Evaluating Linguistic Robustness of Large Language Models for Question Answering: A Study on Consumer Health Queries

Anonymous ACL submission

Abstract

Question-answering (QA) systems powered by Large Language Models (LLMs) increasingly enable interactive access to essential information across diverse domains. However, the robustness of these systems to variations in linguistic style, such as differences in reading level, formality, or domain-specific terminology, remains underexplored. To systematically address this gap, we propose the Style Perturbed Question Answering (SPQA) framework. SPQA systematically perturbs original questions to produce linguistically diverse variants and evaluates model responses to both original and perturbed queries based on correctness, completeness, coherence, and linguistic adaptability. Given the critical importance of accessible and medically accurate health information, we specifically apply SPQA to consumer health QA. Using a scalable evaluation pipeline combining automated style-transfer methods with a rigorously validated GPT-4o-based automated evaluation approach, we benchmark several state-of-the-art LLMs. Our results demonstrate substantial performance declines under realistic stylistic perturbations, highlighting significant challenges related to equity, reliability, and robustness in consumer-facing QA systems, especially in sensitive domains like healthcare.

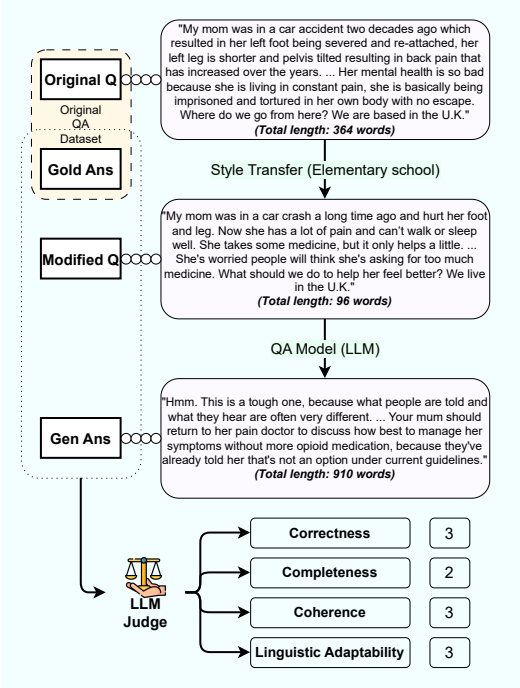


Figure 1: Example of the Style Perturbed Question Answering (SPQA) task. An original consumer health question is linguistically transformed into a specified style, creating a modified QA task. The generated answer (to this modified question) is then evaluated against the gold standard answer (to the original question) based on four criteria: correctness, completeness, coherence and fluency, and linguistic adaptability

1 Introduction

The integration of Large Language Models (LLMs) into consumer-facing question-answering (QA) systems has enabled new, interactive ways for users to access essential information across a broad range of contexts and domains (Yu et al., 2024; Chiang et al., 2024; He et al., 2025). These models interpret and respond to user queries, providing relevant advice and information. However, a substantial challenge remains: users frequently pose questions using diverse tones and linguistic styles, shaped by factors such as emotional state, cultural background, and varying domain literacy (Epner and Baile, 2012; Vela et al., 2022). This phenomenon is especially pronounced in the medical domain, where a broad and heterogeneous audience gives rise to even greater variability in tone (Wang and Zhang, 2024).

Prior research has demonstrated that demographic attributes such as gender, race, and age can lead to disparities in the quality of LLM-generated responses (Qu and Wang, 2024; Gosavi et al., 2024; Shin et al., 2024). Similarly, linguistic variations, including informal language and demographic-specific paraphrasing, adversely affect model com-

prehension, leading to inconsistent interpretations and responses (Arora et al., 2025). Additionally, LLMs are known to experience performance degradation when encountering typographical errors, adversarial attacks, and other forms of input perturbations, significantly impairing their reasoning capabilities (Gan et al., 2024; Li et al., 2024; Wang et al., 2021). These findings highlight the need for systematic evaluations to measure LLM robustness against diverse linguistic inputs, an area that remains underexplored within consumer-facing QA contexts.

To address this gap, we propose a novel evaluation framework: Style Perturbed Question Answering (SPQA) (as shown in Figure 1). SPQA systematically perturbs questions into predefined stylistic variations, generates responses to both the original and perturbed questions, and evaluates these responses according to four comprehensive criteria: correctness, completeness, coherence, and linguistic adaptability. We apply SPQA within the context of consumer health information, given the critical importance of medically accurate and reliable health information. The specific styles explored in this study include reading level, formality spectrum, and domain knowledge and were selected for their relevance to the medical domain and their known influence on information accessibility and health literacy. Our contributions to this research domain are as follows:

1. Robustness Evaluation Framework: We introduce SPQA, a novel framework to systematically evaluate LLM robustness against realistic linguistic variations, an underexplored yet critical aspect of QA.

2. Automated Evaluation with LLM-Judge: We leverage GPT-4o as an automated evaluator, extensively validated against expert human annotations, enabling scalable and reliable QA assessments.

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.

This study advances the understanding of linguistic robustness in QA systems broadly, with particular emphasis on critical challenges in consumer health information contexts. By demonstrating the susceptibility of current LLMs to realistic linguistic variations, our findings underscore significant equity concerns related to the accessibility and reliability of medical information. The proposed SPQA framework thus presents key opportunities to enhance health literacy, promote equitable information access, and ultimately improve health outcomes among diverse populations.

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 specifically targeting open-ended QA for practical, real-world applications remain limited. Recent benchmarks at addressing open-ended QA evaluation include MT-Bench (Bai et al., 2024) for dialogue coherence and Chatbot Arena (Chiang et al., 2024) for pairwise response ranking. There are few other open-ended QA benchmarks as well that focus on complex question answering (Yen et al., 2023; Prabhu and Anand, 2024; Shah et al., 2024). Testing the robustness of LLMs is also quite common. Few works use adversarial attacks (Huang et al., 2024; Singh et al., 2024), while frameworks like RITFIS (Walsh et al., 2024) evaluate model resilience to broader input variations.

2.1.1 Consumer Health QA

Medical QA benchmarks prioritize accuracy and clinical reliability. Notable examples include MedQA (Jin et al., 2020), which targets clinical reasoning, and PubMedQA (Jin et al., 2019), which emphasizes biomedical literature synthesis. MedRedQA, the QA dataset we used in experimentation, evaluates responses to consumer-driven medical inquiries from Reddit (Nguyen et al., 2023), making it particularly relevant to our exploration of consumer health information. Several other works have tried to solve the consumer health QA task (Demner-Fushman et al., 2020; Welivita and Pu, 2023).

2.2 Evaluation Criteria

General domain QA model evaluation typically assesses correctness, completeness, and coherence (Yalamanchili et al., 2024; Liu et al., 2023). Medical QA evaluation additionally considers trustworthiness (Zhu et al., 2020), given the high-six nature of health-related information. However, the concept of linguistic adaptability, measuring how effectively LLMs align their responses with variations in tone and style, remains underexplored, highlighting a significant gap addressed by our proposed SPQA framework.

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. More recent metrics, like BERTScore (Zhang* et al., 2020), incorporate contextual embeddings but primarily measure semantic similarities in topics and themes rather than information accuracy.

LLMs themselves have become increasingly popular as evaluators due to their demonstrated alignment with human judgments across benchmarks. 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. Despite evidence showing models like GPT-4 can reliably assess responses, significant challenges persist within specialized domains such as medical QA, where factual accuracy and nuanced interpretation are paramount.

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. We found few questions to be incomplete and few with missing answers. We randomly sampled questions that were complete and had clean answers. Since the answers in the original dataset are expert verified or expert generated, we used these answers as the gold standard in our experiments.

The resulting filtered dataset comprises 470 samples, split into two parts: `SYSTEM-VAL` and `QA-BENCH`. 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 validated 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}:

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

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, complete, and linguistically adaptable despite these perturbations.

3.3 Automated Style Transfer (AST)

3.3.1 AST Framework

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

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.	(Adlakha et al., 2024; Yalamanchili et al., 2024; Scialom et al., 2021) define correctness as the factual alignment of generated responses with ground-truth data in QA tasks.
Completeness	Evaluates what portion of the question is fully answered by the LLM-generated response.	(Yalamanchili et al., 2024; Xu et al., 2023; Scialom et al., 2021) examines the comprehensiveness of long-form answers, analyzing whether the responses fully address the posed questions without omitting essential information.
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, formality, 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 §6 for details)

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 employed four previously validated sub-categories representing a wide spectrum of reading 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. 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.

3.3.2 AST Validation

We validated each perturbation through a rigorous human validation process involving five health-informatics graduate students from a reputable university in the USA. Each perturbed question Q^* in the `SYSTEM-VAL` subset was 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 (of the original question) and assessed based on four criteria: *correctness*, *completeness*, *coherence and fluency*, and *linguistic adaptability*. Table 1 provides

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 complex 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 detailed explanations suited to subject-matter experts.
	layperson	Text using everyday vocabulary and clear explanations geared toward a general audience.

Table 2: Definitions of each style transfer category

detailed definitions of these criteria. *Correctness* measures the factual correctness and accuracy of the response, considering the gold-standard answer as factually correct. *Completeness* evaluates what portion of the question is fully addressed by the generated answer. *Coherence and fluency* assesses the grammatical correctness and logical coherence 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 allows us to systematically assess whether QA systems not only provide accurate and comprehensive answers but also context-sensitive responses, thereby enhancing accessibility and usability.

Each criterion is scored using a standardized 3-point Likert scale (1–3). Figure 6 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, we evaluated 10 different LLMs from four different model families.

3.4.1 Validation of LLM-Judge

To validate the reliability of our automated LLM-Judge, we conducted a structured annotation study involving three medical students as annotators. 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-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 and the remaining 100 samples used for validating its reliability (see §4.2 for results). 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[1]. 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. Responses were evaluated us-

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

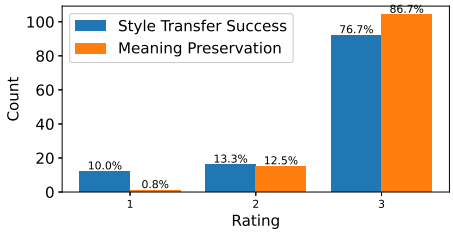


Figure 2: Distribution of ratings for Question Style Transfer Validation where 3 indicates successful, 2 indicates somewhat successful and 1 indicates failure

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. For this task, Llama has poor agreement with humans deeming it unsuitable for usage as an LLM-Judge

ing 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.

In our experiments, we used zero-shot prompting to the models using HuggingFace. Therefore, we did not require any fine-tuning step. We used an A100 GPU with 80GB VRAM for inference. The average inference time for the larger models was 5 hours for each variant. For smaller models, the inference time was around 2 hours per variant.

4 Results

4.1 Style Transfer 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 success rate in this validations confirms that our style transfer methods consistently preserves meaning and effectively performs the intended linguistic perturbation on the original questions.

4.2 LLM-Judge Validation Results

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. The observed values indicated moderate agreement (Kuckartz et al., 2013), reflecting the inherent complexity and subjectivity involved in evaluating nuanced linguistic adaptations open-ended QA and medical QA contexts.

Despite modest absolute agreement scores, the consistency between human annotators and the LLM-Judge indicates that the automated evaluation closely mirrors human judgment. 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 §5). 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 metrics and models there are statistically significant performance decreases.

Across all models and metrics, the quality of the answers generated for stylistically altered questions significantly decreased 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.

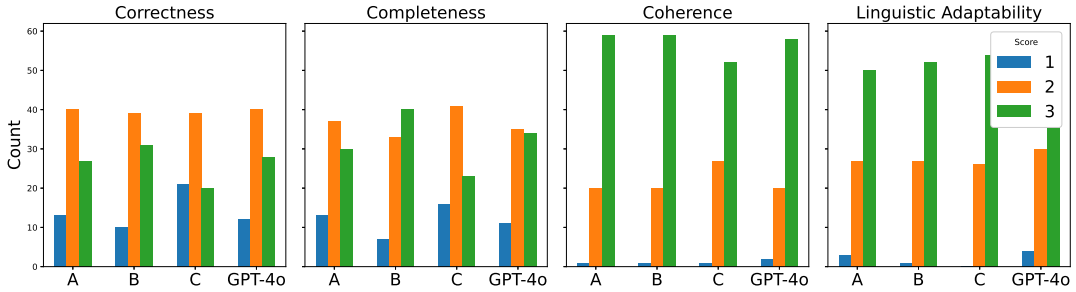


Figure 3: Score distribution for the human annotators (marked as A, B, and C) and the LLM-Judge (GPT-4o) across the four evaluation criteria, indicating similar scoring patterns between humans and the LLM-Judge

			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 statistically significance with $p < 0.05$. (See Figure 5 for full results and Figures 7, 8, and 9 for significance test results)

In contrast, coherence remained relatively stable, indicating that models maintain fluent output even when misinterpreting question intent. This is consistent with the known ability of LLMs to generate fluent text.

Impact of Linguistic Axes

We further analyzed these performance drops to identify patterns. Figure 4 presents the average performance drop 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. The largest models in each family achieved the highest scores but larger models were more vulnerable to performance drop. For example, `DeepSeek-R1-Distilled-Llama3-70B` achieved the highest baseline scores on original questions but experienced disproportionately greater performance drops under stylistic perturbations. Sim-

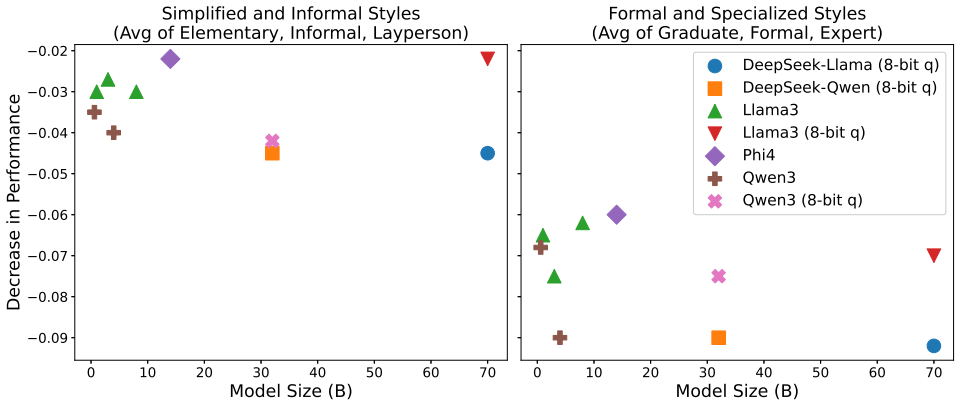


Figure 4: Average performance drop (across 4 metrics) for evaluated LLMs, indicating that larger models are more susceptible to performance degradation. Performance decline is more pronounced for formal and specialized stylistic variants compared to simplified styles

ilarly, `DeepSeek-R1-Distilled-Llama3-70B` experienced marked losses under expert and formal styles, indicating brittleness despite its size. In comparison, `Llama3-70B-Instruct`, though similar in size, performed marginally better on linguistic adaptability, potentially due to its additional instruction tuning.

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

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. 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 the SPQA framework, a systematic method for evaluating linguistic robustness in question-answering systems powered by LLMs. SPQA systematically assesses how stylistic variations in questions impact QA model performance across multiple evaluation dimensions. By rigorously validating both the automated style transformations and the automated evaluation mechanism against expert human annotation, this work establishes a robust foundation for comprehensive and scalable robustness evaluation in QA tasks.

While broadly applicable, we applied SPQA to consumer health QA, revealing vulnerabilities in current LLMs when processing stylistic variations reflecting real-world linguistic diversity. These findings raise concerns about robustness across diverse populations, particularly affecting those with limited health literacy. 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.

Limitations

This study has several limitations. First, errors introduced during the question style-transfer step could potentially cascade into subsequent stages. Although validation indicated that stylistic perturbations preserved original question meaning over 99% of the time, occasional failures in achieving exact stylistic adherence could still impact downstream results. Second, evaluating the quality of generated answers using human annotation revealed inherent subjectivity and ambiguity in judgments related to correctness, completeness, coherence, and linguistic adaptability. While the automated LLM-Judge demonstrated performance comparable to human evaluators, systematic errors or biases inherent to GPT-4o could influence evaluation outcomes, potentially affecting result validity. Third, the current evaluation is limited to a single consumer health QA dataset. Additional experiments across other datasets and application domains are necessary to fully assess the generalizability and robustness of the SPQA framework. Finally, the reliance on a single pretrained model (GPT-4o) for both stylistic perturbations and evaluation may introduce implicit biases or performance limitations unique to that model, warranting future assessments with additional models.

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.

Pulkit Arora, Akbar Karimi, and Lucie Flek. 2025. Exploring robustness of llms to sociodemographically-conditioned paraphrasing. *arXiv preprint arXiv:2501.08276*.

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.

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*.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

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.

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. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 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).

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.

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*.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

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

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*.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*, 57(11):299.

Anuradha Welivita and Pearl Pu. 2023. A survey of consumer health question answering systems. *Ai Magazine*, 44(4):482–507.

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.

Figure 5 represents the prompt provided to the LLMs to generate the answers to the questions. Figure 6 represents the prompt 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.

Medical QA Evaluation Prompt (USER MESSAGE)

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

Figure 5: Prompt for QA Models

Medical QA Evaluation Prompt (SYSTEM MESSAGE)

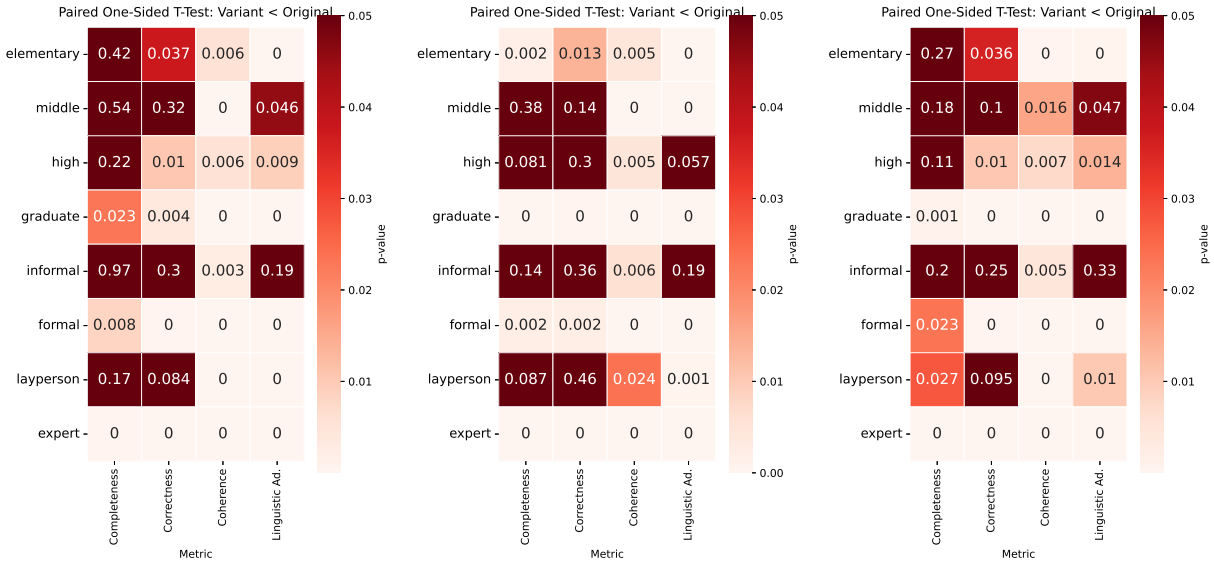
```
You are a helpful assistant that evaluates medical QA samples. For each sample, you must evaluate the generated answer (A_gen) as a response to the modified question (Q_mod). Use the gold
standard answer (A_gold) as the medically accurate information regarding the topic. In this QA task, the generated 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 1 (lowest) to 3 (highest).
  1. Correctness: What portion of A_gen is accurate in answering the modified question Q_mod? Consider A_gold to be medically accurate.
  - Score Definitions:
      CR3: Considering A_gold as medically accurate, the generated answer is mostly correct in answering Q_mod.
      CR2: The generated answer is partially correct, including some medically sound details but missing or misrepresenting key elements needed to fully address Q_mod.
      CR1: Comparing with A_gold, the generated answer is completely incorrect. It fails to capture the relevant medical information needed to address Q_mod.
  2. Completeness: What portion of the queries made in Q_mod are answered by A_gen? Do not penalize for incorrect answers.
  - Score Definitions:
      CM3: A_gen answers the important queries asked in Q_mod.
      CM2: The generated answer covers some essential details but omits other essential elements needed for a complete answer.
      CM1: The generated answer is largely incomplete, missing most of the essential details required to fully address Q_mod.
  3. Fluency and Coherence: How well is A_gen written in terms of language fluency and logical structure?
  - Score Definitions:
      FC3: The generated answer is highly fluent and logically structured.
      FC2: The generated answer is moderately fluent but contains noticeable issues.
      FC1: The generated answer is poorly written, with significant grammatical and structural issues.
  4. Linguistic Adaptability: How effectively does A_gen adopt the style and tone of the modified question (Q_mod)?
  - Score Definitions:
      LA3: The generated answer fully captures the tone and style of Q_mod.
      LA2: The generated answer partially matches the tone or style.
      LA1: The generated answer fails to match the tone or style of Q_mod.

  Return your evaluation in JSON format as follows:
  {
      "correctness": ,
      "completeness": ,
      "fluency_and_coherence": ,
      "linguistic_adaptability":
  }
  Ensure that your output contains only the JSON object.
```

Figure 6: System Prompt for LLM-Judge and instructions for annotators validating the LLM-Judge

B Additional results

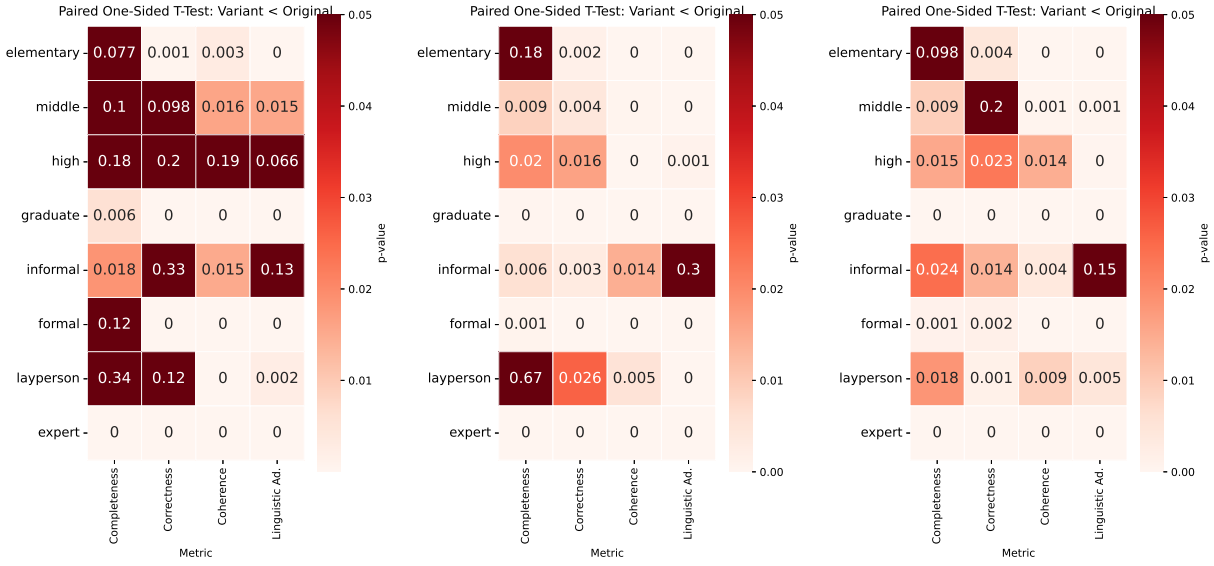
B.1 Significance Test

Section 4 mentions that a significance test was performed. Figures 7, 8, and 9 represent heatmaps of the detailed results from the significance test.



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

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



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

Figure 8: 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

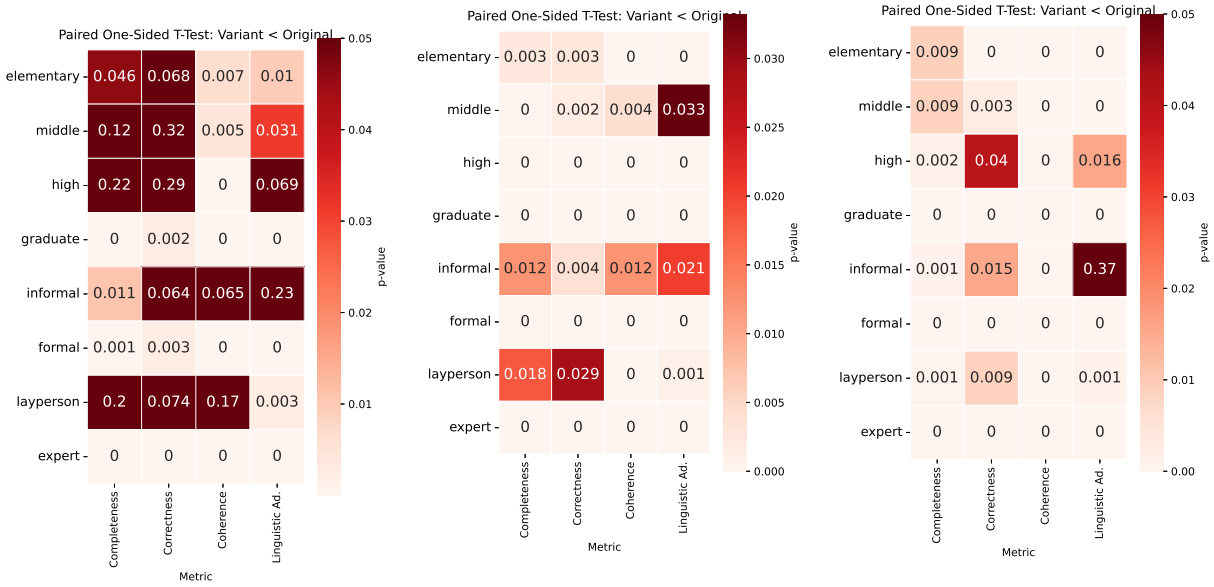
Table 5: Full results table. † indicates models with 8-bit quantization.

B.2 Full Result

Table 5 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.

C 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.



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

Figure 9: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)