lead to measurable performance degradation, even when semantic intent is preserved during perturbation. The largest performance drops occur in answer correctness and completeness, while models also show limited ability to adapt their style to match the input. These findings underscore the risk of inequitable information delivery and highlight the need for accessibility-aware QA evaluation.

1 Introduction

Large Language Models (LLMs) have rapidly become central to consumer-facing question-answering (QA) systems, offering users quick and interactive access to information across a wide range of domains (Yu et al., 2024; Chiang et al., 2024; He et al., 2025). With this growing adoption, these systems are increasingly relied upon in critical areas such as healthcare, where users expect accurate and actionable guidance. However, as the user base becomes more diverse, linguistic variability in queries arising from differences in literacy, tone, and background knowledge presents a substantial challenge (Epner and Baile, 2012; Vela et al., 2022). Such diversity often affects the accessibility and reliability of responses, creating disparities in the quality of information retrieved. Despite its practical significance, this issue remains largely understudied, and existing evaluations rarely assess whether QA models can adapt to stylistic and readability differences, even when the underlying information need is unchanged.

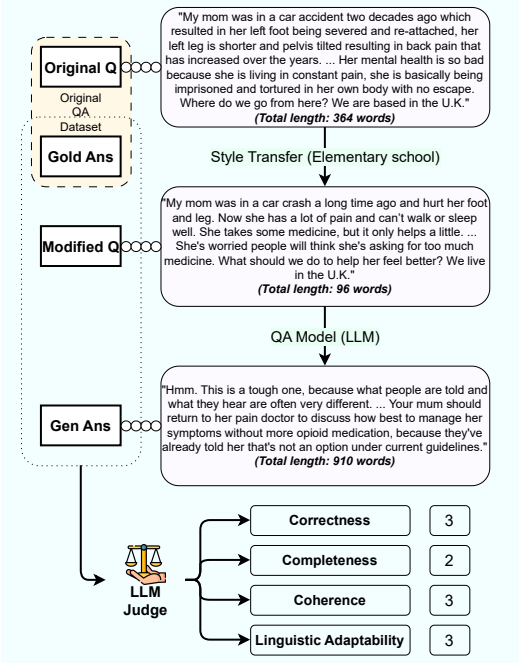


Figure 1: Example of the Style Perturbed Question Answering (SPQA) task

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 70–86
November 4-9, 2025

While prior research has documented that demographic and stylistic factors influence model outputs (Qu and Wang, 2024; Gosavi et al., 2024), most evaluations have focused on narrow dimensions such as adversarial perturbations or typographical noise (Gan et al., 2024; Li et al., 2024; Wang et al., 2021). These approaches do not address real-world variability in question phrasing that affects user comprehension and system accessibility. Current QA assessments typically emphasize correctness and completeness but overlook whether responses maintain quality and align with the style of diverse user queries. We term this as *linguistic adaptability*, and it is essential for ensuring accessible information delivery and mitigating inequities in high-stakes domains like health communication.

To address this gap, we introduce Style Perturbed Question Answering (SPQA), an evaluation framework and benchmark (Figure 1). SPQA systematically perturbs user questions along predefined stylistic axes - A) reading level, B) formality, and C) domain knowledge, while preserving semantic intent. For each perturbed question, responses are evaluated against four comprehensive criteria: correctness, completeness, coherence and fluency, and linguistic adaptability. All perturbations are grounded in human annotations to ensure realism, and evaluations are conducted using a human-validated LLM-as-Judge for scalability and reliability. SPQA serves as a general framework for evaluating QA robustness under stylistic variation and provides a structured protocol for stress-testing QA systems under stylistic diversity, operationalizing accessibility as an evaluation dimension. In this work, we apply SPQA to consumer health question answering, where the selected stylistic axes capture key factors influencing comprehension and accessibility. Our key contributions are summarized below:

1. Readability-Aware Evaluation Framework: We introduce SPQA [1], a framework for evaluating QA performance under readability- and style-controlled perturbations, addressing an underexplored dimension of accessibility in QA.

2. Human-Grounded Perturbations with Automated Evaluation: We generate stylistic question variants informed by human annotations and evaluate responses using GPT-4o as an automated judge, validated against expert annotations.

3. Comprehensive LLM Benchmarking: We benchmark major LLMs (Llama, DeepSeekR1, Qwen, and Phi) across multiple configurations, revealing their performance sensitivities to linguistic perturbations.

4. Focus on Consumer Health: We apply SPQA specifically to consumer health QA, emphasizing implications for health literacy, accessibility, and equity in medical information provision.

SPQA provides a systematic approach to evaluating QA systems under stylistic and linguistic variations, extending efforts in text simplification and readability assessment. Our experiments across ten state-of-the-art LLMs reveal that stylistic perturbations lead to measurable and statistically significant performance degradation, even when the underlying question intent remains unchanged. The largest declines occur in correctness and completeness, while linguistic adaptability remains a persistent challenge, indicating that models often fail to align their response style with user phrasing. Performance drops are particularly pronounced for graduate-level and expert-style variants, underscoring risks for both low-literacy and highly specialized users. These findings highlight the urgency of accessibility-aware evaluations to ensure equitable health information delivery.

2 Related Work

2.1 Open-ended QA Benchmarks for LLMs

LLMs are evaluated using a range of benchmarks that assess language understanding (Hendrycks et al., 2020; Bommasani et al., 2023), factual knowledge (Lin et al., 2021; Kwiatkowski et al., 2019; Thorne et al., 2018), reasoning (Zellers et al., 2019; Ghazal et al., 2017), and question answering (Abacha et al., 2017). While QA models frequently use multiple-choice question datasets like ARC (Clark et al., 2018), benchmarks targeting open-ended QA for practical, real-world applications remain limited because of the difficulty in evaluation. The prominent open-ended QA works (Yen et al., 2023; Prabhu and Anand, 2024; Shah et al., 2024). have used benchmarks like MT-Bench (Bai et al., 2024) for dialogue coherence and Chatbot Arena (Chiang et al., 2024) for pairwise response ranking. Medical QA benchmarks prioritize accuracy and clinical reliability. Notable datasets include MedQA (Jin et al., 2020), and PubMedQA (Jin et al., 2019). MedRedQA (Nguyen et al., 2023), the QA dataset we used in experimentation, evaluates

[1] Code: github.com/mushfiqur11/spqa

responses to consumer-driven medical inquiries from Reddit, making it highly relevant for studying linguistic and stylistic variability in real-world health queries.

To address robustness, literature use techniques like, adversarial attacks (Huang et al., 2024; Singh et al., 2024), and specialized frameworks like RITFIS (Walsh et al., 2024). However, existing benchmarks rarely assess whether QA models maintain performance under stylistic variation, an essential dimension of accessibility and linguistic robustness.

2.2 Evaluation Criteria and Accessibility

Evaluation in QA traditionally emphasizes correctness, completeness, and coherence (Yalamanchili et al., 2024; Liu et al., 2023a), while medical QA additionally incorporates trustworthiness (Zhu et al., 2020). Literature has explored simplifying biomedical text for lay readers (Shardlow et al., 2024; Ondov et al., 2022; Rahman et al., 2024; Štajner et al., 2022) and transferring domain-specific language into more comprehensible forms. Such work advances text generation for readability but leaves unanswered how QA models respond to inputs that vary in readability and style, which SPQA explicitly evaluates.

2.2.1 Automated Metrics and LLM-Judge

Traditional QA metrics like BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) rely on n-gram overlap, limiting their ability to capture deeper semantic nuances or stylistic alignment. Embedding-based measures, like BERTScore (Zhang* et al., 2020), incorporate contextual embeddings but primarily measure semantic similarities in topics and themes rather than information accuracy. Recently, LLM-based evaluators have shown strong alignment with human judgments (Chiang et al., 2024; Bai et al., 2024; Dubois et al., 2024). Chatbot Arena (Chiang et al., 2024), MT-Bench (Bai et al., 2024), and AlpacaEval (Dubois et al., 2024) utilize LLM-based ranking systems for dialogue evaluation. GPT-4 has demonstrated moderate to strong correlation with human ratings in natural language generation tasks, with Spearman coefficients around 0.51–0.65 (Liu et al., 2023b; Sottana et al., 2023) and high interrater reliability with intraclass correlation (ICC) scores between 0.94 and 0.99 (Hackl et al., 2023). These findings suggest that LLM-Judge setups can serve as practical and scalable proxies for human assessment. However, applying such evaluators in readability-aware and health-sensitive QA contexts remains underexplored. Unlike prior LLM-based evaluation frameworks that focus primarily on general response quality or user preference, SPQA extends the LLM-as-Judge paradigm to explicitly address readability- and accessibility-aware QA performance using a human-validated setup that combines scalability with rigor informed by domain expertise and health question answering.

3 Methods

3.1 Dataset

For dataset preparation, we utilized MedRedQA (Nguyen et al., 2023), a large QA dataset comprising 51,000 consumer questions and their corresponding expert answers. During initial inspection, we identified a small number of incomplete questions or entries missing answers. To ensure data quality, we dropped the entries with incomplete questions or ill-formatted or incomplete expert answers. We randomly sampled 470 data points from 1000+ such clean QA pairs. Since the answers in the original dataset are expert verified or expert generated, we used these answers as the gold standard in our experiments.

We split our filtered dataset (of 470 samples) into two parts: `SYSTEM-VAL` (N=120) and `QA-BENCH` (N=350). In the `SYSTEM-VAL` subset, each of the 120 samples was assigned one of the eight perturbation types, resulting in 15 instances per perturbation type. These samples were used to validate the style transfer process and LLM-Judge (see §3.4.1). The `QA-BENCH` subset includes 350 unique original questions, each transformed into all eight stylistic variations, alongside the original version, totaling 3,150 QA pairs.

3.2 Task Formulation

The primary objective of QA systems is to generate accurate, informative, and contextually appropriate responses to user questions. Formally, this QA task is represented as the mapping function:

$$f : Q \rightarrow A' \tag{1}$$

where f denotes an LLM-based QA model that generates an answer A' given an input question Q. The quality of the generated answer is evaluated via a scoring function g, which compares the model-generated answer A' against a gold-standard, expert-validated answer A_{gold}:

Criteria	Definition (This Work)	Prior Work and Their Definition
Correctness	Measures the factual correctness and accuracy of the LLM generated response considering the gold answer as factually correct.	Literature defines correctness as the factual alignment of generated responses with ground-truth data in QA tasks (Adlakha et al., 2024; Yalamanchili et al., 2024; Scialom et al., 2021).
Completeness	Evaluates what portion of the question is fully answered by the LLM-generated response.	Literature examines the comprehensiveness of long-form answers, analyzing whether the responses fully address the posed questions without omitting essential information (Yalamanchili et al., 2024; Xu et al., 2023; Scialom et al., 2021).
Coherence and Fluency	Assesses the grammatical correctness and logical coherence of the generated response.	In literature, coherence is defined as response consistency, while fluency is defined as grammatical correctness and naturalness (Zhong et al., 2022).
Linguistic Adaptability	Measures how well an LLM adjusts its response based on variations in tone, and user expertise while preserving factuality.	No prior works systematically define this; our study introduces this criterion to assess LLM robustness to stylistic perturbations.

Table 1: Evaluation criteria used in this study for the perturbed QA task (See §A for details)

$$g(Q, A_{gold}, A') \tag{2}$$

To systematically evaluate how linguistic variations affect QA performance, we formulate a modified QA task by linguistically perturbing the original question Q, generating a transformed question Q^*. The new task now becomes:

$$st : Q \rightarrow Q^* \implies f^* : Q^* \rightarrow A' \tag{3}$$

consequently, the evaluation function is adjusted accordingly:

$$g(Q^*, A_{gold}, A') \tag{4}$$

Importantly, while Q^* differs from the original question in phrasing, tone, complexity, or style, the semantic intent remains constant. The gold-standard answer A_{gold} is based on the original question Q, emphasizing the necessity to verify the model-generated answer remains accurate and complete, while appropriately adapting its linguistic style to the perturbed input.

3.3 Automated Style Transfer (AST)

The SPQA framework is broadly applicable across various QA domains, with the specific linguistic styles requiring careful selection based on the target task and domain context. Because relevant linguistic styles vary significantly by domain, each application of SPQA must identify style dimensions critical to effective communication within that context.

In this study, we specifically apply SPQA to consumer health QA, given the critical importance of providing medically accurate, reliable, and easily understandable health information to diverse user populations. To systematically assess QA robustness within this domain, we selected three linguistic dimensions, for which we identified eight distinct style variations: *reading level*, *formality spectrum*, and *domain-knowledge level* (see Table 2). These dimensions were specifically selected for their relevance to the consumer health context and their known influence on information accessibility and health literacy.

For the reading level dimension, we used four commonly employed sub-categories spanning a broad range of text complexity levels: elementary, middle school, high school, and graduate school (Petersen and Ostendorf, 2007; Balyan et al., 2020). Variations in formality (*formal* vs. *informal*) and domain knowledge (*domain expert* vs. *layperson*) were similarly incorporated to reflect the realistic range of ways consumers engage with health information, from casual and accessible to highly specialized and formal. Additional or alternative stylistic dimensions can be integrated based on the specific QA task or domain context.

The linguistic perturbations were generated via a zero-shot prompting approach utilizing `GPT-4o-2024-08-06`. Given an original question Q, the model produced transformed versions Q^* that preserved the semantic intent while varying linguistically according to the specified stylistic criteria.

3.3.1 Validating AST Framework

We validated each perturbation through a rigorous human validation process involving five health-informatics from George Mason University's College of Public Health (Appendix §B.1). Each perturbed question Q^* in the `SYSTEM-VAL` subset was

Domain	Category	Definition
Grade levels	elementary	Text written with very basic vocabulary and simple sentence structures, as used by an elementary school student.
	middle	Text written with basic but varied vocabulary and slightly longer sentences, reflecting a middle school student's style.
	high	Text featuring advanced vocabulary and sentence structures typical of a high school student.
	graduate	Text employing specialized terminology and dense, academic sentences characteristic of a graduate student.
Formality spectrum	formal	Text using precise grammar and elevated word choice appropriate for a professional report.
	informal	Text using casual phrasing and contractions common in everyday conversation.
Domain knowledge levels	domain-expert	Text incorporating field-specific terms and explanations suited for subject-matter experts.
	layperson	Text using everyday vocabulary and clear explanations geared toward a general audience.

Table 2: Definitions of each style transfer category

at first doubly annotated and then independently adjudicated for evaluation on a 3-point Likert scale using two criteria:

- **Style Transfer Success:** The degree to which the intended linguistic transformation (e.g., adjusting formality or reading level) was successfully implemented.
- **Meaning Preservation:** The extent to which the original medical meaning and intent of the question were preserved after perturbation.

During annotation, the annotators were not told what specific stylistic perturbation was performed on a given sample. This quality-control step ensured that observed performance differences across perturbations genuinely reflected model sensitivity to linguistic variations rather than unintended semantic changes.

3.4 LLM-Judge

A comprehensive and scalable evaluation of LLM-based QA systems using the SPQA framework requires an automated evaluation approach closely aligned with human judgments. To achieve this, we implemented an automated evaluation mechanism using GPT-4o as an *LLM-Judge*. Each generated answer was compared with the gold answer associated with the original question and assessed based on four criteria: *correctness*, *completeness*, *coherence and fluency*, and *linguistic adaptability*. Table 1 provides detailed definitions of these criteria. *Correctness* assesses the factual accuracy of the model-generated response, using the gold-standard answer as the reference. *Completeness* assesses the extent to which the generated answer fully addresses the information needs expressed in the question. *Coherence and fluency* assess the grammatical quality, clarity, and logical flow of the generated answer. These three criteria are widely used in literature. *Linguistic adaptability*, a new criterion introduced in this study, evaluates how effectively a system adjusts the tone, formality, and style of its responses to align with the linguistic style of the input questions. Within health contexts, including patient-facing applications and educational tools, misaligned tone or style can undermine comprehension and negatively impact user experience (Okoso et al., 2025). Incorporating linguistic adaptability into our evaluation extends conventional QA assessment beyond factual and structural quality to include responsiveness to user style and context, thereby advancing the accessibility and inclusiveness of QA systems.

Each criterion is scored using a standardized 3-point Likert scale (1–3). Figure A presents the final zero-shot prompt used in the system. This prompt was refined based on 20 selected samples from the `SYSTEM-VAL` subset. Using these criteria and the LLM-as-Judge setup, we evaluated the outputs of 10 different LLMs from four model families as consumer health QA systems.

3.4.1 Validating LLM-Judge

We evaluated the reliability of the automated LLM-as-Judge through an annotation study conducted with three medical student annotators (Appendix §)[2]. Annotators evaluated 120 selected QA pairs, each comprising a stylistically perturbed question (Q^*), the original expert answer (A_{gold}), and the model-generated answer (A'), using the same four evaluation criteria and Likert scale as the LLM-

[2]This is different from the annotation described in §3.3.1

Judge. Annotation occurred in four rounds: an initial calibration round, where each annotator evaluated eight samples followed by a training session to align scoring practices, and three subsequent rounds. The resulting 120 annotated samples were randomly split into two subsets, with 20 samples reserved for refining the LLM-Judge prompt (see §4.2 for results). To ensure the reliability of the automated evaluation, the LLM-Judge was tested using the held-out validation set (100 samples) that was never seen during prompt tuning. The judge operated under blinded conditions, where it was not informed whether a question was the original or style-transferred version. This procedure minimized bias and potential information leakage. Evaluation prompts were designed to independently compare generated and gold answers for each question instance. This structured process ensures rigorous assessment of the automated evaluation mechanism, enabling reliable identification of LLM strengths and weaknesses across realistic linguistic variations.

3.5 QA Benchmarking and Exp. Setup

Using our SPQA framework, we evaluated ten state-of-the-art LLM variants from four LLM families: Phi-4, Llama3, Qwen3, and DeepSeek-R1-Distilled[3]. Each model generated answers for the same set of 350 consumer health questions in their original forms and across eight stylistically transformed variants, resulting in 3,150 total generated answers per model. Each model's generated response was compared against the gold-standard expert answer associated with the original question, regardless of stylistic perturbation. This ensured that all evaluations measured factual consistency and completeness relative to the same ground truth rather than stylistic similarity alone. Responses were evaluated using GPT-4o as an automated judge, scoring each answer on four criteria, *correctness*, *completeness*, *coherence*, and *linguistic adaptability*, using a 3-point Likert scale. These 3-point Likert scores were scaled and normalized to a 0-1 scale for ease of comparison.

All models were evaluated in a zero-shot setting via HuggingFace without additional fine-tuning. Inference was performed on an A100 GPU (80 GB VRAM), with average runtime per variant of approximately five hours for larger models and two hours for smaller ones.

[3]For the DeepSeek model, we exclusively utilized locally downloaded pretrained weights without employing any external API, in compliance with institutional and state requirements.

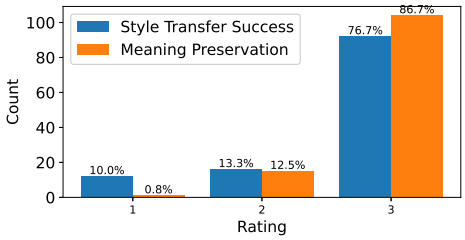


Figure 2: Distribution of ratings for Question Style Transfer Validation where 3 indicates successful, 2 indicates somewhat successful and 1 indicates failure

4 Results

4.1 AST Validation Results

Figure 2 presents the final adjudicated results from validating the stylistic transformations applied specifically to the questions. The results demonstrate that only 10.0% of the style-transferred questions did not fully achieve the desired stylistic modifications, and just 0.8% failed to retain the original meaning of the question. The high validation success indicates that the style transfer process reliably preserves meaning while effectively applying the intended stylistic modifications.

4.2 LLM-Judge Validation Results

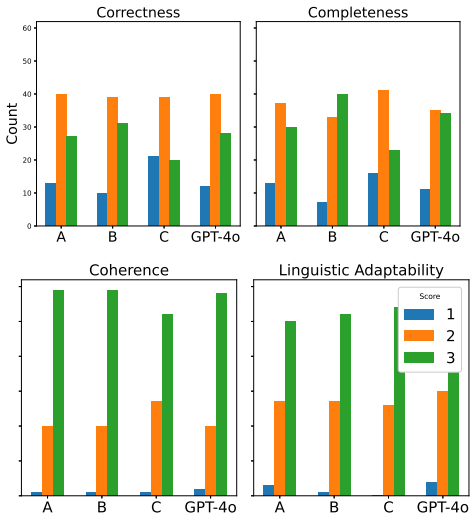


Figure 3: Score distribution for the human annotators (marked as A, B, and C) and the LLM-Judge (GPT-4o) across the four evaluation criteria, indicating similar scoring patterns between humans and the LLM-Judge

The moderate agreement score (0.47 Pearson correlation) among the human annotators (see Ta-

Agreement Type	Pearson Correlation (r)	Cohen's Kappa (κ)
Human vs. Human (avg)	0.47	0.39
Human vs. GPT-4o (LLM-Judge)	0.36	0.33
Human vs. Llama3-70B-Inst.	0.23	0.18

Table 3: The agreement scores between human experts and the LLM-Judge are moderate. Human vs human agreement and human vs LLM-Judge agreement are quite similar indicating reliability of performance from the LLM-Judge

ble 3) indicate the inherent complexity and subjectivity involved in evaluating nuanced linguistic adaptations in open-ended QA and medical QA contexts. So, GPT-4o's score of 0.36 when compared to human judgment (which is also considered as a moderate correlation (Kuckartz et al., 2013)) makes it a decent choice for LLM-Judge. The Cohen's Kappa among humans (0.39) and between GPT-4o and humans (0.33) are not too far as well which further strengthens the claim for GPT-4o as an LLM-Judge.

Inter-annotator agreement among human annotators, as well as alignment between human annotators and the automated LLM-Judge, was assessed using Pearson correlation coefficients and Cohen's Kappa scores.

Figure 3 presents the distribution of Likert scores for human annotators and the LLM-Judge across each evaluation criterion. This comparative analysis supports the reliability and suitability of the LLM-Judge for automated evaluation in nuanced medical QA tasks.

4.3 QA Benchmarking

Overall Degradation Across Styles

Table 4 provides results from the best performing models from each LLM family (full table in §6). The table shows the normalized scores for the original questions and the performance change for each stylistic variant compared to the original scores. To assess the significance of this performance drop, we performed a paired t-test with the null hypothesis of no performance degradation. Fields marked with $*$ indicate statistically significant decreases ($p < 0.05$).

Across all models and metrics, the quality of the answers generated for stylistically altered questions significantly decreased (with statistical significance) compared to answers generated for original questions. These declines were most prominent for correctness and completeness, suggesting that models either misinterpreted the question or failed to provide adequate information. Linguistic adaptability, a criterion introduced in our SPQA framework to assess how well answer style matches question style, also showed substantial drops, suggesting models often fail to adjust their response style when question phrasing shifts. In contrast, coherence remained relatively stable, consistent with the known ability of LLMs to produce fluent text even when misinterpreting question intent.

Impact of Linguistic Axes

We further analyzed these performance drops to identify patterns. Figure 4 presents the average performance change across models, computed as the difference between the mean score on original questions and the mean score on stylistically altered variants. The results are grouped into two broader variants: (1) a simplified and informal style, averaging elementary, informal, and layperson variants; and (2) a formal and specialized style, averaging graduate, formal, and expert variants, representing advanced and specialized language usage.

As represented in the figure, the overall degradation in performance is higher in formal and specialized styles compared to simple and informal styles. This result was consistent for all ten LLM variants that we used in our experimentation.

Comparative Model Performance

All ten models demonstrated susceptibility to style-induced performance degradation, although the degree varied by model size and training approach. Larger models achieved higher baseline accuracy but were also more sensitive to stylistic perturbations. For example, `DeepSeek-R1-Distilled-Llama3-70B` achieved the highest baseline scores on original questions but experienced disproportionately greater performance drops under stylistic perturbations. Similarly, `DeepSeek-R1-Distilled-Qwen` experienced marked losses under formal and specialized styles, indicating brittleness despite its size. In comparison, `Llama3-70B-Instruct` and `Qwen-3`, though similar in size to their R1 counterparts, performed marginally better on linguistic adaptability.

Mid-sized models like `Phi-4` exhibited more stable performance across styles, albeit with lower baseline performance. `Qwen3-0.6B`, the smallest

			Drop in performance compared to original							
			Grade Level				Formality Spectrum		Domain-knowledge	
Model	Metric	Original	Elementary	Middle	High	Graduate	Informal	Formal	Layperson	Expert
DS-Llama3-70B†	Coherence	0.71	-0.06*	-0.04*	-0.05*	-0.08*	-0.03*	-0.08*	-0.06*	-0.12*
	Completeness	0.5	-0.04*	-0.05*	-0.05*	-0.07*	-0.04*	-0.05*	-0.03*	-0.11*
	Correctness	0.62	-0.04*	-0.04*	-0.06*	-0.07*	-0.04*	-0.06*	-0.03*	-0.11*
	Linguistic Ad.	0.63	-0.06*	-0.03*	-0.07*	-0.13*	-0.03*	-0.1*	-0.05*	-0.14*
DS-Qwen3-32B†	Coherence	0.73	-0.06*	-0.07*	-0.05*	-0.1*	-0.05*	-0.09*	-0.07*	-0.12*
	Completeness	0.48	-0.03*	-0.03*	-0.04*	-0.06*	-0.04*	-0.04*	-0.04*	-0.07*
	Correctness	0.61	-0.05*	-0.04*	-0.02*	-0.07*	-0.03*	-0.07*	-0.03*	-0.09*
	Linguistic Ad.	0.64	-0.07*	-0.06*	-0.03*	-0.13*	0.0	-0.11*	-0.05*	-0.11*
Phi4	Coherence	0.69	-0.03*	-0.03*	-0.05*	-0.06*	-0.02*	-0.07*	-0.01*	-0.08*
	Completeness	0.44	-0.02*	-0.01	-0.01	-0.05*	-0.03*	-0.04*	-0.01	-0.05*
	Correctness	0.56	-0.02	-0.01	-0.01	-0.04*	-0.02	-0.04*	-0.02	-0.05*
	Linguistic Ad.	0.66	-0.04*	-0.03*	-0.02	-0.08*	-0.01	-0.07*	-0.05*	-0.08*
Qwen3-32B†	Coherence	0.7	-0.06*	-0.05*	-0.03*	-0.09*	-0.04*	-0.08*	-0.04*	-0.09*
	Completeness	0.5	-0.02	-0.03*	-0.03*	-0.05*	-0.03*	-0.04*	-0.03*	-0.08*
	Correctness	0.61	-0.04*	-0.01	-0.02*	-0.05*	-0.03*	-0.04*	-0.04*	-0.07*
	Linguistic Ad.	0.64	-0.09*	-0.06*	-0.06*	-0.13*	-0.02	-0.08*	-0.05*	-0.09*

Table 4: Normalized mean scores of the best performing models from each family (Rounded to 2 Decimal Places). Except for a few cases, all models have performed worse in case of the linguistic variants compared to the original. († indicates 8-bit quantization). $*$ indicates statistical significance with $p < 0.05$. (See Figure 6 for full results and Figures 5, 6, and 7 for significance test results)

model, had the smallest absolute drop but also the lowest original performance. Interestingly, its resilience to informal and layperson styles may reflect its reduced specialization, leading to more consistent outputs (Yang et al., 2025).

These observations suggest that model scale and advanced training techniques (like Reinforcement Learning with Human Feedback (RLHF)), although beneficial for original phrasing, may amplify sensitivity to stylistic shifts. Instruction tuning may reinforce specific interaction norms that break down under atypical inputs.

Implications for Equity and Robustness

These results raise pressing concerns regarding QA robustness in real-world deployments. While the largest performance drops occurred with formal and expert-style queries, there was still notable degradation for simplified and informal styles. Users with low literacy or non-native speakers may frame queries in simplified or unconventional ways. Our findings show that such phrasing, though semantically equivalent, often results in lower answer quality. Conversely, expert users posing technically precise questions also receive degraded responses, an especially problematic outcome in clinical settings. The performance degradation likely stems from the models' reliance on surface-level linguistic patterns during fine-tuning, which reduces their ability to generalize across stylistically distinct but semantically equivalent inputs.

This dual vulnerability suggests that current LLMs may be more proficient with specific styles, likely shaped by standard web-based corpora and fine-tuning data that emphasize neutral, well-formed text (Cao et al., 2025). As a result, models fail to generalize across diverse communication styles, reducing their utility for a broad population.

5 Conclusion and Future Work

This study introduces SPQA, a framework and benchmark for evaluating linguistic robustness in question-answering systems under controlled stylistic perturbations. SPQA systematically assesses how stylistic variations in questions impact QA model performance across multiple evaluation dimensions. Unlike prior evaluations focused primarily on accuracy under standard inputs, SPQA captures a critical but overlooked dimension: the ability of models to deliver consistent, accessible answers across diverse linguistic contexts.

While broadly applicable, we applied SPQA to consumer health QA, revealing vulnerabilities in current LLMs when processing stylistic variations reflecting real-world linguistic diversity. Our experiments across ten state-of-the-art LLMs demonstrate that stylistic changes, even when semantic meaning is preserved, result in measurable per-

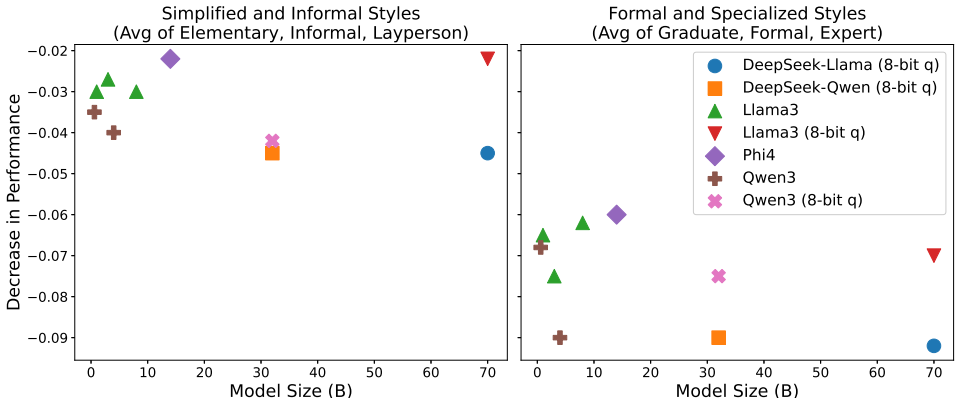


Figure 4: Average performance drop (across 4 metrics) for evaluated LLMs, indicating that larger models are more susceptible to performance degradation. Performance decline is more pronounced for formal and specialized stylistic variants compared to simplified styles

formance degradation in answer generation. The largest declines occur in correctness and completeness, and models frequently fail to align response style with question style. These findings reveal systematic risks to health information accessibility, affecting both users with limited literacy and expert users with specialized language needs, thereby reinforcing the urgency of equity-aware QA evaluations. SPQA provides a pathway toward addressing this gap by offering a systematic, human-grounded, and readability-aware evaluation protocol validated through LLM-as-Judge. Future research should extend SPQA to additional domains, including multimodal inputs, spoken interactions, and low-resource languages. Performance improvements may be achieved through adaptive prompting, style-diverse data augmentation, and patient-centered metrics. This work underscores the need for robust evaluation frameworks to ensure equitable access to reliable information for all. Future extensions should also incorporate readability-based user modeling to guide real-world deployment strategies.

Lay Summary

People ask medical questions in many different ways. Some use plain and simple words. Others use more formal wording or include medical terminology. This variability is important because an AI system might work well with one style but struggle with another. If an AI gives good answers to formal questions but fails for simple ones (or vice versa), then it may not serve all users equally.

To explore this issue, we create a formal process that can be followed step-by-step to evaluate how AI systems respond to changes in language style. We start with real medical questions and rewrite each into several versions that differ in reading level, tone, and technical detail, while keeping the original meaning. We call this benchmark SPQA. By doing this, we can test whether AI models remain accurate and helpful no matter how a question is phrased.

We evaluate ten leading language models on every rewritten version and compare their responses across four dimensions: correctness, completeness, fluency, and how well they match the style of the question. Our results show that style variation has a clear impact on answer quality. Models frequently lose accuracy and completeness when questions are highly formal or contain dense medical jargon, and the same question can produce different answers depending on how it is phrased. Some models are more sensitive to these shifts than others, and most do not adjust their response style to match the user's expression. These findings show that current systems handle writing styles unevenly, which may disadvantage some users.

Our work emphasizes the need for health question-answering systems that give reliable, inclusive, and understandable answers for everyone, regardless of how they phrase their questions.

Our work highlights the importance of developing health question-answering systems that provide reliable, fair, and easy-to-understand information for all users, regardless of how they write their questions.

Limitations

This study has several limitations. Methodologically, errors introduced during the style-transfer step could propagate through subsequent stages,

although validation showed that meaning was preserved in over 99% of cases. Occasional deviations from the intended style may still influence downstream outcomes. In addition, evaluating generated answers through human and LLM-based scoring introduces subjectivity. While the automated LLM-Judge achieved performance comparable to human annotators, moderate agreement levels suggest residual bias or inconsistency that could affect result validity. Future work should explore hybrid or multi-judge evaluation strategies to improve reliability.

Beyond these methodological constraints, the validation group consisted of medical students and domain experts, limiting sociolinguistic diversity and potentially affecting generalization across populations. The current evaluation also focuses on a single consumer health QA dataset; broader experimentation across datasets and domains is needed to establish the generalizability of the SPQA framework. Finally, reliance on GPT-4o for both style transfer and evaluation may introduce model-specific biases, underscoring the importance of future replication with diverse model architectures.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Renu Balyan, Kathryn S McCarthy, and Danielle S McNamara. 2020. Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education*, 30(3):337–370.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2025. Out of style: Rag's fragility to linguistic variation. *arXiv preprint arXiv:2504.08231*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457v1*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Daniel E Epner and Walter F Baile. 2012. Patient-centered care: the key to cultural competence. *Annals of oncology*, 23:iii33–iii42.

Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. Reasoning robustness of LLMs to adversarial typographical errors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10449–10459, Miami, Florida, USA. Association for Computational Linguistics.

Ahmad Ghazal, Todor Ivanov, Pekka Kostamaa, Alain Crolotte, Ryan Voong, Mohammed Al-Kateb, Waleed Ghazal, and Roberto V. Zicari. 2017. Bigbench v2: The new and improved bigbench. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1225–1236.

Purva Prasad Gosavi, Vaishnavi Murlidhar Kulkarni, and Alan F Smeaton. 2024. Capturing bias diversity in llms. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 593–598. IEEE.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In *Frontiers in Education*, volume 8, page 1272229. Frontiers Media SA.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are AI-generated text detectors robust to adversarial perturbations? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6005–6024, Bangkok, Thailand. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Udo Kuckartz, Stefan Rädiker, Thomas Ebert, and Julia Schehl. 2013. *Statistik: eine verständliche Einführung*. Springer-Verlag.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2024. Evaluating the instruction-following robustness of large language models to prompt injection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 557–568, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Ayano Okoso, Keisuke Otaki, Satoshi Koide, and Yukino Baba. 2025. Impact of tone-aware explanations in recommender systems. *ACM Trans. Recomm. Syst.*, 3(4).

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sarah Elizabeth Petersen and Mari Ostendorf. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplication for Bilingual Education*. Citeseer.

Venktesh V Deepali Prabhu and Avishek Anand. 2024. Dexter: A benchmark for open-domain complex question answering using llms. *arXiv preprint arXiv:2406.17158*.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shalin Shah, Srikanth Ryali, and Ramasubbu Venkatesh. 2024. Multi-document financial question answering using llms. *arXiv preprint arXiv:2411.07264*.

Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, editors. 2024. *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*. Association for Computational Linguistics, Miami, Florida, USA.

Ayush Singh, Navpreet Singh, and Shubham Vatsal. 2024. Robustness of llms to perturbations in text. *arXiv preprint arXiv:2407.08989*.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu, editors. 2022. *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual).

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. 2022. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43(1):477–501.

Matthew Walsh, David Schulker, and Shing hon Lau. 2024. Beyond Capable: Accuracy, Calibration, and Robustness in Large Language Models.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Amulya Yalamanchili, Bishwambhar Sengupta, Joshua Song, Sara Lim, Tarita O. Thomas, Bharat B. Mittal, Mohamed E. Abazeed, and P. Troy Teo. 2024. Quality of large language model responses to radiation oncology patient care questions. *JAMA Network Open*, 7(4):e244630–e244630.

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

Howard Yen, Tianyu Gao, Jinhyuk Lee, and Danqi Chen. 2023. MoQA: Benchmarking multi-type open-domain question answering. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 8–29, Toronto, Canada. Association for Computational Linguistics.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 895–900. IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A LLM Judge Prompts

Since we used a zero-shot LLM-Judge, it was essential to have a rigorously engineered prompt for different phases of our workflow. Appendix A represents the user-prompt provided to the LLMs to generate the answers to the questions. Appendix A represents the system prompt with all the necessary definitions provided to the LLM-Judge. These were also used as the base instructions for the annotators validating the LLM-Judge. Keeping the instructions same, we ensured fair ground for the LLM-Judge and human experts.

User-prompt for LLM-Judge

```
"""
Evaluate the following QA sample:
Modified Question (Q_mod): [SEP] {question} [SEP]
Generated Answer (A_gen): [SEP] {answer} [SEP]
Gold Standard Answer (A_gold): [SEP] {gold} [SEP]
"""
```

System-prompt for LLM-Judge

```
"""
You are a helpful assistant that evaluates medical QA samples.
For each sample, you must evaluate the generated answer (A_gen) as a response to the modified question (Q_mod). Use the gold standard answer (A_gold) as the medically accurate information regarding the topic.
In this QA task, the generated answer (A_gen) and the gold standard answer (A_gold) may come from different linguistic distribution. Do not penalize A_gen for being linguistically different from A_gold.

Use the following four criteria. For each criterion, assign a score from 0 (lowest) to 2 (highest).

1. Correctness: Considering A_gold as medically correct, what portion of A_gen is accurate in answering the Q_mod? A_gen can be significantly different from A_gold.
- Score Definitions:
    CR0: A_gen is completely incorrect. It does not have any medically accurate information.
    CR1: A_gen is mostly incorrect. It has very little medically correct advice or information.
    CR2: The generated answer is mostly correct.

2. Completeness: What portion of the queries made in Q_mod are answered by A_gen?
- Score Definitions:
    CM0: A_gen is absolutely irrelevant and does not answer any of the queries made in Q_mod.
    CM1: A_gen is somewhat incomplete, missing the essential details required to address Q_mod.
    CM2: A_gen answers Q_mod sufficiently. Important parts of the question in Q_mod is addressed by A_gen.

3. Fluency and Coherence: How well is A_gen written in terms of language fluency and logical structure?
- Score Definitions:
    FC0: A_gen is poorly written, with significant grammatical and structural issues.
    FC1: A_gen is somewhat fluent but contains noticeable issues and grammatical errors.
    FC2: A_gen is fluent and logically structured.

4. Linguistic Adaptability: How effectively does A_gen adopt the style and tone of the modified question (Q_mod)?
- Score Definitions:
    LA0: A_gen fails to match the tone or style of Q_mod and would be totally unreadable for the user asking Q_mod.
    LA1: A_gen somewhat matches the tone or style but would not be fully legible for the user asking Q_mod.
    LA2: A_gen is appropriate and easy to read for someone who asked the question Q_mod.
```

```
Return your evaluation in JSON format as follows:

{
    "correctness": <rating as an integer>,
    "completeness": <rating as an integer>,
    "fluency_and_coherence": <rating as an integer>,
    "linguistic_adaptability": <rating as an integer>
}

Ensure that your output contains only the JSON object.
"""
```

B Human Annotation Details

This appendix provides details of the two independent annotation streams that supported this work. Both streams involved the same set of 120 question–answer pairs but differed in purpose, annotator expertise, and evaluation criteria.

B.1 Annotation Stream 1: AST Validation

Objective: Assess whether stylistic perturbations (generated via GPT-4o) successfully applied the intended style changes (reading level, formality, patient background knowledge) while preserving original meaning.

Annotator Profile: Five health informatics graduate students with training in health data interpretation.

Procedure:

- Each question–variant pair was evaluated independently by two annotators and adjudicated by a third for disagreements.
- Annotations were based on two criteria using a 3-point Likert scale:
  1. *Style Transfer Success* – Did the variant reflect the assigned stylistic dimension?
  2. *Meaning Preservation* – Was the original medical intent maintained?
- Random audits (15% of samples) were conducted for quality control.

Outcome: High validation accuracy was achieved (Style Transfer Success: 76.7%, Meaning Preservation: 86.7%), confirming fidelity of stylistic transformations.

B.2 Annotation Stream 2: LLM-Judge Validation

Objective: Validate the reliability of GPT-4o as an automated judge by comparing its ratings to human annotations for QA responses.

Annotator Profile: Three medical students (clinical track) with prior training in patient communication.

Procedure:

- Annotators rated the same 120 question–answer pairs on four evaluation criteria:
  1. Correctness
  2. Completeness
  3. Coherence and Fluency
  4. Linguistic Adaptability
- Process included a calibration round (8 samples), followed by three main annotation rounds after a training session.
- Agreement metrics were computed using Pearson correlation and Cohen's Kappa.

Outcome: Human–LLM alignment showed moderate agreement (Pearson's $r = 0.36$, Cohen's $\kappa = 0.33$), supporting the use of LLM-as-Judge for scalable evaluation.

B.3 Summary of Annotation Resources

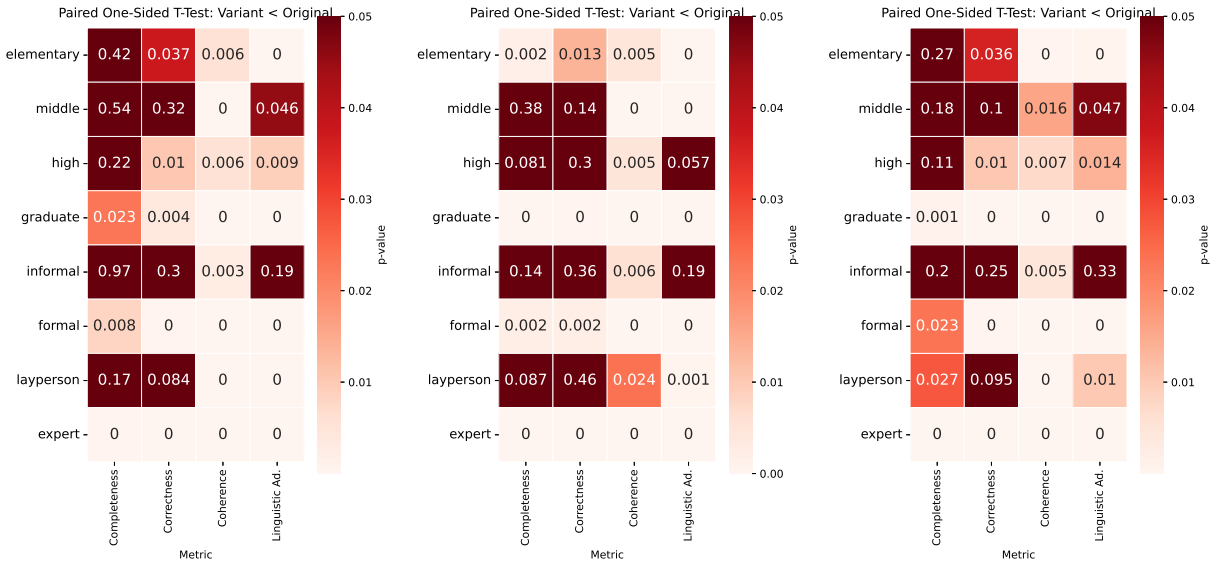
- Total samples annotated: 120 (used for both streams).
- Annotators: 5 health informatics students (style validation) and 3 medical students (evaluation validation).

Table 5: Summary of Human Annotation Streams

Stream	Annotator Group	Goal	Criteria
Style Transfer Validation	5 Health Informatics Students	Validate stylistic perturbations	Style Transfer Success, Meaning Preservation
Evaluation Rubric Validation	3 Medical Students	Validate automated evaluation rubric	Correctness, Completeness, Coherence and Fluency, Linguistic Adaptability

C Additional results

C.1 Significance Test



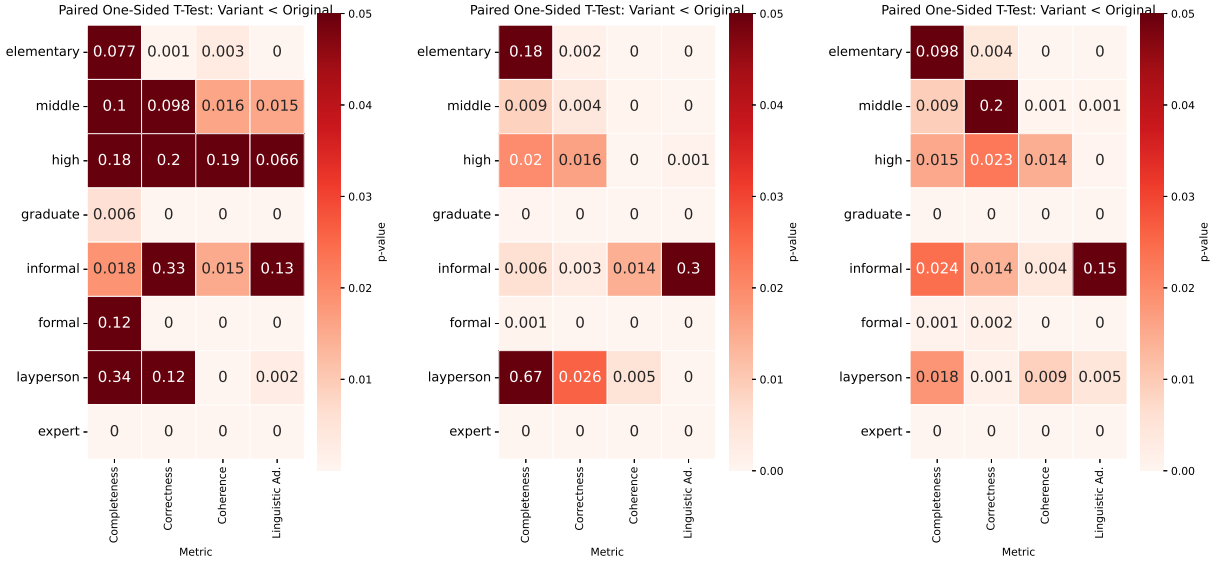
(a) `Llama3.2-1B-Instruct` (b) `Llama3.2-3B-Instruct` (c) `Llama3.1-8B-Instruct`

Figure 5: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)

Section 4 mentions that a significance test was performed. Figures 5, 6, and 7 represent heatmaps of the detailed results from the significance test.

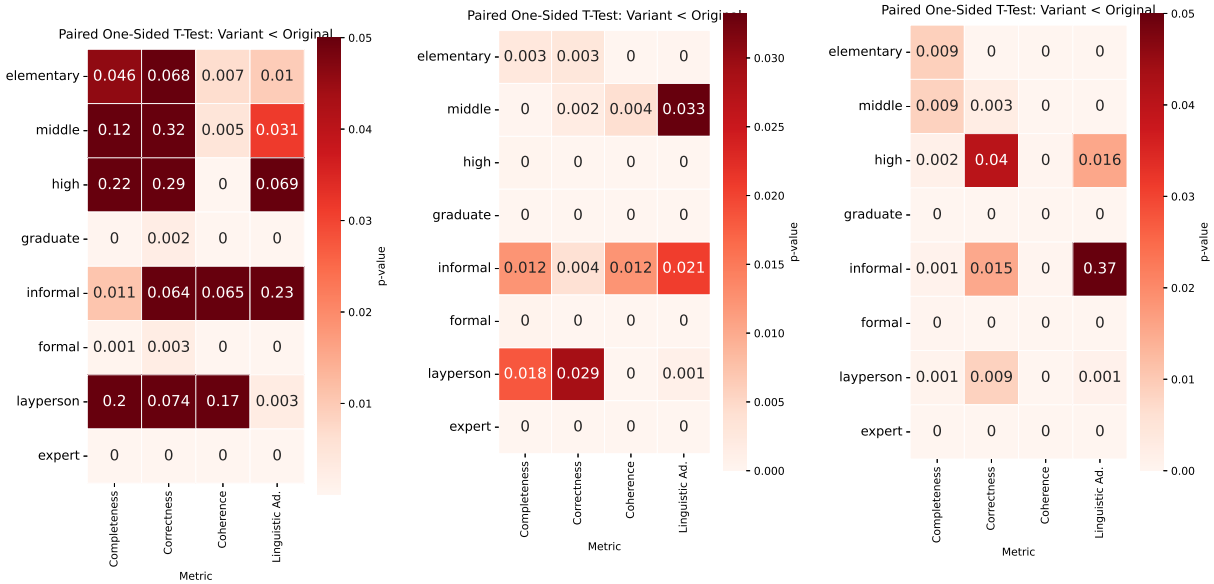
C.2 Full Result

Table 6 represents the complete results table with all the models we have used in our experimentation. A shorter and more concise version of this table has been presented in the main paper.



(a) Qwen3-0.6B (b) Qwen3-4B (c) Qwen3-32B

Figure 6: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)



(a) Phi4 (b) DeepSeek-R1-Distilled-Llama70B (c) DeepSeek-R1-Distilled-Qwen3-32B

Figure 7: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)

			Drop in performance compared to original							
			Grade Level				Formality Spectrum		Domain-knowledge	
Model	Metric	Original	Elementary	Middle	High	Graduate	Informal	Formal	Layperson	Expert
DS-Llama-70B†	Coherence	0.71	-0.06	-0.04	-0.05	-0.08	-0.03	-0.08	-0.06	-0.12
DS-Llama-70B†	Completeness	0.5	-0.04	-0.05	-0.05	-0.07	-0.04	-0.05	-0.03	-0.11
DS-Llama-70B†	Correctness	0.62	-0.04	-0.04	-0.06	-0.07	-0.04	-0.06	-0.03	-0.11
DS-Llama-70B†	Linguistic Ad.	0.63	-0.06	-0.03	-0.07	-0.13	-0.03	-0.1	-0.05	-0.14
DS-Qwen-32B†	Coherence	0.73	-0.06	-0.07	-0.05	-0.1	-0.05	-0.09	-0.07	-0.12
DS-Qwen-32B†	Completeness	0.48	-0.03	-0.03	-0.04	-0.06	-0.04	-0.04	-0.04	-0.07
DS-Qwen-32B†	Correctness	0.61	-0.05	-0.04	-0.02	-0.07	-0.03	-0.07	-0.03	-0.09
DS-Qwen-32B†	Linguistic Ad.	0.64	-0.07	-0.06	-0.03	-0.13	0.0	-0.11	-0.05	-0.11
Llama3-1B	Coherence	0.71	-0.04	-0.05	-0.03	-0.08	-0.04	-0.06	-0.06	-0.09
Llama3-1B	Completeness	0.41	0.0	0.0	-0.01	-0.02	0.03	-0.03	-0.01	-0.05
Llama3-1B	Correctness	0.54	-0.03	-0.01	-0.03	-0.04	-0.01	-0.05	-0.02	-0.06
Llama3-1B	Linguistic Ad.	0.67	-0.08	-0.03	-0.04	-0.11	-0.02	-0.07	-0.07	-0.11
Llama3-3B	Coherence	0.71	-0.04	-0.05	-0.04	-0.1	-0.04	-0.08	-0.03	-0.11
Llama3-3B	Completeness	0.43	-0.04	0.0	-0.02	-0.05	-0.01	-0.04	-0.01	-0.06
Llama3-3B	Correctness	0.54	-0.03	-0.01	-0.01	-0.05	0.0	-0.04	0.0	-0.06
Llama3-3B	Linguistic Ad.	0.68	-0.07	-0.06	-0.03	-0.1	-0.01	-0.09	-0.05	-0.12
Llama3-8B	Coherence	0.71	-0.05	-0.03	-0.03	-0.08	-0.04	-0.07	-0.06	-0.1
Llama3-8B	Completeness	0.43	-0.01	-0.01	-0.02	-0.04	-0.01	-0.03	-0.03	-0.05
Llama3-8B	Correctness	0.56	-0.03	-0.02	-0.03	-0.06	-0.01	-0.05	-0.02	-0.07
Llama3-8B	Linguistic Ad.	0.66	-0.05	-0.03	-0.04	-0.07	-0.01	-0.07	-0.04	-0.07
Llama3-70B†	Coherence	0.69	-0.04	-0.04	-0.01	-0.06	-0.02	-0.06	-0.03	-0.08
Llama3-70B†	Completeness	0.45	-0.01	-0.01	-0.01	-0.05	0.0	-0.05	0.0	-0.07
Llama3-70B†	Correctness	0.57	-0.03	-0.03	-0.02	-0.06	-0.01	-0.06	-0.02	-0.07
Llama3-70B†	Linguistic Ad.	0.67	-0.07	-0.03	-0.03	-0.08	-0.01	-0.08	-0.04	-0.11
Phi4	Coherence	0.69	-0.03	-0.03	-0.05	-0.06	-0.02	-0.07	-0.01	-0.08
Phi4	Completeness	0.44	-0.02	-0.01	-0.01	-0.05	-0.03	-0.04	-0.01	-0.05
Phi4	Correctness	0.56	-0.02	-0.01	-0.01	-0.04	-0.02	-0.04	-0.02	-0.05
Phi4	Linguistic Ad.	0.66	-0.04	-0.03	-0.02	-0.08	-0.01	-0.07	-0.05	-0.08
Qwen3-0.6B	Coherence	0.69	-0.04	-0.03	-0.01	-0.07	-0.03	-0.07	-0.06	-0.09
Qwen3-0.6B	Completeness	0.46	-0.02	-0.02	-0.01	-0.03	-0.03	-0.01	0.0	-0.07
Qwen3-0.6B	Correctness	0.59	-0.05	-0.02	-0.02	-0.06	-0.01	-0.05	-0.02	-0.08
Qwen3-0.6B	Linguistic Ad.	0.63	-0.08	-0.04	-0.03	-0.09	-0.02	-0.07	-0.06	-0.11
Qwen3-4B	Coherence	0.72	-0.07	-0.05	-0.06	-0.1	-0.03	-0.1	-0.04	-0.12
Qwen3-4B	Completeness	0.48	-0.02	-0.04	-0.03	-0.05	-0.04	-0.05	0.0	-0.07
Qwen3-4B	Correctness	0.62	-0.04	-0.04	-0.03	-0.08	-0.04	-0.06	-0.03	-0.1
Qwen3-4B	Linguistic Ad.	0.65	-0.09	-0.07	-0.05	-0.11	-0.01	-0.09	-0.06	-0.13
Qwen3-32B†	Coherence	0.7	-0.06	-0.05	-0.03	-0.09	-0.04	-0.08	-0.04	-0.09
Qwen3-32B†	Completeness	0.5	-0.02	-0.03	-0.03	-0.05	-0.03	-0.04	-0.03	-0.08
Qwen3-32B†	Correctness	0.61	-0.04	-0.01	-0.02	-0.05	-0.03	-0.04	-0.04	-0.07
Qwen3-32B†	Linguistic Ad.	0.64	-0.09	-0.06	-0.06	-0.13	-0.02	-0.08	-0.05	-0.09

Table 6: Full results table. † indicates models with 8-bit quantization.

D Declaration of use of Generative AI

During the preparation of this manuscript, the authors used ChatGPT to obtain editorial assistance focused on writing clarity and proofreading. All scientific content, including analyses and interpretations, was developed independently by the authors. The authors carefully reviewed and revised the text following the use of these tools and assume full responsibility for the integrity and accuracy of the final manuscript.