Out of Style: RAG's Fragility to Linguistic Variation

Tianyu Cao[1*] **Neel Bhandari**[1*] **Akhila Yerukola**[1] **Akari Asai**[12] **Maarten Sap**[1]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Allen Institute for AI
{tianyuca, neelbhan, ayerukol, aasai, msap2}@cs.cmu.edu

Abstract

Despite the impressive performance of Retrieval-augmented Generation (RAG) systems across various NLP benchmarks, their robustness in handling real-world user-LLM interaction queries remains largely underexplored. This presents a critical gap for practical deployment, where user queries exhibit greater linguistic variations and can trigger cascading errors across interdependent RAG components. In this work, we systematically analyze how varying four linguistic dimensions (*formality, readability, politeness, and grammatical correctness*) impact RAG performance. We evaluate two retrieval models and nine LLMs, ranging from 3 to 72 billion parameters, across four information-seeking Question Answering (QA) datasets. Our results reveal that linguistic reformulations significantly impact both retrieval and generation stages, leading to a relative performance drop of up to 40.41% in Recall@5 scores for less formal queries and 38.86% in answer match scores for queries containing grammatical errors. Notably, RAG systems exhibit greater sensitivity to such variations compared to LLM-only generations, highlighting their vulnerability to error propagation due to linguistic shifts. These findings highlight the need for improved robustness techniques to enhance reliability in diverse user interactions. [1]

1 Introduction

Retrieval-augmented Generation (RAG) systems enhance Large Language Models (LLMs) by integrating external knowledge retrieval, grounding their output in factual context to improve accuracy and reliability (Lewis et al., 2021; Gao et al., 2024). However, their widespread integration into real-world applications (K2view, 2024) introduces potential challenges regarding robustness to linguistic variations. Users bring varied backgrounds, domains, and cultural contexts that naturally produce linguistic differences in their queries (Park et al., 2024; Li et al., 2020; Lorenzo-Dus and Bou-Franch, 2013). As Figure 1 illustrates, different from the carefully curated queries from traditional NLP benchmarks, real-world user-LLM queries tend to be less formal and frequently contain grammatical inconsistencies (Ouyang et al., 2023).

Failing to account for these linguistic variations risks excluding a broad segment of users from effective interaction, especially for users whose linguistic expressions fall outside the narrow patterns these systems are tuned on (Liang et al., 2023). Moreover, unlike standalone LLMs, RAG systems incorporate multiple interdependent components, making them susceptible to cascading errors arising at both the retrieval and generation stages (Asai et al., 2023; Yoran et al., 2024a; Kim et al., 2025). *A truly robust RAG system should maintain consistent retrieval effectiveness and generation quality across the full spectrum of user linguistic variations.*

We present a large-scale systematic investigation of how variations in linguistic characteristics affect the robustness of RAG systems. We target diverse and prevalent variations commonly found in real-world user inputs that meaningfully challenge RAG systems (Park et al., 2024; Ouyang et al., 2023), namely, **formality**, **readability**, **politeness**, and **grammatical correctness**. These choices ensure our analysis covers stylistic, pragmatic, and structural aspects aligned with practical usage. By automatically rewriting queries across these dimensions, we analyze how linguistic variations impact each RAG system component, as well as potential cascading errors throughout the pipeline. Our evaluation encompasses two retrieval models, namely Contriever (Izacard et al., 2021) and ModernBERT (Nussbaum et al., 2024), and nine LLMs

[*] Equal contribution.
[1] Code is available at https://github.com/Springcty/RAG-fragility-to-linguistic-variation.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 280–318

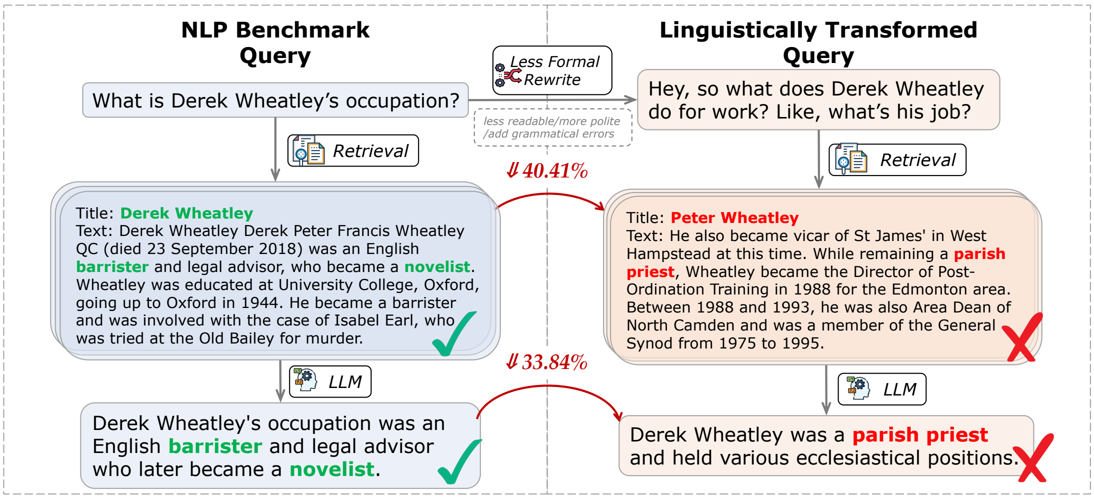

Figure 1: **RAG systems demonstrate overall performance degradation when queries are rewritten to be less formal, more polite, less readable, and with grammatical errors.** For the traditional NLP query (left), the RAG systems successfully retrieve related information and generate the correct answer, while the less formal queries (right) retrieve incorrect information. The linguistic variation on formality causes significant performance drops: 40.41% decrease in Recall@5 and 33.38% decrease in answer match (AM) score on the MS MARCO dataset.

from three families (Llama 3.1, Qwen 2.5, Gemma 2) of varying scales across four open-domain Question Answering (QA) datasets: PopQA (Mallen et al., 2023), EntityQuestions (Sciavolino et al., 2022), MS MARCO (Bajaj et al., 2018), and Natural Questions (Kwiatkowski et al., 2019).

Our experiments reveal significant vulnerabilities in RAG systems to linguistic variations. Retrieval analysis shows an average 15% relative performance degradation across all datasets and linguistic dimensions, with grammatical modifications most severely impacting recall while politeness variations show minimal effect. Generation analysis demonstrates average decreases of 16.52% (AM), 41.15% (EM), and 19.60% (F1) across all experimental settings. Notably, increasing the LLM scales doesn't always help mitigate these performance gaps.

Furthermore, RAG systems exhibit greater vulnerability to linguistic variations than LLM-only generations, suggesting cascading errors across components. The PopQA dataset shows a 22.53% average performance drop in RAG systems versus only 10.78% in LLM-only generations. These findings highlight the urgent need to improve the robustness of retrieval components when handling linguistically varied queries. We also find that while advanced RAG techniques, e.g., query expansion with HyDE (Gao et al., 2022) and documents reranking, tend to improve overall performance, they remain similarly vulnerable to linguistic variations.

In summary, we present the first systematic analysis of the robustness of RAG systems to linguistic query reformulation. Our results demonstrate that despite strong performance on standard benchmarks, they remain fragile to inevitable real-world linguistic variations. These findings highlight the need for enhanced robustness techniques to improve reliability across diverse user interactions and inform design principles for next-generation RAG systems.

2 Related Work

Robustness of retrieval systems to linguistic variations. Prior research investigated retrieval system robustness to noisy queries (Campos et al., 2023; Chen et al., 2022, 2023b) and specific linguistic variations including word substitutions (Wu et al., 2022), aspect changes and paraphrasing (Penha et al., 2022), typographical errors (Zhuang and Zuccon, 2021), and grammatical variations (Long et al., 2024). Our work provides the first holistic evaluation of RAG systems' robustness to diverse linguistic variations and uncovers cascading failures across the entire RAG pipeline.

Robustness of language models to linguistic variations. Prior research has examined impacts of syntactic perturbations (Moradi and Samwald, 2021; Singh et al., 2024), round-trip translation (Bhandari and Chen, 2023), politeness variations (Yin et al., 2024), equivalent queries (Cao et al., 2024) and scale of model size on robustness to grammatical errors (Yin et al., 2020; Hagen et al., 2024). Rawte et al. (2023) examined how formality and readability affect LLM performance in isolation. In contrast, our research investigates a broader spectrum of linguistic variations, generates data for each variation using LLM-based rewrites, and studies their compounding effects throughout the end-to-end RAG pipeline comprehensively.

Robustness of RAG systems. While RAG systems demonstrate impressive performance (Lewis et al., 2021; Gao et al., 2024) and reduced hallucinations (Mallen et al., 2023), vulnerabilities exist with increasing retrieval context noise (Chen et al., 2023a), irrelevant contexts (Yoran et al., 2024b), and noise impacts (Fang et al., 2024). Yang et al. (2025) analyzed how spurious features affect RAG perform. Cho et al. (2024) introduced document-level perturbations and evaluated RAG's vulnerability to noisy documents. RbFT (Tu et al., 2025) introduce specialized training objectives to improve consistency under corrupted or adversarial retrieval contexts. These studies focus primarily on retrieved content noise rather than initial *query*; our work uniquely demonstrates how diverse linguistic variations in user queries compound throughout the RAG pipeline, exposing critical vulnerabilities in systems serving diverse users.

3 Robustness Evaluation Approach

In our work, we explore the impact of the following linguistic aspects: **Formality**, **Readability**, **Politeness** and **Grammatical Correctness - Round-Trip Translation and Typos**. We explore these linguistic queries as they are essential dimensions of language variation that are prevalent and significant in real-world RAG interactions. We extend Rawte et al. (2023)'s findings on linguistic variations and LLM hallucinations by synthetically generating queries across four linguistic dimensions and four datasets to analyze each RAG pipeline component comprehensively. We first formulate our task, followed by defining each of our linguistic characteristics (Section 3.1), and elaborating on our query rewriting design (Section 3.2).

Task formulation. Given a seed dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots\}$, where x_i and y_i indicate i-th input and output, respectively, we reformulate each query $x_i \to x'_i$ based on four linguistic aspects. A robust RAG system, composed of a retriever $\mathcal{R}$ and a generator $\mathcal{G}$ operating on corpus $\mathcal{C}$, should maintain performance when processing linguistically varied inputs that are semantically equivalent, i.e., where the gold output remains unchanged semantically. For the retrieval component, we expect retrieved documents $\mathbf{D}_i = \mathcal{R}(x_i, \mathcal{C})$ and $\mathbf{D}'_i = \mathcal{R}(x'_i, \mathcal{C})$ to both contain the answer y_i. For generation, given retrieved documents $\mathbf{D}_i, \mathbf{D}'_i$, a robust system should produce: $\mathcal{G}(x_i, \mathbf{D}_i) \approx \mathcal{G}(x'_i, \mathbf{D}'_i) \approx y_i$.

3.1 Linguistic Variations

Formality. Formality in language lacks universal definition (Pavlick and Tetreault, 2016; Mosquera and Moreda, 2011; Fang and Cao, 2009), but encompasses situational factors (Hovy, 1987; Lahiri et al., 2011), grammar quality (Peterson et al., 2011), and specific linguistic elements like contractions (Heylighen and Dewaele, 1999). We quantify formality using the RoBERTa-based formality classifier from (Babakov et al., 2023).

Readability. Readability quantifies text comprehensibility through linguistic complexity. We employ the widely-used Flesch Reading Ease Score (FRES; Flesch 1948) to assess readability (Rawte et al., 2023; Han et al., 2024), defined in A.1.

Politeness. Politeness is a sociocultural phenomenon defined as showing consideration of others (Wang, 2014). We calculate politeness scores using Polite Guard (Intel, 2024), an open-source NLP model from Intel that's fine-tuned from BERT to classify text into four levels: *polite*, *somewhat polite*, *neutral*, and *impolite*.

Grammatical correctness. In our work, we define grammatical correctness as the preservation of both grammaticality (Chomsky, 2002) and semantic fidelity. We alter the grammatical correctness through two approaches, inspired by Yin et al. (2020); Zhuang and Zuccon (2021); Lichtarge et al. (2019): (1) **Typos**, where random addition, deletion, or substitution operations are applied at a 20% probability per word, requiring an edit distance of at least 1; and (2) **Round-trip translation (RTT)** via English-Afrikaans-English using EasyNMT with opus-mt model, requiring the output to not be the exact same as the original query.

3.2 Query Rewriting

We systematically reformulate queries $x_i \to x'_i$ across targeted linguistic dimensions while ensuring all rewrites satisfy dimension-specific thresholds and preserve semantic consistency (Appendix A.1).

Linguistic Dimension (Dataset)	Example Rewrites
Politeness ↑ (MS MARCO)	**Original:** complex carbohydrates are stored in animals in the form of **Rewritten:** Would you be so kind as to share how complex carbohydrates are stored in animals?
Readability ↓ (Natural Questions)	**Original:** who stars in the new movie the post **Rewritten:** In the upcoming cinematic production titled "The Post," which individuals have been cast in leading roles?
Formality ↓ (PopQA)	**Original:** Who is the author of Dolores Claiborne? **Rewritten:** Hey, do you happen to know who wrote Dolores Claiborne? I'm kinda curious!
Grammar: RTT ↓ (EntityQuestions)	**Original:** Which company is HMS Blankney produced by? **Rewritten:** What company is producing HMS Blankey?
Grammar: Typos ↓ (Natural Questions)	**Original:** when did the japanese river otter **become** extinct **Rewritten:** when did the japanese river otter **ecome** extinct

Table 1: Examples of query rewrites across different linguistic dimensions and datasets. Each pair shows the original query and its rewritten form.

Linguistic reformulation. For each distinct dimension (formality, readability, and politeness), we sample $5,000$ original queries and use GPT-4o-mini (OpenAI et al., 2024) to generate rewrites using three different prompts (detailed in Appendix I).[2] We design distinct query sets across all datasets to be less formal, less readable, or more polite—directions chosen because the well-curated original datasets contained predominantly formal, readable, and neutral language, allowing our modifications to explore more realistic linguistic variations. This creates $15,000$ rewritten samples per dimension for each dataset, over which we average results to account for prompt stochasticity. For grammatical correctness, we similarly use $5,000$ original queries but apply deterministic transformations through typo introduction and round-trip translation as described in Section 3.1, creating $5,000$ modified samples for each approach. Example rewrites are shown in Table 1.

Semantic consistency. We define a rewritten query x'_i as semantically consistent with the original query x_i if it preserves the original pragmatic intent, i.e., a human annotator would assign the same gold answer y_i to both queries. This operational definition reflects the requirements that linguistic variation should not alter the underlying information need. To strictly enforce this, we apply a two-stage verification procedure. First, we automatically filter rewrites using a sentence-level semantic similarity threshold (> 0.7) computed with MPNet-v2 (Song et al., 2020). Second, we conduct qualitative evaluation through manual annotation on 250 queries across all linguistic variations, finding that 94.67% of the rewritten queries preserve the original answer. Further details of the annotation are provided in Appendix A.2.

For each linguistic dimension and dataset, we construct controlled comparative datasets $\mathcal{D}' = \{(x'_1, y_1), (x'_2, y_2), \ldots\}$, enabling direct measurement of performance changes attributable solely to linguistic variation.

4 Experimental Setups

4.1 Benchmarks

In this work, we use four open-domain QA datasets as seed datasets $\mathcal{D}$ and evaluate the effects how linguistic variations of the original queries affect RAG systems. PopQA (Mallen et al., 2023) is a large-scale entity-centric open-domain QA dataset about entities with a wide variety of popularity. EntityQuestions (Sciavolino et al., 2022) is a set of simple, entity-rich questions based on facts of Wikidata. MS MARCO (Bajaj et al., 2018) contains questions derived from real user search queries from Bing's search logs. Natural Questions (Kwiatkowski et al., 2019) contains questions consisting of real, anonymized, aggregated queries to the Google search engine. Both PopQA and EntityQuestions consist of well-structured, standardized, and simple queries, while queries from MS MARCO and Natural Questions exhibit free-form and arbitrary. We evaluate using PopQA's test split, EntityQuestions' dev split, and both dev and test splits from MS MARCO and Natural Questions. The retrieval is performed on the Wikipedia

[2]To verify that our experimental results are robust to rewriting models, we generate 500 rewritten queries from the PopQA dataset using Llama-3.1-70B-Instruct. Shown in Appendix C, results confirm our findings with GPT-4o-mini.

passage set used in DPR[3] for the PopQA, EntityQuestions, and Natural Questions datasets, while the MS MARCO dataset uses its corresponding passage dataset (Bajaj et al., 2018).

4.2 Models

Retrieval. We use two neural retrieval systems, namely **Contriever** (`facebook/contriever`; Izacard et al. 2022) and **ModernBERT Embed** (`nomic-ai/modernbert-embed-base`; Warner et al. 2024; Nussbaum et al. 2024). Contriever is an unsupervised dense retriever built from BERT base architecture and is pre-trained using a contrastive learning framework. ModernBERT Embed is an embedding model trained from ModernBERT-base (Warner et al., 2024), bringing the new advances of ModernBERT to embeddings. We use the retrieval system implementation by Izacard et al. (2022) for both retrieval models.[4]

Generation. We evaluate the nine most advanced open-source instruction-tuned LLMs from three model families with various scales: **Llama 3.1** (Grattafiori et al. 2024; 8 and 70 billion), **Qwen 2.5** (Yang et al. 2024; Team 2024; 3, 7, 32, and 72 billion), and **Gemma 2** (Team et al. 2024; 2, 9, and 27 billion). Detailed hyperparameter settings are included in the Appendix B.1.

We use few-shot prompting to ensure that the model outputs are in the correct format. For each dataset and linguistic characteristic, we include two question-answer pairs in the context: one random original query with its answer, and its corresponding linguistically rewritten version with the same answer. This balanced approach exposes the model to both original and rewritten query formats to ensure fairness. We also include the top five retrieved passages in the context. Detailed prompts are provided in Appendix J.

4.3 Metrics

Retrieval. We employ **Recall@k (R@k)** (Karpukhin et al., 2020), which calculates the fraction of the retrieved documents containing gold answers. We use $k = 5$ as our primary setup, and evaluate the effect of varying k in our analysis.

Generation. The generation stage is assessed using a comprehensive set of metrics: **Answer Match (AM)** measures the percentage of the predictions that any substring of the prediction is an exact match of any of the ground truth answers, **Exact Match (EM)** measures the percentage of predictions that exactly match any of the ground truth answers, and **F1 Score (F1)** captures the harmonic mean of precision and recall in generated responses. We mainly report AM scores in the main paper because it better reflects answer correctness under linguistic and stylistic variaion. The full results could be found in Appendix F. As a complement to standard metrics, we also use **LLM-as-a-Judge** with GPT-5-mini (OpenAI, 2025).

5 RAG Robustness Experimental Results

In this section, we progress from component-level to end-to-end RAG system analysis, followed by an assessment of advanced techniques (query expansion and re-ranking) and their ability to mitigate performance drops when faced with linguistic variations.

5.1 Retrieval Analysis

We conduct a comprehensive analysis of retrieval systems performance across two candidate retrievers: Contriever and ModernBERT Embed. The results are shown in Table 2.

Linguistics	Retriever	PopQA	Entity	MARCO	NQ	Δ Q-len
Readability	Contriever	18.45 (0.61)	8.61 (0.64)	**21.10** (0.25)	7.69 (0.60)	5.23
	ModernBERT	**17.73** (0.65)	13.17 (0.61)	14.58 (0.40)	10.08 (0.65)	
Gram. (RTT)	Contriever	**29.00** (0.59)	14.57 (0.68)	9.14 (0.34)	23.50 (0.62)	-0.26
	ModernBERT	**29.68** (0.62)	14.85 (0.66)	17.54 (0.32)	18.24 (0.67)	
Gram. (Typos)	Contriever	27.79 (0.59)	14.80 (0.68)	15.53 (0.34)	**30.83** (0.62)	0.02
	ModernBERT	**22.48** (0.62)	11.01 (0.66)	12.30 (0.32)	13.45 (0.67)	
Formality	Contriever	19.96 (0.70)	10.71 (0.68)	**40.41** (0.25)	15.35 (0.65)	13.65
	ModernBERT	13.67 (0.74)	8.05 (0.68)	**15.55** (0.40)	9.51 (0.69)	
Politeness	Contriever	8.30 (0.62)	1.70 (0.67)	**16.44** (0.26)	1.16 (0.60)	7.29
	ModernBERT	**10.70** (0.67)	3.39 (0.67)	5.18 (0.40)	4.97 (0.65)	

Table 2: Relative retrieval performance drop (%) in R@5 on rewritten queries across datasets. (Original scores) shown in gray parentheses. Bold indicates **the largest degradation value** per retriever. Δ **Q-len** represents average query token-length change. **Query linguistic variations degrades retrieval performance consistently across all linguistic characteristics.**

Query variations based on linguistic dimensions degrade retrieval performance. Our analysis (Table 2) reveals significant linguistic fragility in retrieval systems, with performance degradation averaging 16.7% for Contriever and 13.3% for ModernBERT across all modifications. We hypothesize that ModernBERT's increased robustness

[3] https://dl.fbaipublicfiles.com/dpr/wikipedia_split/psgs_w100.tsv.gz
[4] https://github.com/facebookresearch/contriever

could largely be attributed to its diverse pre-training data, which forces the model to learn to process non-standard sequences and enhances its resilience to linguistic noise. Additionally, the masked language modeling objective utilized by ModernBERT could effectively train the model to denoise and reconstruct context from corrupted inputs. Results show highest sensitivity on PopQA (19.78% average impact), particularly to grammatical transformations (29.34% from RTT). MS MARCO exhibits the second-highest impact (16.78%), with striking sensitivity to formality changes (40.41% with Contriever), suggesting retrieval systems may be implicitly optimized for specific linguistic patterns, limiting effectiveness when handling diverse query variations.

Grammatical variations have the highest impact on retrieval performance. On average, grammaticality rewrites emerge as the most impactful linguistic variation on recall performance. Round-trip translation degrades recall by an average of 19.56% across all datasets. Interestingly, ModernBERT shows greater vulnerability to these structural transformations (20.12% drop) compared to Contriever (19% drop). Typographical errors present another significant challenge, causing an average recall reduction of 18.51%. However, the retrievers display opposite behavior patterns with typos: Contriever exhibits substantially lower robustness (22.22% drop) than ModernBERT (14.81% drop). This suggests that ModernBERT's diverse training data mixture likely enables it to develop greater robustness to character-level grammatical perturbations compared to Contriever.

Politeness variations have minimal impact on retrieval performance. Politeness variations have the least impact on retrieval performance, with an average recall drop of only 6.48% across all datasets and retrievers. This stands in stark contrast to grammatical variations (19.56%) and typos (18.51%). The minimal effect is most evident in Natural Questions with Contriever (1.16%) and EntityQuestions with Contriever (1.70%). This suggests that retrieval models effectively filter out social courtesy markers while preserving their focus on the query's core semantic content and keywords, maintaining robust performance despite changes in query politeness level.

Retrieval performance drops independent of query length. Our analysis, as shown in Table 2 demonstrates that query length changes do not directly correlate with retrieval performance. Queries with increased formality showed substantial length increases (+13.65 tokens) yet produced inconsistent performance impacts across datasets. Conversely, round-trip translated queries were marginally shorter (-0.26 tokens) but consistently caused significant performance degradation. This indicates retrieval models respond more to linguistic quality (grammatical correctness, readability) than to query length itself, highlighting the need for systems robust to linguistic variations rather than optimized for specific query lengths. We confirm this by analyzing semantic preservation using LLM-as-a-judge in Table 4 and a human evaluation of results in Table 5.

Scaling up number of documents improves performance. Table 2 shows performance degradation (ΔR@K) decreasing as K increases, indicating linguistic perturbations cause relevant documents to slide down rather than disappear from the ranked list. As an example, for rewrites with typos, Δ R@K for Contriever decreases from 22.2 at R@5 to 12.0 at R@100, and for ModernBERT from 20.1 to 9.8. This is detailed further in Appendix E.1. This ranking deterioration forces downstream language models to operate with suboptimal information, potentially compromising response quality. We further investigate this hypothesis in Section 5.4.2 by examining if rerankers can improve recall scores for Top-5 retrieved documents.

5.2 Generation Analysis

The RAG experiment results on ModernBERT retriever and nine LLMs are presented in Table 3 with answer match (AM) scores. Table 4 shows the results with LLM-as-a-Judge evaluation. Overall, RAG systems show performance degradation on all linguistic variations.

RAG systems are sensitive to linguistic variations. As illustrated in Table 3, across all datasets, we observe a noticeable overall degradation in performance when queries are rewritten to become less formal, more polite, less readable, or have grammatical errors. Across all datasets, linguistic dimensions, and experimental settings, we found average drops of 16.52% (AM), 41.15% (EM), and 19.60% (F1). The PopQA dataset shows the highest sensitivity to all linguistic variations, with an average performance drop of 18.64% on AM scores. Particularly notable were the effects of reduced read-

Model \ Rew. %↓ (Ori.)	Readability				Grammatical Correctness							
	PopQA	NQ	MARCO	Entity	PopQA		NQ		MARCO		Entity	
					RTT	Typos	RTT	Typos	RTT	Typos	RTT	Typos
gemma-2-2b-it	**20.71** (0.52)	16.82 (0.45)	17.44 (0.20)	16.43 (0.51)	**32.73**	**20.26** (0.49)	28.09	10.44 (0.43)	29.81	10.19 (0.16)	22.47	10.36 (0.54)
gemma-2-9b-it	**17.12** (0.54)	13.95 (0.49)	11.13 (0.20)	15.25 (0.54)	**33.90**	**19.60** (0.51)	24.50	8.19 (0.48)	25.69	6.98 (0.16)	22.92	10.39 (0.56)
gemma-2-27b-it	**16.15** (0.56)	11.33 (0.52)	8.33 (0.20)	12.68 (0.55)	**32.81**	**18.21** (0.53)	24.03	6.81 (0.51)	27.71	3.79 (0.15)	20.73	10.69 (0.58)
Llama-3.1-8B-Instruct	**17.27** (0.55)	15.07 (0.52)	15.23 (0.22)	13.30 (0.52)	**35.27**	**21.95** (0.50)	27.45	10.54 (0.50)	32.37	12.37 (0.19)	25.06	12.53 (0.56)
Llama-3.1-70B-Instruct	**17.66** (0.58)	15.77 (0.54)	13.30 (0.22)	16.09 (0.55)	**35.30**	**19.49** (0.54)	25.43	8.80 (0.52)	24.83	5.38 (0.17)	21.61	12.71 (0.57)
Qwen2.5-3B-Instruct	22.35 (0.47)	18.65 (0.42)	13.27 (0.21)	**24.14** (0.46)	**32.89**	**23.77** (0.44)	30.74	15.65 (0.44)	32.14	12.31 (0.17)	24.53	13.20 (0.49)
Qwen2.5-7B-Instruct	19.50 (0.53)	17.15 (0.48)	11.55 (0.21)	**20.35** (0.51)	**34.06**	**22.67** (0.49)	29.67	13.19 (0.49)	24.28	6.64 (0.17)	21.77	11.99 (0.55)
Qwen2.5-32B-Instruct	17.00 (0.56)	14.01 (0.52)	**32.10** (0.28)	18.41 (0.53)	**34.11**	**19.69** (0.52)	24.80	9.28 (0.52)	26.67	7.31 (0.16)	20.96	11.62 (0.57)
Qwen2.5-72B-Instruct	16.25 (0.56)	13.62 (0.54)	8.36 (0.20)	**17.03** (0.54)	**33.68**	**19.27** (0.53)	24.28	7.74 (0.53)	27.47	8.22 (0.16)	19.56	10.35 (0.58)
Avg	**18.22** (0.54)	15.15 (0.50)	14.52 (0.22)	17.07 (0.52)	**33.86**	**20.55** (0.51)	26.55	10.07 (0.49)	27.88	8.13 (0.17)	22.18	11.54 (0.56)

Model \ Rew. %↓ (Ori.)	Formality				Politeness			
	PopQA	NQ	MARCO	Entity	PopQA	NQ	MARCO	Entity
gemma-2-2b-it	13.64 (0.61)	12.44 (0.46)	**23.31** (0.17)	7.66 (0.60)	8.37 (0.52)	4.42 (0.44)	**11.16** (0.20)	2.96 (0.59)
gemma-2-9b-it	11.67 (0.64)	9.95 (0.50)	**19.68** (0.17)	8.85 (0.62)	**8.54** (0.54)	3.60 (0.48)	8.19 (0.20)	3.01 (0.61)
gemma-2-27b-it	12.21 (0.65)	9.92 (0.52)	**19.13** (0.17)	7.19 (0.63)	**7.93** (0.56)	3.38 (0.51)	6.13 (0.20)	1.97 (0.61)
Llama-3.1-8B-Instruct	12.23 (0.65)	10.51 (0.52)	**17.59** (0.20)	7.16 (0.62)	7.59 (0.55)	3.46 (0.51)	**9.90** (0.23)	3.50 (0.59)
Llama-3.1-70B-Instruct	11.48 (0.67)	12.09 (0.54)	**19.00** (0.20)	7.19 (0.64)	7.00 (0.57)	6.29 (0.53)	**9.89** (0.23)	3.61 (0.61)
Qwen2.5-3B-Instruct	10.97 (0.58)	11.38 (0.45)	**20.87** (0.18)	7.51 (0.56)	11.54 (0.47)	5.91 (0.42)	**12.30** (0.21)	7.74 (0.55)
Qwen2.5-7B-Instruct	12.99 (0.63)	11.46 (0.50)	**19.84** (0.18)	8.08 (0.61)	**10.76** (0.54)	5.05 (0.48)	7.08 (0.21)	4.69 (0.60)
Qwen2.5-32B-Instruct	11.34 (0.65)	7.85 (0.53)	**14.80** (0.18)	6.24 (0.62)	**8.65** (0.56)	4.76 (0.52)	6.66 (0.21)	5.64 (0.60)
Qwen2.5-72B-Instruct	10.68 (0.66)	9.11 (0.55)	**20.72** (0.18)	5.75 (0.63)	7.36 (0.56)	5.45 (0.54)	**16.62** (0.20)	3.86 (0.62)
Avg	11.91 (0.64)	10.52 (0.51)	**19.44** (0.18)	7.29 (0.61)	8.64 (0.54)	4.70 (0.49)	**9.77** (0.21)	4.11 (0.60)

Table 3: RAG performance on answer match (AM) scores using ModernBERT retriever with the Gemma 2, Llama 3.1, and Qwen 2.5 model families across four datasets. Results show relative percentage performance degradation on rewritten queries (Rew. % ↓) and the original query performance (Ori.) within parentheses in gray. For RTT, it has the same original scores as Typos. **The largest degradation value** among four datasets is in bold. **All systems exhibit performance drops across all linguistic variations and datasets.**

Dataset	Readability	Formality	Politeness	Gram-RTT	Gram-Typos
PopQA	19.41%	11.79%	6.75%	35.60%	21.20%
NQ	13.12%	4.27%	4.38%	22.74%	5.63%
MARCO	10.56%	4.05%	1.75%	15.21%	5.32%
Entity	13.20%	5.06%	2.86%	22.18%	10.49%

Table 4: RAG performance in LLM-as-a-Judge evaluation (ModernBERT + Qwen2.5-72B-Instruct), with relative percentage performance degradation. **LLM-as-A-Judge evaluation shows a consistent degradation pattern.**

ability (18.22% degradation) and round-trip translation (33.86% degradation). These findings suggest that while the RAG systems perform well on standard NLP benchmarks with structured queries, they remain vulnerable to common linguistic variations. The full experiment results are shown in Appendix F.

Politeness reformulations yield different impacts on AM and EM scores. When queries are rephrased to be more polite, we find that the AM scores remain relatively similar to those of the original, with less than 10% change for all datasets. However, there are significant drops in exact match (EM) scores. Specifically, we observe 44.57% and 18.32% drops in EM scores for queries from the Natural Questions and PopQA datasets, respectively. Although the rewritten queries preserve the original pragmatic intent, they introduce longer phrasing and auxiliary words that shift embeddings and sometimes nudge generators toward more verbose outputs. As a result, they still cause slight regressions in the RAG.

The round-trip translation errors and typos highlight different sensitivities. Round-trip translation, which introduces structural sentence transformations, generally causes notable decreases across all datasets, showing 33.86%, 26.55%, 27.88%, and 22.18% drops in AM scores in the PopQA, Natural Questions, MS MARCO, and EntityQuestions, respectively. In contrast, typos, mainly introducing surface-level grammatical errors, produce moderate but less drastic performance degradation. This finding is consistent with the retrieval experiment results, suggesting that RAG systems are more vulnerable to structural transformations than superficial grammatical mistakes.

5.3 Retrieval Method and LLM Scale Influence

In this section, we are going to investigate the influence of different retrieval methods and LLM scales on the robustness of the RAG systems. The main results are shown in Figure 2.

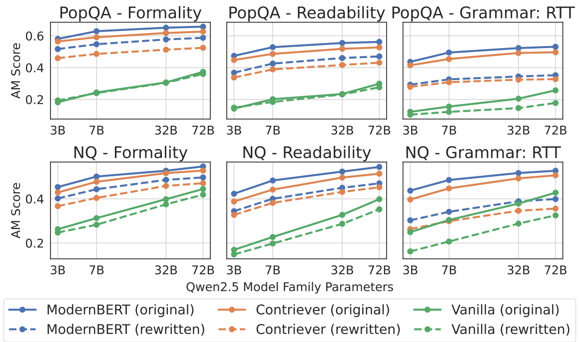

Figure 2: PopQA and Natural Questions (NQ) LLMs scaling results, augmented with ModernBERT, Contriever, and LLM-only generation (Vanilla). **Retrieval-augmented generation is more sensitive to linguistic variations than the LLM-only generation.**

RAG systems with ModernBERT retrieval show greater robustness to linguistic variations. As shown in Figure 2, generation results based on ModernBERT retrieval consistently outperform those with Contriever retrieval across both original and rewritten queries. Notably, RAG systems with ModernBERT demonstrate superior robustness to linguistic variations, exhibiting an average performance drop of only 19.52% on rewritten queries compared to 24.38% for Contriever. This suggests ModernBERT retrieval maintains better semantic understanding when handling linguistically varied inputs.

RAG systems show higher sensitivity to linguistic variations than LLM-only generations. For PopQA, we observe an average performance drop of 22.53% across all linguistic variations and both retrieval models, while LLM-only generations experience only a 10.78% reduction. Even more striking, Figure 2 shows that there is barely any performance difference in LLM-only generation on formality rewrites, which suggests that errors are cascaded from the retrieval component to the generation component in the RAG system. These findings further indicate that retrieval components represent the primary vulnerability in RAG systems when handling linguistic variations.

LLM scaling doesn't always help with mitigating performance gaps in RAG systems. Notably, the performance gap between original and rewritten queries narrows for formality and readability variations as LLMs scale up (see Figure 2). Specifically, PopQA shows reduced degradation on less readable queries from 22.35% at 3B to 16.25% at 72B parameter. This suggests that larger models can extract relevant information better from retrieved contexts. However, this scaling benefit remains selective and limited; for round-trip translation variations, the performance gap actually widens with increased model size. This counterintuitive finding may be attributed to the structural transformations introduced during translation that become more problematic for larger models attempting more precise reasoning.

Human annotator evaluation aligns with our automated evaluation metrics To validate our automated metrics and control for potential confounds such as length or verbosity, two independent annotators evaluated 250 samples across EntityQuestions (100 samples) and NaturalQuestions (150 samples). For typos and RTT variations, rewrites with more than 2 character edits to any entity were marked as not preserving semantic meaning. As shown in Table 5, semantic preservation rates ranged from 86.7% to 100% with high inter-annotator agreement (96.7%-100%), confirming that observed performance differences stem from linguistic variation rather than semantic drift. Full annotation protocols are provided in Section A.2.

5.4 Exploring the Robustness of Advanced RAG Systems

Many modern RAG systems include more components than simply retrieval and generation, which aim to make them more useful for users (Gao et al., 2022). In this section, we explore the possibility of using a simple **query-expansion** step to fix the vulnerability on linguistic variations and whether the addition of **reranking** improves RAG robustness. Detailed results of the experiment can be found in Appendices G and H.

5.4.1 Query Expansion

We evaluate Hypothetical Document Embeddings (HyDE; Gao et al. 2022) for query expansion. Figure 3 reveals that HyDE improves ModernBERT's

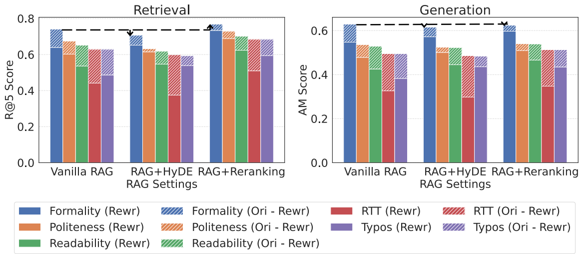

Figure 3: Retrieval (ModernBERT, R@5 Score) and generation (Qwen2.5-7B-Instruct, AM Score) performance across different RAG settings on PopQA. **We find that (1) adding HyDE and Rerank to the RAG pipeline improves the robustness to linguistic variations, but still lags behind original queries in performance. (2) HyDE improves robustness but slightly reduces performance on original queries. (3) Reranking improves performance on both original and rewritten queries.**

retrieval on linguistically varied queries (readability, typos, formality, politeness) by 2.61% on average, but severely impairs performance on round-trip translated queries (11% decrease). For original queries, HyDE consistently reduces ModernBERT's retrieval effectiveness by 5.43%. Similarly, generation quality increases for rewritten queries across PopQA (3.77%) but decreases for original queries (1.92%). These findings suggest HyDE provides insufficient benefits for ModernBERT, underscoring the need for more effective query expansion methods.

5.4.2 Reranker

The retriever must be efficient for large document collections containing millions of entries, although it may sometimes retrieve irrelevant candidates. To address this, we incorporate a Cross-Encoder-based re-ranker to significantly enhance the quality of final answers. Specifically, we employ the MS MARCO Cross-Encoders developed by Reimers and Gurevych (2019) to re-rank passages retrieved by ModernBERT and Contriever. As illustrated in Figure 3, re-ranking substantially improves retrieval performance, particularly for rewritten queries, achieving an average improvement of 16.56% compared to only 7.40% for original queries. In contrast, generation results show a modest improvement of 1.83% for original queries and a more substantial improvement of 8.50% for rewritten queries. These findings suggest that the effectiveness of re-ranking is especially pronounced when handling rewritten queries and highlight the importance of improving the robustness of retrieval systems.

6 Conclusion

We conduct the first large-scale, systematic investigation into how linguistic variations—specifically formality, readability, politeness, and grammatical correctness—impact the robustness of RAG systems. Our analysis reveals that both the retrieval and generation components suffer performance degradation when faced with linguistic variations. Notably, RAG systems exhibit greater vulnerability to linguistic variations compared to LLM-only generations, indicating potential cascading errors within the retrieval-generation pipeline. Crucially, increasing the scale of LLMs does not consistently mitigate these robustness issues, and even advanced retrieval techniques like HyDE and reranking show similar susceptibility. These findings highlight the need to develop strategies that ensure reliable performance across linguistically varied queries, guiding future improvements of real-world RAG systems.

Limitations

While our choice of linguistic dimensions cover a broad spectrum of stylistic, pragmatic, and structural variations, other relevant factors such as dialect, idiomatic expressions, or domain-specific terminology could be explored in future work. Our grammatical variations (typos and round-trip translation) represent a subset of potential errors; other categories such as verb form or agreement errors (Dahlmeier et al., 2013) may present greater challenges and warrant future investigation. We conducted query rewriting using two LLMs (GPT-4o-mini and Llama-3.1-70B-Instruct) and observed similar vulnerabilities; future studies may verify the generalizability of these findings using a broader range of rewriting methods. Additionally, while we explored widely used methods such as query expansion and reranking to test for mitigation strategies, more comprehensive approaches, including training models explicitly on diverse, linguistically varied data, remain important avenues for future research.

Acknowledgments

This research was supported in part by the National Science Foundation under grant 2230466 and in part by DSO National Laboratories.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Neel Bhandari and Pin-Yu Chen. 2023. Lost in translation: Generating adversarial examples robust to round-trip translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE.

Daniel Campos, ChengXiang Zhai, and Alessandro Magnani. 2023. Noise-robust dense retrieval via contrastive alignment post training. *Preprint*, arXiv:2304.03401.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *Preprint*, arXiv:2406.10248.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. *Preprint*, arXiv:2309.01431.

Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023b. Dealing with textual noise for robust and effective bert re-ranking. *Information Processing & Management*, 60(1):103135.

Xuanang Chen, Jian Luo, Ben He, Le Sun 0001, and Yingfei Sun. 2022. Towards robust dense retrieval via local ranking alignment. In *IJCAI*, pages 1980–1986.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. 2024. Typos that broke the rag's back: Genetic attack on rag pipeline by simulating documents in the wild via low-level perturbations. *Preprint*, arXiv:2404.13948.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Alex Chengyu Fang and Jing Cao. 2009. Adjective density as a text formality characteristic for automatic text classification: A study based on the British National Corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 130–139, Hong Kong. City University of Hong Kong.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *Preprint*, arXiv:2405.20978.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *Preprint*, arXiv:2212.10496.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tim Hagen, Harrisen Scells, and Martin Potthast. 2024. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4283–4296, Miami, Florida, USA. Association for Computational Linguistics.

Yu Han, Aaron Ceross, and Jeroen H. M. Bergmann. 2024. The use of readability metrics in legal text: A systematic literature review. *Preprint*, arXiv:2411.09497.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1).

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Intel. 2024. Intel/polite-guard.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

K2view. 2024. 2024 genai adoption survey.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Taeyoun Kim, Jacob Springer, Aditi Raghunathan, and Maarten Sap. 2025. Mitigating bias in rag: Controlling the embedder. *arXiv*.

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.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020. Studying politeness across cultures using english twitter and mandarin weibo. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–15.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Preprint*, arXiv:2304.02819.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Whispers in grammars: Injecting covert backdoors to compromise dense retrieval systems. *Preprint*, arXiv:2402.13532.

Nuria Lorenzo-Dus and Patricia Bou-Franch. 2013. A cross-cultural investigation of email communication in peninsular spanish and british english: The role of (in) formality and (in) directness. *Pragmatics and Society*, 4(1):1–25.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alejandro Mosquera and Paloma Moreda. 2011. The use of metrics for measuring informality levels in web 2.0 texts. In *Brazilian Symposium in Information and Human Language Technology*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

OpenAI. 2025. GPT-5 System Card. `https://openai.com/index/gpt-5-system-card/`. Accessed: 2025-XX-XX.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. *Preprint*, arXiv:2310.12418.

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanushree Mitra, David Jurgens, and Yulia Tsvetkov. 2024. Valuescope: Unveiling implicit norms and values via return potential model of social interactions. *Preprint*, arXiv:2407.02472.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. *Preprint*, arXiv:2111.13057.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Vipula Rawte, Prachi Priya, S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Amit Sheth, and Amitava Das. 2023. Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *Preprint*, arXiv:2309.11064.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2022. Simple entity-centric questions challenge dense retrievers. *Preprint*, arXiv:2109.08535.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yiteng Tu, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. 2025. Robust fine-tuning for retrieval augmented generation against retrieval defects. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1272–1282, New York, NY, USA. Association for Computing Machinery.

Fang Wang. 2014. A model of translation of politeness based on relevance theory. *Open Journal of social sciences*, 2(9):270–277.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified robustness to word substitution ranking attack for neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2128–2137. ACM.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, Ming Gong, Hengyuan Zhang, and Dongmei Zhang. 2025. Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data. *Preprint*, arXiv:2503.05587.

Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. 2020. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online. Association for Computational Linguistics.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. *Preprint*, arXiv:2402.14531.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024a. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024b. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for bert-based passage retrieval and ranking. *Preprint*, arXiv:2108.12139.

A Linguistic Rewriting Settings

A.1 Linguistic Variations

In this section, we detail our design of linguistic variations as well as our qualitative analysis of the queries we generated.

Query rewrite design : For each linguistic dimension, we establish quantitative thresholds to ensure meaningful rewrites from original query x_i to the rewritten query x_i':

- Formality: x_i' must score below 0.5 probability using our formality classifier
- Readability: x_i' must score below 60 on the Flesch Reading Ease Score
- Politeness: The sum of *somewhat polite* and *polite* score on x_i' must above 0.5 using Polite Guard classification model.
- Grammatical correctness (Typos): x_i' must have edit distance ≥ 1 from x_i and must not have the same GLEU score. Edit distance between two strings is the minimum number of operations required to transform one string into the other.
- Grammatical correctness (Round-trip translation): x_i' must differ from x_i after English-Afrikaans-English translation, constrained by ensuring that both queries do not have the same GLEU score and an edit distance of 0.

Each rewrite for each prompt must adhere to these thresholds in order to develop the final dataset we use. In addition to this, we want to ensure that the rewritten queries are semantically similar to the original query. Therefore, in the query-rewriting process, we add an additional constraint where the sentence similarity between the original and rewritten queries must be greater than 0.7, as assessed by the MPNet-v2 model (Song et al., 2020) in the SentenceBert Library (Reimers and Gurevych, 2019).

Evaluation : Flesch Reading score is defined as follows:

$$\text{FRES} = 206.835 - 1.015 \frac{\text{total words}}{\text{total sentences}} - 84.6 \frac{\text{total syllables}}{\text{total words}} \tag{1}$$

Linguistic Variation	Semantic Preservation (%)	Inter-annotator Agreement (%)
Politeness	100.0	100.0
Formality	98.6	100.0
Readability	96.7	96.7
Typos	95.0	100.0
RTT	95.0	100.0

(a) EntityQuestions

Linguistic Variation	Semantic Preservation (%)	Inter-annotator Agreement (%)
Politeness	100.0	100.0
Formality	90.0	96.7
Readability	100.0	100.0
Typos	86.7	96.7
RTT	96.7	96.7

(b) Natural Questions

Table 5: Percentage of queries that preserve semantic meaning after linguistic rewrites, along with inter-annotator agreement rates across 250 annotated samples.

A.2 Semantic Preservation of Linguistic Variations

To ensure our linguistic rewrites preserve semantic meaning, we conducted a qualitative evaluation where two independent researchers annotated randomly sampled query rewrites per linguistic property from EntityQuestions (100 examples) and Natural Questions (150 examples) datasets, totaling 250 queries.

A.2.1 Annotation Guidelines

Annotators were provided with the following protocol to assess whether each rewritten query preserved the semantic meaning of the original:

- Mark as *semantically preserved* if both queries would elicit identical factual answers
- Mark as *not preserved* if the rewrite changes entities, relationships, or temporal/spatial constraints
- Rule for Typos and RTT: If more than 2 characters are edited in any entity mention, mark as *not preserved*. This threshold helps maintain a balance between introducing realistic linguistic variations and preserving essential meaning.
- Ignore stylistic differences that do not affect the core information need

Annotators worked independently and their judgments were compared to compute inter-annotator agreement.

A.2.2 Results and Analysis

Table 5 shows high semantic preservation rates across all linguistic dimensions, ranging from 86.7% to 100%. The high inter-annotator agreement (96.7%-100%) further validates the reliability of our assessment approach. These results demonstrate the robustness of our rewriting methodology in preserving semantic content while introducing natural linguistic variations. Notably, even with our strict 2-character edit threshold for typos and RTT, we achieve high preservation rates (86.7%-96.7%), indicating that our generation process successfully balances realism with semantic fidelity. This semantic preservation is critical for our experimental design, as it ensures that performance differences observed in RAG systems can be attributed to linguistic variations rather than semantic drift.

B Detailed RAG Experimental Settings

B.1 LLMs Generation Hyperparameter Setting

We set `temperature` as 0.5 and `top_q` 0.90 to strike a balance between creativity and accuracy. The `max_tokens` is set to 128 considering the gold answers length.

C Experiment Results on Llama 3.1 Rewriting Queries

We conducted supplementary experiments by sampling 500 queries from the PopQA dataset. We generated rewritten queries using Llama-3.1-70B-Instruct, employing the same rewriting criteria as previously described. Then, we performed RAG experiments using ModernBERT as the retriever and Qwen2.5-7B-Instruct as the generator. The results are summarized below:

Dimension	**Original**	**Rewritten**	**Δ**
Readability	0.624	0.490	**21.5%**
Formality	0.826	0.746	9.7%
Politeness	0.728	0.626	14.0%

Table 6: Retrieval Performance (R@5)

Dimension	**Original**	**Rewritten**	**Δ**
Readability	0.500	0.372	**25.6%**
Formality	0.776	0.684	11.9%
Politeness	0.604	0.520	13.9%

Table 7: RAG Generation Performance (Answer Match - AM Score)

In general, **the RAG system remains notably sensitive to the linguistic variations introduced by LLaMA**. Specifically, queries rewritten for reduced readability caused the most significant performance degradation, with 21.5% in retrieval accuracy and 25.6% in the AM score. Unlike variations introduced by GPT, the system exhibited greater sensitivity to changes in politeness compared to formality. We will present a comprehensive analysis and more detailed results in the revised paper.

D Data and Code Availability

Our code and rewritten query datasets will be released after the peer review stage under the CC BY 4.0 license. All existing datasets, models, and codes used in this work were employed consistently with their intended research purposes.

E Full Retrieval Experiment Results

In this section we provide the absolute results from Contriever retriever experiments that led to Table 2

Dataset	Linguistics	R@5	R@10	R@20	R@100
EntityQuestions	RTT	0.6838	0.7332	0.7710	0.8422
EntityQuestions	Typos	0.6838	0.7332	0.7710	0.8422
EntityQuestions	Formality	0.6846	0.7292	0.7616	0.8240
EntityQuestions	Politeness	0.6744	0.7218	0.7594	0.8310
EntityQuestions	Readability	0.6434	0.6994	0.7452	0.8354
MS MARCO	RTT	0.3412	0.4068	0.4694	0.6102
MS MARCO	Typos	0.3412	0.4068	0.4694	0.6102
MS MARCO	Formality	0.2534	0.3252	0.4048	0.5598
MS MARCO	Politeness	0.2620	0.3414	0.4152	0.5664
MS MARCO	Readability	0.2512	0.3280	0.4072	0.5718
Natural Questions	RTT	0.6246	0.7118	0.7834	0.8822
Natural Questions	Typos	0.6246	0.7118	0.7834	0.8822
Natural Questions	Formality	0.6472	0.7372	0.7954	0.8868
Natural Questions	Politeness	0.6012	0.6946	0.7588	0.8488
Natural Questions	Readability	0.5978	0.6882	0.7582	0.8536
PopQA	RTT	0.5938	0.6614	0.7148	0.8192
PopQA	Typos	0.5938	0.6614	0.7148	0.8192
PopQA	Formality	0.6974	0.7574	0.8066	0.8760
PopQA	Politeness	0.6220	0.6942	0.7576	0.8534
PopQA	Readability	0.6108	0.6856	0.7368	0.8344

Table 8: Contriever Retrieval performance (R@k) of original queries across datasets and linguistic modifications.

Dataset	Linguistics	R@5	R@10	R@20	R@100
EntityQuestions	RTT	0.5842	0.6428	0.6926	0.7896
EntityQuestions	Typos	0.5826	0.6452	0.6964	0.7836
EntityQuestions	Formality	0.6113	0.6626	0.7025	0.7815
EntityQuestions	Politeness	0.6629	0.7090	0.7507	0.8253
EntityQuestions	Readability	0.5887	0.6471	0.6979	0.7935
MS MARCO	RTT	0.3100	0.3662	0.4328	0.5754
MS MARCO	Typos	0.2882	0.3460	0.4052	0.5440
MS MARCO	Formality	0.1502	0.2020	0.2722	0.4365
MS MARCO	Politeness	0.2189	0.2933	0.3764	0.5448
MS MARCO	Readability	0.1982	0.2716	0.3467	0.5235
Natural Questions	RTT	0.4778	0.5718	0.6506	0.7898
Natural Questions	Typos	0.4320	0.5230	0.6180	0.7730
Natural Questions	Formality	0.5479	0.6440	0.7215	0.8456
Natural Questions	Politeness	0.5942	0.6832	0.7516	0.8443
Natural Questions	Readability	0.5519	0.6481	0.7253	0.8346
PopQA	RTT	0.4216	0.4854	0.5422	0.6654
PopQA	Typos	0.4288	0.4890	0.5482	0.6724
PopQA	Formality	0.5582	0.6269	0.6800	0.7856
PopQA	Politeness	0.5704	0.6426	0.7045	0.8081
PopQA	Readability	0.4981	0.5635	0.6192	0.7332

Table 9: Contriever Retrieval performance (R@k) of rewritten queries across datasets and linguistic modifications

Dataset	Linguistics	R@5	R@10	R@20	R@100
EntityQuestions	RTT	0.6614	0.7184	0.7598	0.8284
EntityQuestions	Typos	0.6614	0.7184	0.7598	0.8284
EntityQuestions	Formality	0.6798	0.7240	0.7558	0.8150
EntityQuestions	Politeness	0.6730	0.7214	0.7594	0.8276
EntityQuestions	Readability	0.6132	0.6758	0.7254	0.8108
MS MARCO	RTT	0.3204	0.3916	0.4574	0.5680
MS MARCO	Typos	0.3204	0.3916	0.4574	0.5680
MS MARCO	Readability	0.3982	0.4818	0.5604	0.6746
MS MARCO	Formality	0.4074	0.4896	0.5644	0.6720
MS MARCO	Politeness	0.4030	0.4840	0.5552	0.6638
Natural Questions	RTT	0.6690	0.7556	0.8110	0.8878
Natural Questions	Typos	0.6690	0.7556	0.8110	0.8878
Natural Questions	Readability	0.6512	0.7300	0.7872	0.8614
Natural Questions	Formality	0.6874	0.7700	0.8246	0.8944
Natural Questions	Politeness	0.6538	0.7326	0.7860	0.8600
PopQA	RTT	0.6280	0.6952	0.7508	0.8344
PopQA	Typos	0.6280	0.6952	0.7508	0.8344
PopQA	Readability	0.6518	0.7168	0.7682	0.8432
PopQA	Formality	0.7408	0.7922	0.8316	0.8832
PopQA	Politeness	0.6744	0.7418	0.7962	0.8688

Table 10: ModernBERT Retrieval performance (R@k) for original queries across datasets and linguistic modifications

Dataset	Linguistics	R@5	R@10	R@20	R@100
PopQA	Readability	0.5363	0.6148	0.6751	0.7796
PopQA	RTT	0.4416	0.5090	0.5668	0.6856
PopQA	Typos	0.4868	0.5646	0.6242	0.7406
PopQA	Formality	0.6395	0.7109	0.7649	0.8493
PopQA	Politeness	0.6023	0.6776	0.7334	0.8317
EntityQuestions	Readability	0.5325	0.5992	0.6607	0.7729
EntityQuestions	RTT	0.5632	0.6260	0.6784	0.7800
EntityQuestions	Typos	0.5886	0.6518	0.7038	0.7918
EntityQuestions	Formality	0.6251	0.6739	0.7156	0.7949
EntityQuestions	Politeness	0.6502	0.7002	0.7436	0.8199
MS MARCO	Readability	0.3401	0.4291	0.5119	0.6461
MS MARCO	RTT	0.2642	0.3340	0.3908	0.5146
MS MARCO	Typos	0.2810	0.3540	0.4188	0.5396
MS MARCO	Formality	0.3441	0.4311	0.5113	0.6479
MS MARCO	Politeness	0.3821	0.4637	0.5385	0.6579
Natural Questions	Readability	0.5855	0.6741	0.7435	0.8393
Natural Questions	RTT	0.5470	0.6416	0.7192	0.8328
Natural Questions	Typos	0.5790	0.6838	0.7512	0.8502
Natural Questions	Formality	0.6220	0.7177	0.7890	0.8762
Natural Questions	Politeness	0.6213	0.7054	0.7701	0.8539

Table 11: ModernBERT Retrieval performance (R@k) for rewritten queries across datasets and linguistic modifications

E.1 Scaling Number of documents

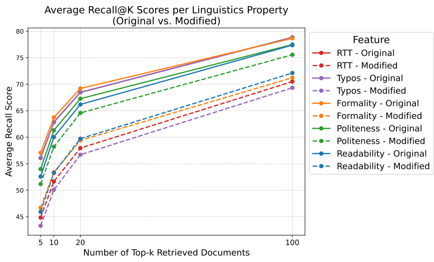

Figure 4: Average Recall@K increase as Number of Top-K Documents increases – Contriever.

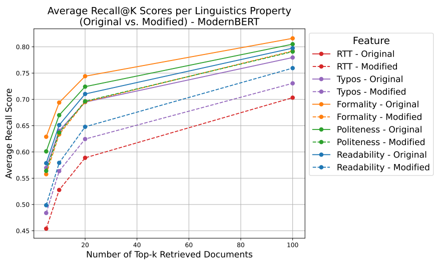

Figure 5: Average Recall@K increase as Number of Top-K Documents increases – ModernBERT.

As mentioned in Section 5.1, the scaling of the number of documents decreases the degradation in performance, but does not mitigate the overall issue. As you can see in Figures 4 and 5, as K increases, higher recall benefits both original and rewritten queries, with performance gaps narrowing as correctly ranked documents appear at lower positions—suggesting linguistic variations primarily affect ranking order rather than complete retrieval failure. This hypothesis is confirmed in Section 5.4.2, where the reranking shows considerable performance gains at R@5, showing that retrieval systems tend to push the correct documents for rewritten queries to lower ranks, and reranking helps prioritize them again, which is demonstrated by the reduction in performance degradation in Figure 3.

F Full RAG Experiment Results

Contriever, Formality, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.5858	0.4719	19.45%	0.4432	0.3742	15.57%	0.1416	0.0955	32.58%	0.6026	0.5446	9.62%
gemma-2-9b-it	0.6006	0.4995	16.83%	0.4784	0.4177	12.68%	0.1420	0.0971	31.64%	0.6148	0.5509	10.39%
gemma-2-27b-it	0.6126	0.5094	16.85%	0.5046	0.4447	11.88%	0.1384	0.1025	25.96%	0.6300	0.5691	9.67%
Llama-3.1-8B-Instruct	0.6122	0.5098	16.73%	0.4930	0.4376	11.24%	0.1580	0.1121	29.03%	0.6156	0.5621	8.70%
Llama-3.1-70B-Instruct	0.6386	0.5297	17.05%	0.5170	0.4331	16.22%	0.1586	0.1057	33.38%	0.6352	0.5747	9.52%
Qwen2.5-3B-Instruct	0.5648	0.4596	18.63%	0.4288	0.3669	14.43%	0.1416	0.1002	29.24%	0.5626	0.5086	9.60%
Qwen2.5-7B-Instruct	0.5910	0.4864	17.70%	0.4774	0.4041	15.35%	0.1544	0.1085	29.71%	0.6106	0.5529	9.44%
Qwen2.5-32B-Instruct	0.6176	0.5128	16.97%	0.5164	0.4586	11.19%	0.1494	0.1115	25.39%	0.6242	0.5720	8.36%
Qwen2.5-72B-Instruct	0.6266	0.5249	16.24%	0.5280	0.4709	10.82%	0.1462	0.1079	26.22%	0.6366	0.5872	7.76%

Contriever, Politeness, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5206	0.4903	5.83%	0.4870	0.4857	0.27%	0.1696	0.1463	13.72%	0.6300	0.5691	9.67%
gemma-2-2b-it	0.4798	0.4508	6.04%	0.4040	0.4147	2.66%	0.1724	0.1400	18.79%	0.6026	0.5446	9.62%
gemma-2-9b-it	0.4974	0.4685	5.80%	0.4560	0.4528	0.70%	0.1736	0.1487	14.36%	0.6148	0.5509	10.39%
Llama-3.1-70B-Instruct	0.5384	0.5059	6.03%	0.4966	0.4811	3.13%	0.1900	0.1528	19.58%	0.6352	0.5747	9.52%
Llama-3.1-8B-Instruct	0.5142	0.4809	6.47%	0.4806	0.4735	1.48%	0.1874	0.1593	15.01%	0.6156	0.5621	8.70%
Qwen2.5-32B-Instruct	0.5184	0.4799	7.42%	0.4948	0.4787	3.25%	0.1776	0.1525	14.15%	0.6242	0.5720	8.36%
Qwen2.5-3B-Instruct	0.4436	0.4026	9.24%	0.3938	0.3830	2.74%	0.1680	0.1423	15.28%	0.5626	0.5086	9.60%
Qwen2.5-72B-Instruct	0.5216	0.4906	5.94%	0.5156	0.4991	3.21%	0.1722	0.1348	21.72%	0.6366	0.5872	7.76%
Qwen2.5-7B-Instruct	0.4830	0.4544	5.92%	0.4402	0.4329	1.67%	0.1814	0.1552	14.44%	0.6106	0.5529	9.44%

Contriever, Readability, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5252	0.4259	18.91%	0.4936	0.4499	8.85%	0.1668	0.1512	9.35%	0.5600	0.5064	9.57%
gemma-2-2b-it	0.4778	0.3762	21.26%	0.4146	0.3559	14.15%	0.1686	0.1413	16.21%	0.5254	0.4660	11.31%
gemma-2-9b-it	0.5026	0.4095	18.53%	0.4602	0.4041	12.20%	0.1638	0.1439	12.13%	0.5488	0.4872	11.22%
Llama-3.1-70B-Instruct	0.5426	0.4414	18.65%	0.5068	0.4383	13.51%	0.1780	0.1482	16.74%	0.5636	0.4947	12.23%
Llama-3.1-8B-Instruct	0.5184	0.4186	19.25%	0.4880	0.4235	13.22%	0.1798	0.1512	15.91%	0.5388	0.4851	9.97%
Qwen2.5-32B-Instruct	0.5184	0.4168	19.60%	0.4964	0.4312	13.13%	0.1794	0.1607	10.41%	0.5442	0.4725	13.18%
Qwen2.5-3B-Instruct	0.4480	0.3379	24.57%	0.3880	0.3272	15.67%	0.1654	0.1371	17.13%	0.4762	0.3840	19.36%
Qwen2.5-72B-Instruct	0.5272	0.4310	18.25%	0.5140	0.4523	12.01%	0.1714	0.1551	9.49%	0.5548	0.4816	13.19%
Qwen2.5-7B-Instruct	0.4864	0.3890	20.02%	0.4418	0.3817	13.60%	0.1770	0.1523	13.97%	0.5154	0.4412	14.40%

Contriever, Round-trip Translation, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.4942	0.3290	33.43%	0.4808	0.3438	28.49%	0.1270	0.0952	25.04%	0.5914	0.4744	19.78%
gemma-2-2b-it	0.4628	0.3088	33.28%	0.4090	0.2700	33.99%	0.1282	0.0856	33.23%	0.5568	0.4378	21.37%
gemma-2-9b-it	0.4746	0.3148	33.67%	0.4512	0.3208	28.90%	0.1304	0.0970	25.61%	0.5672	0.4432	21.86%
Llama-3.1-70B-Instruct	0.5064	0.3288	35.07%	0.1420	0.3276	130.70%	0.1420	0.1050	26.06%	0.5818	0.4606	20.83%
Llama-3.1-8B-Instruct	0.4708	0.3100	34.15%	0.4622	0.3184	31.11%	0.1424	0.0990	30.48%	0.5686	0.4166	26.73%
Qwen2.5-32B-Instruct	0.4922	0.3244	34.09%	0.4926	0.3462	29.72%	0.1376	0.1004	27.03%	0.5830	0.4668	19.93%
Qwen2.5-3B-Instruct	0.4146	0.2792	32.66%	0.3970	0.2632	33.70%	0.1274	0.0856	32.81%	0.5134	0.3940	23.26%
Qwen2.5-72B-Instruct	0.4968	0.3278	34.02%	0.5066	0.3558	29.77%	0.1316	0.0958	27.20%	0.5956	0.4784	19.68%
Qwen2.5-7B-Instruct	0.4544	0.3082	32.17%	0.4474	0.2988	33.21%	0.1406	0.0956	32.01%	0.5630	0.4320	23.27%

Contriever, Typos, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.4942	0.3894	21.21%	0.4808	0.4120	14.31%	0.1270	0.1062	16.38%	0.5914	0.5232	11.53%
gemma-2-2b-it	0.4628	0.3458	25.28%	0.4090	0.3196	21.86%	0.1282	0.1036	19.19%	0.5568	0.4794	13.90%
gemma-2-9b-it	0.4746	0.3660	22.88%	0.4512	0.3684	18.35%	0.1304	0.1078	17.33%	0.5672	0.4974	12.31%
Llama-3.1-70B-Instruct	0.5064	0.3886	23.26%	0.4866	0.3830	21.29%	0.1420	0.1164	18.03%	0.5818	0.4916	15.50%
Llama-3.1-8B-Instruct	0.4708	0.3460	26.51%	0.4622	0.3650	21.03%	0.1424	0.1148	19.38%	0.5686	0.4754	16.39%
Qwen2.5-32B-Instruct	0.4922	0.3654	25.76%	0.4926	0.3992	18.96%	0.1376	0.1146	16.72%	0.5830	0.4920	15.61%
Qwen2.5-3B-Instruct	0.4146	0.3038	26.72%	0.3970	0.2868	27.76%	0.1274	0.0952	25.27%	0.5134	0.4196	18.27%
Qwen2.5-72B-Instruct	0.4976	0.3796	23.71%	0.5066	0.4336	14.41%	0.1316	0.1104	16.11%	0.5936	0.5198	12.43%
Qwen2.5-7B-Instruct	0.4544	0.3348	26.32%	0.4474	0.3434	23.25%	0.1406	0.1124	20.06%	0.5630	0.4866	13.57%

Table 12: RAG experiment results with Contriever as retrieval model on AM scores.

Contriever, Formality, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.1454	0.0151	89.64%	0.0550	0.0205	62.67%	0.0196	0.0058	70.41%	0.1914	0.0330	82.76%
gemma-2-9b-it	0.0214	0.0066	69.16%	0.0196	0.0093	52.72%	0.0340	0.0089	73.92%	0.0888	0.0341	61.64%
gemma-2-27b-it	0.0044	0.0122	177.27%	0.0188	0.0147	21.63%	0.0058	0.0037	36.78%	0.0292	0.0505	72.83%
Llama-3.1-8B-Instruct	0.3282	0.1348	58.93%	0.1482	0.0525	64.60%	0.0216	0.0055	74.69%	0.2896	0.1429	50.67%
Llama-3.1-70B-Instruct	0.3432	0.0884	74.24%	0.1234	0.0369	70.12%	0.0244	0.0069	71.86%	0.2996	0.1177	60.70%
Qwen2.5-3B-Instruct	0.3906	0.1705	56.34%	0.1810	0.0842	53.48%	0.0186	0.0058	68.82%	0.3404	0.2079	38.93%
Qwen2.5-7B-Instruct	0.4242	0.1847	56.45%	0.1892	0.1138	39.85%	0.0352	0.0148	57.95%	0.2862	0.1521	46.87%
Qwen2.5-32B-Instruct	0.3736	0.1257	66.35%	0.1086	0.0609	43.95%	0.0236	0.0105	55.37%	0.3580	0.1762	50.78%
Qwen2.5-72B-Instruct	0.4656	0.1781	61.76%	0.1490	0.0733	50.83%	0.0204	0.0094	53.92%	0.3890	0.1992	48.79%

Contriever, Politeness, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0034	0.0152	347.06%	0.0148	0.0207	40.09%	0.0060	0.0054	10.00%	0.0188	0.0259	37.59%
gemma-2-2b-it	0.0970	0.0410	57.73%	0.0450	0.0424	5.78%	0.0200	0.0082	59.00%	0.1752	0.0358	79.57%
gemma-2-9b-it	0.0170	0.0068	60.00%	0.0130	0.0105	19.49%	0.0512	0.0153	70.18%	0.0912	0.0278	69.52%
Llama-3.1-70B-Instruct	0.1906	0.0875	54.11%	0.1064	0.0436	59.02%	0.0236	0.0082	65.25%	0.2330	0.1326	43.09%
Llama-3.1-8B-Instruct	0.1998	0.1657	17.05%	0.1024	0.0673	34.31%	0.0272	0.0077	71.57%	0.2840	0.2535	10.73%
Qwen2.5-32B-Instruct	0.2020	0.1683	16.70%	0.0546	0.0851	55.80%	0.0192	0.0155	19.44%	0.3344	0.3055	8.65%
Qwen2.5-3B-Instruct	0.2694	0.2473	8.22%	0.1920	0.1695	11.70%	0.0236	0.0159	32.77%	0.3338	0.3331	0.22%
Qwen2.5-72B-Instruct	0.3270	0.2855	12.68%	0.1030	0.1504	46.02%	0.0242	0.0187	22.59%	0.3912	0.3877	0.90%
Qwen2.5-7B-Instruct	0.2736	0.2719	0.61%	0.1700	0.1984	16.71%	0.0384	0.0261	32.12%	0.3160	0.3382	7.03%

Contriever, Readability, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0008	0.0010	25.00%	0.0056	0.0054	3.57%	0.0042	0.0035	17.46%	0.0062	0.0024	61.29%
gemma-2-2b-it	0.0928	0.0661	28.81%	0.0282	0.0414	46.81%	0.0134	0.0205	52.74%	0.0674	0.0692	2.67%
gemma-2-9b-it	0.0114	0.0379	232.16%	0.0072	0.0223	210.19%	0.0454	0.0521	14.68%	0.0262	0.0545	107.89%
Llama-3.1-70B-Instruct	0.2596	0.0733	71.78%	0.0962	0.0543	43.59%	0.0176	0.0071	59.85%	0.1228	0.0776	36.81%
Llama-3.1-8B-Instruct	0.2438	0.1507	38.17%	0.1024	0.0709	30.79%	0.0168	0.0071	57.94%	0.1412	0.1187	15.96%
Qwen2.5-32B-Instruct	0.2188	0.1020	53.38%	0.0472	0.0719	52.26%	0.0126	0.0137	8.99%	0.1516	0.1405	7.30%
Qwen2.5-3B-Instruct	0.2910	0.1398	51.96%	0.1902	0.1187	37.57%	0.0214	0.0152	28.97%	0.1694	0.1279	24.52%
Qwen2.5-72B-Instruct	0.3788	0.2529	33.23%	0.1002	0.1301	29.87%	0.0168	0.0151	10.32%	0.2200	0.2015	8.39%
Qwen2.5-7B-Instruct	0.3244	0.1885	41.90%	0.1476	0.1629	10.39%	0.0348	0.0441	26.82%	0.1354	0.1303	3.79%

Contriever, Round-trip Translation, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0004	0.0012	200.00%	0.0144	0.0164	13.89%	0.0046	0.0058	26.09%	0.0104	0.0116	11.54%
gemma-2-2b-it	0.0836	0.0630	24.64%	0.0474	0.0288	39.24%	0.0126	0.0102	19.05%	0.0738	0.0582	21.14%
gemma-2-9b-it	0.0078	0.0062	20.51%	0.0176	0.0154	12.50%	0.0380	0.0294	22.63%	0.0552	0.0358	35.14%
Llama-3.1-70B-Instruct	0.2350	0.1274	45.79%	0.0324	0.0876	170.37%	0.0172	0.0158	8.14%	0.1742	0.1222	29.85%
Llama-3.1-8B-Instruct	0.2500	0.1400	44.00%	0.1614	0.1072	33.58%	0.0306	0.0150	50.98%	0.2052	0.1302	36.55%
Qwen2.5-32B-Instruct	0.1360	0.0658	51.62%	0.0800	0.0518	35.25%	0.0100	0.0074	26.00%	0.1734	0.1328	23.41%
Qwen2.5-3B-Instruct	0.2104	0.1042	50.48%	0.1652	0.0850	48.55%	0.0202	0.0158	21.78%	0.2234	0.1606	28.11%
Qwen2.5-72B-Instruct	0.3178	0.1694	46.70%	0.1460	0.0844	42.19%	0.0154	0.0116	24.68%	0.2850	0.2220	22.11%
Qwen2.5-7B-Instruct	0.2438	0.1452	40.44%	0.1560	0.1174	24.74%	0.0334	0.0272	18.56%	0.1536	0.1264	17.71%

Contriever, Typos, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0004	0.0010	150.00%	0.0144	0.0124	13.89%	0.0046	0.0024	47.83%	0.0104	0.0154	48.08%
gemma-2-2b-it	0.0836	0.0588	29.67%	0.0474	0.0338	28.69%	0.0126	0.0114	9.52%	0.0738	0.0556	24.66%
gemma-2-9b-it	0.0078	0.0084	7.69%	0.0176	0.0134	23.86%	0.0380	0.0300	21.05%	0.0552	0.0548	0.72%
Llama-3.1-70B-Instruct	0.2350	0.1466	37.62%	0.1302	0.0944	27.50%	0.0172	0.0118	31.40%	0.1742	0.1302	25.26%
Llama-3.1-8B-Instruct	0.2500	0.1676	32.96%	0.1614	0.1106	31.47%	0.0306	0.0186	39.22%	0.2052	0.1742	15.11%
Qwen2.5-32B-Instruct	0.1360	0.0848	37.65%	0.0800	0.0548	31.50%	0.0100	0.0084	16.00%	0.1734	0.1348	22.26%
Qwen2.5-3B-Instruct	0.2104	0.1314	37.55%	0.1652	0.1064	35.59%	0.0202	0.0156	22.77%	0.2234	0.1720	23.01%
Qwen2.5-72B-Instruct	0.3160	0.2100	33.54%	0.1460	0.0962	34.11%	0.0154	0.0110	28.57%	0.2784	0.2264	18.68%
Qwen2.5-7B-Instruct	0.2438	0.1512	37.98%	0.1560	0.1028	34.10%	0.0334	0.0268	19.76%	0.1536	0.1144	25.52%

Table 13: RAG experiment results with Contriever as retrieval model on EM scores.

Contriever, Formality, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.3457	0.2167	37.31%	0.2285	0.1737	23.99%	0.2474	0.1933	21.86%	0.3694	0.2298	37.79%
gemma-2-9b-it	0.2809	0.2274	19.06%	0.2142	0.1911	10.75%	0.2614	0.1988	23.97%	0.2996	0.2436	18.70%
gemma-2-27b-it	0.2756	0.2364	14.24%	0.2193	0.1972	10.10%	0.2432	0.2016	17.10%	0.2573	0.2569	0.12%
Llama-3.1-8B-Instruct	0.4601	0.2842	38.22%	0.3019	0.1921	36.37%	0.2478	0.1910	22.93%	0.4349	0.2964	31.85%
Llama-3.1-70B-Instruct	0.4676	0.2610	44.17%	0.2861	0.1858	35.06%	0.2502	0.1884	24.71%	0.4406	0.2836	35.65%
Qwen2.5-3B-Instruct	0.4801	0.2843	40.77%	0.3212	0.2015	37.27%	0.2401	0.1764	26.55%	0.4530	0.3201	29.34%
Qwen2.5-7B-Instruct	0.5191	0.3203	38.30%	0.3464	0.2469	28.73%	0.2601	0.1970	24.25%	0.4234	0.3041	28.18%
Qwen2.5-32B-Instruct	0.4835	0.2844	41.18%	0.2789	0.2072	25.71%	0.2541	0.2003	21.15%	0.4824	0.3295	31.69%
Qwen2.5-72B-Instruct	0.5521	0.3208	41.90%	0.3159	0.2203	30.26%	0.2525	0.2025	19.80%	0.5197	0.3480	33.05%

Contriever, Politeness, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2376	0.2257	5.00%	0.2153	0.2158	0.22%	0.2418	0.2230	7.79%	0.2395	0.2443	1.99%
gemma-2-2b-it	0.2789	0.2273	18.52%	0.2135	0.2042	4.36%	0.2452	0.2171	11.45%	0.3630	0.2491	31.39%
gemma-2-9b-it	0.2354	0.2031	13.72%	0.2051	0.1961	4.36%	0.2722	0.2271	16.58%	0.2964	0.2465	16.85%
Llama-3.1-70B-Instruct	0.3254	0.2443	24.93%	0.2698	0.2022	25.05%	0.2474	0.2135	13.72%	0.3954	0.3075	22.24%
Llama-3.1-8B-Instruct	0.3312	0.2908	12.19%	0.2593	0.2179	15.98%	0.2504	0.2146	14.29%	0.4461	0.4117	7.70%
Qwen2.5-32B-Instruct	0.3381	0.2963	12.38%	0.2361	0.2474	4.82%	0.2465	0.2241	9.07%	0.4744	0.4482	5.51%
Qwen2.5-3B-Instruct	0.3592	0.3223	10.28%	0.3236	0.2917	9.86%	0.2416	0.2129	11.88%	0.4636	0.4574	1.33%
Qwen2.5-72B-Instruct	0.4211	0.3810	9.51%	0.2784	0.3069	10.26%	0.2503	0.2254	9.96%	0.5414	0.5366	0.89%
Qwen2.5-7B-Instruct	0.3806	0.3632	4.59%	0.3241	0.3359	3.65%	0.2606	0.2300	11.73%	0.4635	0.4804	3.65%

Contriever, Readability, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2390	0.1688	29.36%	0.2089	0.1806	13.55%	0.2299	0.1939	15.65%	0.2213	0.1890	14.59%
gemma-2-2b-it	0.2754	0.1968	28.55%	0.1968	0.1752	10.99%	0.2304	0.1959	14.97%	0.2739	0.2429	11.34%
gemma-2-9b-it	0.2385	0.1894	20.58%	0.2010	0.1846	8.15%	0.2569	0.2245	12.64%	0.2369	0.2295	3.11%
Llama-3.1-70B-Instruct	0.3804	0.2019	46.92%	0.2591	0.1865	28.03%	0.2331	0.1837	21.19%	0.2926	0.2267	22.52%
Llama-3.1-8B-Instruct	0.3682	0.2547	30.83%	0.2568	0.1963	23.54%	0.2304	0.1822	20.92%	0.3168	0.2756	12.99%
Qwen2.5-32B-Instruct	0.3543	0.2240	36.78%	0.2271	0.2059	9.33%	0.2316	0.1958	15.46%	0.3115	0.2857	8.29%
Qwen2.5-3B-Instruct	0.3817	0.2532	33.66%	0.3122	0.2464	21.08%	0.2309	0.1848	19.97%	0.3055	0.2682	12.20%
Qwen2.5-72B-Instruct	0.4681	0.3387	27.64%	0.2716	0.2601	4.24%	0.2376	0.2015	15.16%	0.3820	0.3439	9.96%
Qwen2.5-7B-Instruct	0.4173	0.2902	30.45%	0.2957	0.2902	1.84%	0.2493	0.2212	11.25%	0.3021	0.2848	5.73%

Contriever, Round-trip Translation, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2156	0.1322	38.67%	0.2102	0.1594	24.20%	0.2151	0.1807	15.96%	0.2206	0.1888	14.41%
gemma-2-2b-it	0.2574	0.1672	35.03%	0.2119	0.1448	31.65%	0.2182	0.1789	17.99%	0.2719	0.2216	18.51%
gemma-2-9b-it	0.2116	0.1302	38.45%	0.2041	0.1539	24.60%	0.2389	0.1963	17.84%	0.2550	0.2059	19.25%
Llama-3.1-70B-Instruct	0.3436	0.2038	40.69%	0.2321	0.2024	12.80%	0.2223	0.1857	16.44%	0.3382	0.2582	23.65%
Llama-3.1-8B-Instruct	0.3534	0.2143	39.36%	0.3134	0.2214	29.37%	0.2318	0.1816	21.67%	0.3654	0.2492	31.80%
Qwen2.5-32B-Instruct	0.2837	0.1646	41.99%	0.2496	0.1733	30.57%	0.2199	0.1841	16.27%	0.3320	0.2639	20.52%
Qwen2.5-3B-Instruct	0.3061	0.1789	41.55%	0.2998	0.1825	39.11%	0.2162	0.1720	20.44%	0.3651	0.2859	21.70%
Qwen2.5-72B-Instruct	0.4089	0.2375	41.91%	0.3098	0.2054	33.71%	0.2248	0.1896	15.65%	0.4348	0.3451	20.62%
Qwen2.5-7B-Instruct	0.3480	0.2204	36.66%	0.3090	0.2270	26.53%	0.2389	0.1952	18.27%	0.3230	0.2672	17.27%

Contriever, Typos, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2156	0.1658	23.08%	0.2102	0.1852	11.90%	0.2151	0.1951	9.28%	0.2206	0.2006	9.09%
gemma-2-2b-it	0.2574	0.1878	27.05%	0.2119	0.1701	19.74%	0.2182	0.1975	9.50%	0.2719	0.2308	15.12%
gemma-2-9b-it	0.2116	0.1604	24.20%	0.2041	0.1719	15.76%	0.2389	0.2114	11.52%	0.2550	0.2295	10.00%
Llama-3.1-70B-Instruct	0.3436	0.2450	28.71%	0.2885	0.2269	21.35%	0.2223	0.1979	10.95%	0.3382	0.2710	19.87%
Llama-3.1-8B-Instruct	0.3534	0.2499	29.27%	0.3134	0.2400	23.44%	0.2318	0.2031	12.38%	0.3654	0.3107	14.97%
Qwen2.5-32B-Instruct	0.2837	0.1982	30.13%	0.2496	0.1992	20.18%	0.2199	0.1979	9.99%	0.3320	0.2723	17.98%
Qwen2.5-3B-Instruct	0.3061	0.2093	31.61%	0.2998	0.2132	28.89%	0.2162	0.1874	13.31%	0.3651	0.2920	20.04%
Qwen2.5-72B-Instruct	0.4067	0.2903	28.62%	0.3098	0.2488	19.71%	0.2248	0.2083	7.36%	0.4299	0.3632	15.51%
Qwen2.5-7B-Instruct	0.3480	0.2378	31.67%	0.3090	0.2324	24.79%	0.2389	0.2107	11.82%	0.3230	0.2613	19.11%

Table 14: RAG experiment results with Contriever as retrieval model on F1 scores.

ModernBERT, Formality, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.6126	0.5291	13.64%	0.4608	0.4035	12.44%	0.1690	0.1296	23.31%	0.6012	0.5551	7.66%
gemma-2-9b-it	0.6360	0.5618	11.67%	0.4984	0.4488	9.95%	0.1724	0.1385	19.68%	0.6178	0.5631	8.85%
gemma-2-27b-it	0.6508	0.5713	12.21%	0.5230	0.4711	9.92%	0.1676	0.1355	19.13%	0.6264	0.5813	7.19%
Llama-3.1-8B-Instruct	0.6488	0.5695	12.23%	0.5228	0.4679	10.51%	0.1982	0.1633	17.59%	0.6158	0.5717	7.16%
Llama-3.1-70B-Instruct	0.6722	0.5950	11.48%	0.5378	0.4728	12.09%	0.1986	0.1609	19.00%	0.6350	0.5893	7.19%
Qwen2.5-3B-Instruct	0.5806	0.5169	10.97%	0.4534	0.4018	11.38%	0.1834	0.1451	20.87%	0.5618	0.5196	7.51%
Qwen2.5-7B-Instruct	0.6290	0.5473	12.99%	0.5010	0.4436	11.46%	0.1848	0.1481	19.84%	0.6146	0.5649	8.08%
Qwen2.5-32B-Instruct	0.6512	0.5773	11.34%	0.5274	0.4860	7.85%	0.1766	0.1505	14.80%	0.6206	0.5819	6.24%
Qwen2.5-72B-Instruct	0.6576	0.5874	10.68%	0.5468	0.4970	9.11%	0.1776	0.1408	20.72%	0.6312	0.5949	5.75%

ModernBERT, Politeness, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5632	0.5185	7.93%	0.5132	0.4959	3.38%	0.1958	0.1838	6.13%	0.6138	0.6017	1.97%
gemma-2-2b-it	0.5220	0.4783	8.37%	0.4394	0.4200	4.42%	0.2020	0.1795	11.16%	0.5938	0.5762	2.96%
gemma-2-9b-it	0.5394	0.4933	8.54%	0.4830	0.4656	3.60%	0.2044	0.1877	8.19%	0.6054	0.5872	3.01%
Llama-3.1-70B-Instruct	0.5734	0.5333	7.00%	0.5328	0.4993	6.29%	0.2306	0.2078	9.89%	0.6124	0.5903	3.61%
Llama-3.1-8B-Instruct	0.5498	0.5081	7.59%	0.5080	0.4904	3.46%	0.2302	0.2074	9.90%	0.5912	0.5705	3.50%
Qwen2.5-32B-Instruct	0.5602	0.5117	8.65%	0.5154	0.4909	4.76%	0.2062	0.1925	6.66%	0.6020	0.5681	5.64%
Qwen2.5-3B-Instruct	0.4712	0.4168	11.54%	0.4250	0.3999	5.91%	0.2130	0.1868	12.30%	0.5506	0.5080	7.74%
Qwen2.5-72B-Instruct	0.5626	0.5212	7.36%	0.5398	0.5104	5.45%	0.1970	0.1643	16.62%	0.6182	0.5943	3.86%
Qwen2.5-7B-Instruct	0.5360	0.4783	10.76%	0.4782	0.4541	5.05%	0.2128	0.1977	7.08%	0.5962	0.5683	4.69%

ModernBERT, Readability, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5592	0.4689	16.15%	0.5220	0.4629	11.33%	0.1968	0.1804	8.33%	0.5480	0.4785	12.68%
gemma-2-2b-it	0.5180	0.4107	20.71%	0.4460	0.3710	16.82%	0.2038	0.1683	17.44%	0.5088	0.4252	16.43%
gemma-2-9b-it	0.5394	0.4471	17.12%	0.4926	0.4239	13.95%	0.1970	0.1751	11.13%	0.5356	0.4539	15.25%
Llama-3.1-70B-Instruct	0.5768	0.4749	17.66%	0.5436	0.4579	15.77%	0.2226	0.1930	13.30%	0.5522	0.4633	16.09%
Llama-3.1-8B-Instruct	0.5546	0.4588	17.27%	0.5180	0.4399	15.07%	0.2210	0.1873	15.23%	0.5194	0.4503	13.30%
Qwen2.5-32B-Instruct	0.5550	0.4607	17.00%	0.5234	0.4501	14.01%	0.2800	0.1901	32.10%	0.5310	0.4333	18.41%
Qwen2.5-3B-Instruct	0.4742	0.3682	22.35%	0.4232	0.3443	18.65%	0.2070	0.1795	13.27%	0.4560	0.3459	24.14%
Qwen2.5-72B-Instruct	0.5620	0.4707	16.25%	0.5444	0.4703	13.62%	0.2042	0.1871	8.36%	0.5352	0.4441	17.03%
Qwen2.5-7B-Instruct	0.5286	0.4255	19.50%	0.4832	0.4003	17.15%	0.2118	0.1873	11.55%	0.5104	0.4065	20.35%

ModernBERT, Round-trip Translation, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5304	0.3564	32.81%	0.5052	0.3838	24.03%	0.1530	0.1106	27.71%	0.5836	0.4626	20.73%
gemma-2-2b-it	0.4876	0.3280	32.73%	0.4272	0.3072	28.09%	0.1590	0.1116	29.81%	0.5384	0.4174	22.47%
gemma-2-9b-it	0.5092	0.3366	33.90%	0.4760	0.3594	24.50%	0.1604	0.1192	25.69%	0.5602	0.4318	22.92%
Llama-3.1-70B-Instruct	0.5388	0.3486	35.30%	0.5182	0.3864	25.43%	0.1748	0.1314	24.83%	0.5728	0.4490	21.61%
Llama-3.1-8B-Instruct	0.5030	0.3256	35.27%	0.4954	0.3594	27.45%	0.1860	0.1258	32.37%	0.5554	0.4162	25.06%
Qwen2.5-32B-Instruct	0.5230	0.3446	34.11%	0.5170	0.3888	24.80%	0.1642	0.1204	26.67%	0.5678	0.4488	20.96%
Qwen2.5-3B-Instruct	0.4366	0.2930	32.89%	0.4372	0.3028	30.74%	0.1674	0.1136	32.14%	0.4908	0.3704	24.53%
Qwen2.5-72B-Instruct	0.5314	0.3524	33.68%	0.5272	0.3992	24.28%	0.1558	0.1130	27.47%	0.5798	0.4664	19.56%
Qwen2.5-7B-Instruct	0.4950	0.3264	34.06%	0.4854	0.3414	29.67%	0.1656	0.1254	24.28%	0.5522	0.4320	21.77%

ModernBERT, Typos, AM Score												
	PopQA			**Natural Questions**			**MS MARCO**			**EntityQuestions**		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.5304	0.4338	18.21%	0.5052	0.4708	6.81%	0.1530	0.1472	3.79%	0.5836	0.5212	10.69%
gemma-2-2b-it	0.4876	0.3888	20.26%	0.4272	0.3826	10.44%	0.1590	0.1428	10.19%	0.5384	0.4826	10.36%
gemma-2-9b-it	0.5092	0.4094	19.60%	0.4760	0.4370	8.19%	0.1604	0.1492	6.98%	0.5602	0.5020	10.39%
Llama-3.1-70B-Instruct	0.5388	0.4338	19.49%	0.5182	0.4726	8.80%	0.1748	0.1654	5.38%	0.5728	0.5000	12.71%
Llama-3.1-8B-Instruct	0.5030	0.3926	21.95%	0.4954	0.4432	10.54%	0.1860	0.1630	12.37%	0.5554	0.4858	12.53%
Qwen2.5-32B-Instruct	0.5230	0.4200	19.69%	0.5170	0.4690	9.28%	0.1642	0.1522	7.31%	0.5678	0.5018	11.62%
Qwen2.5-3B-Instruct	0.4366	0.3328	23.77%	0.4372	0.3688	15.65%	0.1674	0.1468	12.31%	0.4908	0.4260	13.20%
Qwen2.5-72B-Instruct	0.5314	0.4290	19.27%	0.5272	0.4864	7.74%	0.1558	0.1430	8.22%	0.5798	0.5198	10.35%
Qwen2.5-7B-Instruct	0.4950	0.3828	22.67%	0.4854	0.4214	13.19%	0.1656	0.1546	6.64%	0.5522	0.4860	11.99%

Table 15: RAG experiment results with ModernBERT as retrieval model on AM scores.

ModernBERT, Formality, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.1554	0.0167	89.23%	0.0558	0.0204	63.44%	0.0220	0.0066	70.00%	0.1964	0.0299	84.76%
gemma-2-9b-it	0.0282	0.0072	74.47%	0.0210	0.0091	56.83%	0.0358	0.0096	73.18%	0.0992	0.0356	64.11%
gemma-2-27b-it	0.0052	0.0105	101.28%	0.0202	0.0141	30.36%	0.0096	0.0039	59.03%	0.0320	0.0496	55.00%
Llama-3.1-8B-Instruct	0.3500	0.1525	56.42%	0.1642	0.0551	66.42%	0.0286	0.0115	59.67%	0.2886	0.1429	50.47%
Llama-3.1-70B-Instruct	0.3748	0.1008	73.11%	0.1342	0.0375	72.03%	0.0324	0.0123	61.93%	0.3088	0.1203	61.03%
Qwen2.5-3B-Instruct	0.4016	0.1943	51.63%	0.1878	0.0921	50.98%	0.0248	0.0095	61.56%	0.3384	0.2162	36.11%
Qwen2.5-7B-Instruct	0.4554	0.2097	53.96%	0.2116	0.1267	40.11%	0.0460	0.0224	51.30%	0.2936	0.1527	48.00%
Qwen2.5-32B-Instruct	0.3986	0.1289	67.67%	0.1096	0.0643	41.36%	0.0260	0.0139	46.67%	0.3562	0.1733	51.34%
Qwen2.5-72B-Instruct	0.4924	0.2003	59.33%	0.1612	0.0794	50.74%	0.0254	0.0129	49.34%	0.3920	0.1996	49.08%

ModernBERT, Politeness, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0038	0.0124	226.32%	0.0132	0.0233	76.26%	0.0080	0.0056	30.00%	0.0166	0.0294	77.11%
gemma-2-2b-it	0.1046	0.0437	58.19%	0.0476	0.0434	8.82%	0.0218	0.0101	53.52%	0.1774	0.0355	79.97%
gemma-2-9b-it	0.0218	0.0077	64.83%	0.0122	0.0111	8.74%	0.0514	0.0151	70.69%	0.0932	0.0333	64.31%
Llama-3.1-70B-Instruct	0.2240	0.0986	55.98%	0.1106	0.0487	55.94%	0.0274	0.0129	53.04%	0.2366	0.1333	43.65%
Llama-3.1-8B-Instruct	0.2158	0.1763	18.32%	0.1006	0.0710	29.42%	0.0312	0.0135	56.84%	0.2864	0.2413	15.74%
Qwen2.5-32B-Instruct	0.2198	0.1857	15.50%	0.0608	0.0904	48.68%	0.0192	0.0165	13.89%	0.3330	0.2845	14.55%
Qwen2.5-3B-Instruct	0.2916	0.2636	9.60%	0.2076	0.1827	11.98%	0.0298	0.0213	28.41%	0.3230	0.3073	4.85%
Qwen2.5-72B-Instruct	0.3540	0.3109	12.18%	0.1166	0.1603	37.51%	0.0276	0.0224	18.84%	0.3928	0.3777	3.84%
Qwen2.5-7B-Instruct	0.3170	0.2979	6.01%	0.1968	0.2121	7.76%	0.0474	0.0355	25.18%	0.3184	0.3227	1.36%

ModernBERT, Readability, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0010	0.0013	26.67%	0.0068	0.0049	27.45%	0.0054	0.0032	40.74%	0.0046	0.0027	42.03%
gemma-2-2b-it	0.0852	0.0649	23.87%	0.0316	0.0445	40.93%	0.0166	0.0204	22.89%	0.0660	0.0655	0.71%
gemma-2-9b-it	0.0156	0.0433	177.35%	0.0086	0.0254	195.35%	0.0460	0.0554	20.43%	0.0316	0.0561	77.64%
Llama-3.1-70B-Instruct	0.2854	0.0745	73.91%	0.1008	0.0564	44.05%	0.0208	0.0089	57.05%	0.1214	0.0753	38.00%
Llama-3.1-8B-Instruct	0.2542	0.1536	39.58%	0.1096	0.0707	35.52%	0.0230	0.0092	60.00%	0.1346	0.1112	17.38%
Qwen2.5-32B-Instruct	0.2272	0.1111	51.09%	0.0556	0.0771	38.61%	0.0100	0.0164	64.00%	0.1442	0.1272	11.79%
Qwen2.5-3B-Instruct	0.3164	0.1418	55.18%	0.2158	0.1288	40.32%	0.0298	0.0206	30.87%	0.1588	0.1095	31.02%
Qwen2.5-72B-Instruct	0.4048	0.2793	30.99%	0.1112	0.1393	25.30%	0.0220	0.0211	4.24%	0.2174	0.1915	11.93%
Qwen2.5-7B-Instruct	0.3544	0.2018	43.06%	0.1648	0.1814	10.07%	0.0408	0.0569	39.54%	0.1336	0.1189	10.98%

ModernBERT, Round-trip Translation, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0006	0.0012	100.00%	0.0172	0.0182	5.81%	0.0068	0.0050	26.47%	0.0078	0.0086	10.26%
gemma-2-2b-it	0.0754	0.0616	18.30%	0.0474	0.0328	30.80%	0.0154	0.0124	19.48%	0.0736	0.0500	32.07%
gemma-2-9b-it	0.0098	0.0118	20.41%	0.0186	0.0178	4.30%	0.0416	0.0298	28.37%	0.0574	0.0378	34.15%
Llama-3.1-70B-Instruct	0.2532	0.1396	44.87%	0.1416	0.1120	20.90%	0.0248	0.0202	18.55%	0.1774	0.1250	29.54%
Llama-3.1-8B-Instruct	0.2660	0.1550	41.73%	0.1792	0.1284	28.35%	0.0402	0.0232	42.29%	0.1968	0.1274	35.26%
Qwen2.5-32B-Instruct	0.1398	0.0706	49.50%	0.0820	0.0572	30.24%	0.0126	0.0102	19.05%	0.1686	0.1160	31.20%
Qwen2.5-3B-Instruct	0.2282	0.1198	47.50%	0.1818	0.1096	39.71%	0.0276	0.0182	34.06%	0.2036	0.1344	33.99%
Qwen2.5-72B-Instruct	0.3356	0.1938	42.25%	0.1618	0.0984	39.18%	0.0208	0.0144	30.77%	0.2700	0.2146	20.52%
Qwen2.5-7B-Instruct	0.2756	0.1582	42.60%	0.1676	0.1390	17.06%	0.0418	0.0306	26.79%	0.1522	0.1196	21.42%

ModernBERT, Typos, EM Score

	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.0006	0.0008	33.33%	0.0172	0.0094	45.35%	0.0068	0.0052	23.53%	0.0078	0.0088	12.82%
gemma-2-2b-it	0.0754	0.0642	14.85%	0.0474	0.0408	13.92%	0.0154	0.0156	1.30%	0.0736	0.0542	26.36%
gemma-2-9b-it	0.0098	0.0096	2.04%	0.0186	0.0132	29.03%	0.0416	0.0342	17.79%	0.0574	0.0532	7.32%
Llama-3.1-70B-Instruct	0.2532	0.1634	35.47%	0.1416	0.1006	28.95%	0.0248	0.0196	20.97%	0.1774	0.1358	23.45%
Llama-3.1-8B-Instruct	0.2660	0.1850	30.45%	0.1792	0.1358	24.22%	0.0402	0.0322	19.90%	0.1968	0.1826	7.22%
Qwen2.5-32B-Instruct	0.1398	0.0940	32.76%	0.0820	0.0642	21.71%	0.0126	0.0112	11.11%	0.1686	0.1344	20.28%
Qwen2.5-3B-Instruct	0.2282	0.1568	31.29%	0.1818	0.1448	20.35%	0.0276	0.0268	2.90%	0.2036	0.1758	13.65%
Qwen2.5-72B-Instruct	0.3356	0.2424	27.77%	0.1618	0.1144	29.30%	0.0208	0.0160	23.08%	0.2700	0.2288	15.26%
Qwen2.5-7B-Instruct	0.2756	0.1832	33.53%	0.1676	0.1290	23.03%	0.0418	0.0342	18.18%	0.1522	0.1180	22.47%

Table 16: RAG experiment results with ModernBERT as retrieval model on EM scores.

ModernBERT, Formality, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-2b-it	0.3666	0.2525	31.14%	0.2346	0.1833	21.88%	0.2800	0.2276	18.73%	0.3696	0.2327	37.05%
gemma-2-9b-it	0.3018	0.2642	12.48%	0.2217	0.2007	9.46%	0.2927	0.2429	17.01%	0.3082	0.2511	18.55%
gemma-2-27b-it	0.2965	0.2718	8.33%	0.2273	0.2053	9.67%	0.2746	0.2416	12.03%	0.2574	0.2625	1.96%
Llama-3.1-8B-Instruct	0.4871	0.3223	33.84%	0.3205	0.2005	37.43%	0.2880	0.2396	16.81%	0.4291	0.2985	30.45%
Llama-3.1-70B-Instruct	0.4998	0.2968	40.63%	0.2993	0.1972	34.10%	0.2931	0.2467	15.83%	0.4452	0.2891	35.07%
Qwen2.5-3B-Instruct	0.4911	0.3256	33.71%	0.3366	0.2176	35.34%	0.2789	0.2208	20.84%	0.4476	0.3300	26.27%
Qwen2.5-7B-Instruct	0.5522	0.3675	33.45%	0.3655	0.2679	26.72%	0.3030	0.2440	19.48%	0.4276	0.3098	27.55%
Qwen2.5-32B-Instruct	0.5138	0.3227	37.20%	0.2845	0.2176	23.52%	0.2814	0.2356	16.26%	0.4781	0.3325	30.46%
Qwen2.5-72B-Instruct	0.5824	0.3645	37.42%	0.3298	0.2298	30.33%	0.2848	0.2374	16.64%	0.5187	0.3516	32.22%

ModernBERT, Politeness, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2609	0.2398	8.08%	0.2241	0.2211	1.33%	0.2740	0.2577	5.96%	0.2376	0.2450	3.09%
gemma-2-2b-it	0.3058	0.2442	20.16%	0.2256	0.2057	8.84%	0.2790	0.2551	8.58%	0.3653	0.2425	33.61%
gemma-2-9b-it	0.2625	0.2182	16.86%	0.2150	0.2006	6.70%	0.3068	0.2649	13.66%	0.3005	0.2472	17.74%
Llama-3.1-70B-Instruct	0.3630	0.2654	26.88%	0.2822	0.2108	25.30%	0.2901	0.2616	9.82%	0.4009	0.3021	24.65%
Llama-3.1-8B-Instruct	0.3572	0.3119	12.67%	0.2635	0.2222	15.69%	0.2931	0.2596	11.43%	0.4443	0.3962	10.83%
Qwen2.5-32B-Instruct	0.3720	0.3239	12.95%	0.2478	0.2536	2.34%	0.2747	0.2530	7.90%	0.4738	0.4264	10.01%
Qwen2.5-3B-Instruct	0.3851	0.3427	11.02%	0.3434	0.3042	11.40%	0.2806	0.2531	9.81%	0.4545	0.4304	5.31%
Qwen2.5-72B-Instruct	0.4562	0.4112	9.85%	0.2949	0.3168	7.44%	0.2827	0.2571	9.09%	0.5433	0.5237	3.61%
Qwen2.5-7B-Instruct	0.4284	0.3926	8.37%	0.3486	0.3507	0.59%	0.3010	0.2737	9.06%	0.4653	0.4628	0.53%

ModernBERT, Readability, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2571	0.1885	26.69%	0.2185	0.1850	15.36%	0.2555	0.2220	13.10%	0.2156	0.1820	15.59%
gemma-2-2b-it	0.2911	0.2113	27.41%	0.2097	0.1815	13.45%	0.2595	0.2220	14.44%	0.2660	0.2256	15.20%
gemma-2-9b-it	0.2613	0.2084	20.24%	0.2122	0.1946	8.32%	0.2868	0.2596	9.49%	0.2378	0.2210	7.09%
Llama-3.1-70B-Instruct	0.4085	0.2158	47.16%	0.2710	0.1934	28.63%	0.2673	0.2209	17.38%	0.2904	0.2170	25.28%
Llama-3.1-8B-Instruct	0.3897	0.2701	30.69%	0.2701	0.1983	26.57%	0.2662	0.2148	19.30%	0.3075	0.2561	16.71%
Qwen2.5-32B-Instruct	0.3757	0.2450	34.78%	0.2402	0.2137	11.02%	0.2595	0.2248	13.36%	0.3007	0.2617	12.97%
Qwen2.5-3B-Instruct	0.4077	0.2696	33.89%	0.3384	0.2608	22.93%	0.2664	0.2218	16.74%	0.2895	0.2402	17.02%
Qwen2.5-72B-Instruct	0.4982	0.3719	25.35%	0.2898	0.2733	5.68%	0.2665	0.2322	12.85%	0.3741	0.3225	13.80%
Qwen2.5-7B-Instruct	0.4524	0.3145	30.49%	0.3179	0.3090	2.81%	0.2815	0.2612	7.21%	0.2977	0.2639	11.34%

ModernBERT, Round-trip Translation, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2346	0.1468	37.42%	0.2223	0.1773	20.26%	0.2463	0.2101	14.69%	0.2178	0.1841	15.47%
gemma-2-2b-it	0.2677	0.1798	32.85%	0.2190	0.1626	25.74%	0.2474	0.2075	16.13%	0.2622	0.2088	20.39%
gemma-2-9b-it	0.2319	0.1455	37.25%	0.2156	0.1696	21.36%	0.2714	0.2255	16.93%	0.2552	0.2037	20.20%
Llama-3.1-70B-Instruct	0.3676	0.2230	39.34%	0.3063	0.2401	21.60%	0.2631	0.2195	16.56%	0.3380	0.2552	24.49%
Llama-3.1-8B-Instruct	0.3743	0.2317	38.10%	0.3304	0.2491	24.59%	0.2720	0.2136	21.47%	0.3535	0.2459	30.45%
Qwen2.5-32B-Instruct	0.3013	0.1772	41.17%	0.2589	0.1915	26.01%	0.2480	0.2108	14.98%	0.3233	0.2455	24.07%
Qwen2.5-3B-Instruct	0.3244	0.1936	40.31%	0.3226	0.2174	32.61%	0.2568	0.2062	19.71%	0.3431	0.2576	24.93%
Qwen2.5-72B-Instruct	0.4344	0.2658	38.80%	0.3276	0.2301	29.76%	0.2563	0.2176	15.11%	0.4209	0.3337	20.73%
Qwen2.5-7B-Instruct	0.3830	0.2390	37.59%	0.3266	0.2633	19.38%	0.2757	0.2262	17.93%	0.3189	0.2576	19.23%

ModernBERT, Typos, F1 Score												
	PopQA			Natural Questions			MS MARCO			EntityQuestions		
	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta	Original	Rewritten	Delta
gemma-2-27b-it	0.2346	0.1890	19.43%	0.2223	0.2031	8.64%	0.2463	0.2343	4.84%	0.2178	0.1980	9.09%
gemma-2-2b-it	0.2677	0.2134	20.31%	0.2190	0.1956	10.65%	0.2474	0.2353	4.89%	0.2622	0.2312	11.85%
gemma-2-9b-it	0.2319	0.1833	20.93%	0.2156	0.1956	9.30%	0.2714	0.2552	5.97%	0.2552	0.2316	9.24%
Llama-3.1-70B-Instruct	0.3676	0.2716	26.12%	0.3063	0.2570	16.10%	0.2631	0.2466	6.27%	0.3380	0.2782	17.68%
Llama-3.1-8B-Instruct	0.3743	0.2796	25.32%	0.3304	0.2796	15.38%	0.2720	0.2529	7.03%	0.3535	0.3190	9.76%
Qwen2.5-32B-Instruct	0.3013	0.2270	24.65%	0.2589	0.2282	11.86%	0.2480	0.2357	4.95%	0.3233	0.2756	14.75%
Qwen2.5-3B-Instruct	0.3244	0.2375	26.78%	0.3226	0.2702	16.25%	0.2568	0.2398	6.63%	0.3431	0.2970	13.43%
Qwen2.5-72B-Instruct	0.4344	0.3334	23.23%	0.3276	0.2752	15.99%	0.2563	0.2433	5.06%	0.4209	0.3654	13.20%
Qwen2.5-7B-Instruct	0.3830	0.2768	27.73%	0.3266	0.2767	15.29%	0.2757	0.2575	6.59%	0.3189	0.2654	16.78%

Table 17: RAG experiment results with ModernBERT as retrieval model on F1 scores.

G Full Query Expansion Experiment Results

The following tables provide results of Retrieval and Generation components across MS MARCO and PopQA.

Dataset	Linguistics	R@5	R@10	R@20	R@100
MS MARCO	RTT	37.10	43.74	50.56	63.76
MS MARCO	Typos	37.02	43.66	50.40	63.72
MS MARCO	Readability	29.22	37.34	45.60	61.00
MS MARCO	Formality	28.92	36.88	45.16	60.50
MS MARCO	Politeness	29.44	37.54	45.68	60.52
PopQA	RTT	64.50	69.72	74.18	83.32
PopQA	Typos	64.42	69.60	73.82	83.36
PopQA	Readability	64.30	69.62	74.70	84.28
PopQA	Formality	72.82	77.68	81.30	88.00
PopQA	Politeness	66.74	71.62	76.50	85.94

Table 18: Contriever retrieval results with RAG+HyDE for original queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
MS MARCO	RTT	32.80	38.74	44.46	57.70
MS MARCO	Typos	36.00	42.10	49.36	62.84
MS MARCO	Readability	26.14	34.88	42.74	58.52
MS MARCO	Formality	25.66	33.06	41.38	58.60
MS MARCO	Politeness	28.44	36.28	44.58	59.44
PopQA	RTT	44.62	49.06	53.80	65.72
PopQA	Typos	59.72	64.58	69.50	80.30
PopQA	Readability	59.00	63.96	68.72	79.34
PopQA	Formality	68.60	72.82	76.42	84.20
PopQA	Politeness	65.18	70.06	74.60	83.70

Table 19: Contriever retrieval results with RAG+HyDE for rewritten queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
PopQA	RTT	0.5984	0.6496	0.6980	0.7936
PopQA	Typos	0.5928	0.6444	0.6962	0.7970
PopQA	Readability	0.6176	0.6720	0.7170	0.8164
PopQA	Formality	0.7064	0.7518	0.7938	0.8670
PopQA	Politeness	0.6322	0.6780	0.7272	0.8278
MS MARCO	RTT	0.2994	0.3672	0.4362	0.5546
MS MARCO	Typos	0.3030	0.3698	0.4376	0.5560
MS MARCO	Readability	0.3830	0.4616	0.5414	0.6608
MS MARCO	Formality	0.3816	0.4594	0.5380	0.6520
MS MARCO	Politeness	0.3800	0.4580	0.5314	0.6490

Table 20: ModernBERT retrieval results with RAG+HyDE for original queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
PopQA	RTT	0.3750	0.4230	0.4678	0.5802
PopQA	Typos	0.5388	0.5940	0.6538	0.7584
PopQA	Readability	0.5468	0.5904	0.6394	0.7402
PopQA	Formality	0.6520	0.6970	0.7404	0.8228
PopQA	Politeness	0.6148	0.6640	0.7094	0.8064
MS MARCO	RTT	0.2438	0.3020	0.3694	0.4986
MS MARCO	Typos	0.2876	0.3546	0.4280	0.5438
MS MARCO	Readability	0.3462	0.4326	0.5172	0.6480
MS MARCO	Formality	0.3532	0.4376	0.5122	0.6402
MS MARCO	Politeness	0.3724	0.4510	0.5232	0.6454

Table 21: ModernBERT retrieval results with RAG+HyDE for rewritten queries

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.1466	0.1014	30.83%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.1418	0.1050	25.95%
MS MARCO	RTT	gemma-2-9b-it	0.1366	0.0966	29.28%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.1466	0.1378	6.00%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.1418	0.1328	6.35%
MS MARCO	Typos	gemma-2-9b-it	0.1366	0.1284	6.00%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.1626	0.1430	12.05%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.1556	0.1312	15.68%
MS MARCO	Formality	gemma-2-9b-it	0.1468	0.1296	11.72%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.1866	0.1782	4.50%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.1880	0.1750	6.91%
MS MARCO	politeness	gemma-2-9b-it	0.1742	0.1634	6.20%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.1926	0.1698	11.84%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.1884	0.1786	5.20%
MS MARCO	Readability	gemma-2-9b-it	0.1796	0.1648	8.24%
PopQA	RTT	Llama-3.1-8B-Instruct	0.5194	0.3242	37.58%
PopQA	RTT	Qwen2.5-7B-Instruct	0.5160	0.3260	36.82%
PopQA	RTT	gemma-2-9b-it	0.5266	0.3340	36.57%
PopQA	Typos	Llama-3.1-8B-Instruct	0.5194	0.4698	9.55%
PopQA	Typos	Qwen2.5-7B-Instruct	0.5160	0.4710	8.72%
PopQA	Typos	gemma-2-9b-it	0.5266	0.4842	8.05%
PopQA	Formality	Llama-3.1-8B-Instruct	0.6400	0.5984	6.50%
PopQA	Formality	Qwen2.5-7B-Instruct	0.6300	0.5908	6.22%
PopQA	Formality	gemma-2-9b-it	0.6348	0.5954	6.21%
PopQA	politeness	Llama-3.1-8B-Instruct	0.5564	0.5384	3.24%
PopQA	politeness	Qwen2.5-7B-Instruct	0.5364	0.5214	2.80%
PopQA	politeness	gemma-2-9b-it	0.5464	0.5198	4.87%
PopQA	Readability	Llama-3.1-8B-Instruct	0.5558	0.5086	8.49%
PopQA	Readability	Qwen2.5-7B-Instruct	0.5492	0.4840	11.87%
PopQA	Readability	gemma-2-9b-it	0.5524	0.4970	10.03%

Table 22: RAG experiment results with query expansion and Contriever as retrieval model on AM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.0318	0.0166	47.80%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.0358	0.0244	31.84%
MS MARCO	RTT	gemma-2-9b-it	0.0350	0.0254	27.43%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.0318	0.0286	10.06%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.0358	0.0294	17.88%
MS MARCO	Typos	gemma-2-9b-it	0.0350	0.0314	10.29%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.0234	0.0098	58.12%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.0356	0.0240	32.58%
MS MARCO	Formality	gemma-2-9b-it	0.0324	0.0152	53.09%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.0286	0.0104	63.64%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.0434	0.0328	24.42%
MS MARCO	politeness	gemma-2-9b-it	0.0462	0.0132	71.43%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.0182	0.0094	48.35%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.0374	0.0558	49.20%
MS MARCO	Readability	gemma-2-9b-it	0.0428	0.0554	29.44%
PopQA	RTT	Llama-3.1-8B-Instruct	0.2926	0.1594	45.52%
PopQA	RTT	Qwen2.5-7B-Instruct	0.2886	0.1628	43.59%
PopQA	RTT	gemma-2-9b-it	0.0076	0.0066	13.16%
PopQA	Typos	Llama-3.1-8B-Instruct	0.2926	0.2406	17.77%
PopQA	Typos	Qwen2.5-7B-Instruct	0.2886	0.2368	17.95%
PopQA	Typos	gemma-2-9b-it	0.0076	0.0068	10.53%
PopQA	Formality	Llama-3.1-8B-Instruct	0.3792	0.2488	34.39%
PopQA	Formality	Qwen2.5-7B-Instruct	0.4646	0.3114	32.97%
PopQA	Formality	gemma-2-9b-it	0.0202	0.0124	38.61%
PopQA	politeness	Llama-3.1-8B-Instruct	0.2424	0.1970	18.73%
PopQA	politeness	Qwen2.5-7B-Instruct	0.3188	0.3166	0.69%
PopQA	politeness	gemma-2-9b-it	0.0170	0.0058	65.88%
PopQA	Readability	Llama-3.1-8B-Instruct	0.2868	0.1770	38.28%
PopQA	Readability	Qwen2.5-7B-Instruct	0.3780	0.2338	38.15%
PopQA	Readability	gemma-2-9b-it	0.0114	0.0590	417.54%

Table 23: RAG experiment results with query expansion and Contriever as retrieval model on EM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.2412	0.1894	21.49%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.2453	0.2037	16.93%
MS MARCO	RTT	gemma-2-9b-it	0.2458	0.1995	18.81%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.2412	0.2323	3.71%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.2453	0.2357	3.89%
MS MARCO	Typos	gemma-2-9b-it	0.2458	0.2381	3.12%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.2593	0.2270	12.44%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.2680	0.2354	12.14%
MS MARCO	Formality	gemma-2-9b-it	0.2685	0.2353	12.39%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.2591	0.2323	10.34%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.2741	0.2516	8.21%
MS MARCO	politeness	gemma-2-9b-it	0.2790	0.2426	13.05%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2401	0.1955	18.56%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2618	0.2476	5.42%
MS MARCO	Readability	gemma-2-9b-it	0.2647	0.2431	8.19%
PopQA	RTT	Llama-3.1-8B-Instruct	0.3983	0.2358	40.79%
PopQA	RTT	Qwen2.5-7B-Instruct	0.4005	0.2398	40.12%
PopQA	RTT	gemma-2-9b-it	0.2372	0.1391	41.37%
PopQA	Typos	Llama-3.1-8B-Instruct	0.3983	0.3483	12.55%
PopQA	Typos	Qwen2.5-7B-Instruct	0.4005	0.3508	12.41%
PopQA	Typos	gemma-2-9b-it	0.2372	0.2174	8.32%
PopQA	Formality	Llama-3.1-8B-Instruct	0.5024	0.3968	21.02%
PopQA	Formality	Qwen2.5-7B-Instruct	0.5602	0.4461	20.36%
PopQA	Formality	gemma-2-9b-it	0.2958	0.2812	4.94%
PopQA	politeness	Llama-3.1-8B-Instruct	0.3779	0.3398	10.07%
PopQA	politeness	Qwen2.5-7B-Instruct	0.4316	0.4259	1.34%
PopQA	politeness	gemma-2-9b-it	0.2605	0.2277	12.59%
PopQA	Readability	Llama-3.1-8B-Instruct	0.4127	0.2956	28.37%
PopQA	Readability	Qwen2.5-7B-Instruct	0.4781	0.3552	25.70%
PopQA	Readability	gemma-2-9b-it	0.2633	0.2344	10.98%

Table 24: RAG experiment results with query expansion and Contriever as retrieval model on F1 scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.1706	0.1222	28.37%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.1634	0.1216	25.58%
MS MARCO	RTT	gemma-2-9b-it	0.1560	0.1138	27.05%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.1746	0.1642	5.96%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.1592	0.1522	4.40%
MS MARCO	Typos	gemma-2-9b-it	0.1578	0.1462	7.35%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.1896	0.1710	9.81%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.1778	0.1568	11.81%
MS MARCO	Formality	gemma-2-9b-it	0.1620	0.1448	10.62%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.2174	0.2068	4.88%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.2042	0.1956	4.21%
MS MARCO	politeness	gemma-2-9b-it	0.1910	0.1828	4.29%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2126	0.1896	10.82%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2096	0.1908	8.97%
MS MARCO	Readability	gemma-2-9b-it	0.1924	0.1792	6.86%
PopQA	RTT	Llama-3.1-8B-Instruct	0.4892	0.2890	40.92%
PopQA	RTT	Qwen2.5-7B-Instruct	0.4858	0.2982	38.62%
PopQA	RTT	gemma-2-9b-it	0.4982	0.3002	39.74%
PopQA	Typos	Llama-3.1-8B-Instruct	0.4910	0.4394	10.51%
PopQA	Typos	Qwen2.5-7B-Instruct	0.4838	0.4356	9.96%
PopQA	Typos	gemma-2-9b-it	0.4984	0.4524	9.23%
PopQA	Formality	Llama-3.1-8B-Instruct	0.6202	0.5764	7.06%
PopQA	Formality	Qwen2.5-7B-Instruct	0.6154	0.5718	7.08%
PopQA	Formality	gemma-2-9b-it	0.6168	0.5720	7.26%
PopQA	politeness	Llama-3.1-8B-Instruct	0.5418	0.5156	4.84%
PopQA	politeness	Qwen2.5-7B-Instruct	0.5242	0.5008	4.46%
PopQA	politeness	gemma-2-9b-it	0.5314	0.5026	5.42%
PopQA	Readability	Llama-3.1-8B-Instruct	0.5366	0.4662	13.12%
PopQA	Readability	Qwen2.5-7B-Instruct	0.5226	0.4448	14.89%
PopQA	Readability	gemma-2-9b-it	0.5306	0.4578	13.72%

Table 25: RAG experiment results with query expansion and ModernBERT as retrieval model on AM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.0292	0.0218	25.34%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.0380	0.0348	8.42%
MS MARCO	RTT	gemma-2-9b-it	0.0392	0.0270	31.12%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.0358	0.0316	11.73%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.0416	0.0322	22.60%
MS MARCO	Typos	gemma-2-9b-it	0.0400	0.0334	16.50%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.0288	0.0142	50.69%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.0396	0.0322	18.69%
MS MARCO	Formality	gemma-2-9b-it	0.0362	0.0154	57.46%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.0278	0.0116	58.27%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.0454	0.0400	11.89%
MS MARCO	politeness	gemma-2-9b-it	0.0504	0.0144	71.43%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.0210	0.0092	56.19%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.0422	0.0608	44.08%
MS MARCO	Readability	gemma-2-9b-it	0.0486	0.0598	23.05%
PopQA	RTT	Llama-3.1-8B-Instruct	0.2826	0.1408	50.18%
PopQA	RTT	Qwen2.5-7B-Instruct	0.2866	0.1430	50.10%
PopQA	RTT	gemma-2-9b-it	0.0082	0.0066	19.51%
PopQA	Typos	Llama-3.1-8B-Instruct	0.2768	0.2328	15.90%
PopQA	Typos	Qwen2.5-7B-Instruct	0.2758	0.2194	20.45%
PopQA	Typos	gemma-2-9b-it	0.0066	0.0086	30.30%
PopQA	Formality	Llama-3.1-8B-Instruct	0.3728	0.2400	35.62%
PopQA	Formality	Qwen2.5-7B-Instruct	0.4604	0.3054	33.67%
PopQA	Formality	gemma-2-9b-it	0.0162	0.0118	27.16%
PopQA	politeness	Llama-3.1-8B-Instruct	0.2424	0.1956	19.31%
PopQA	politeness	Qwen2.5-7B-Instruct	0.3240	0.3130	3.40%
PopQA	politeness	gemma-2-9b-it	0.0166	0.0062	62.65%
PopQA	Readability	Llama-3.1-8B-Instruct	0.2726	0.1550	43.14%
PopQA	Readability	Qwen2.5-7B-Instruct	0.3552	0.2126	40.15%
PopQA	Readability	gemma-2-9b-it	0.0122	0.0544	345.90%

Table 26: RAG experiment results with query expansion and ModernBERT as retrieval model on EM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.2593	0.2080	19.76%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.2697	0.2247	16.69%
MS MARCO	RTT	gemma-2-9b-it	0.2669	0.2176	18.48%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.2667	0.2544	4.62%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.2717	0.2583	4.95%
MS MARCO	Typos	gemma-2-9b-it	0.2677	0.2540	5.12%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.2855	0.2559	10.36%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.2939	0.2657	9.59%
MS MARCO	Formality	gemma-2-9b-it	0.2877	0.2528	12.14%
MS MARCO	politeness	Llama-3.1-8B-Instruct	0.2818	0.2558	9.22%
MS MARCO	politeness	Qwen2.5-7B-Instruct	0.2967	0.2745	7.49%
MS MARCO	politeness	gemma-2-9b-it	0.2966	0.2608	12.07%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2613	0.2135	18.29%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2825	0.2686	4.95%
MS MARCO	Readability	gemma-2-9b-it	0.2847	0.2634	7.47%
PopQA	RTT	Llama-3.1-8B-Instruct	0.3849	0.2103	45.37%
PopQA	RTT	Qwen2.5-7B-Instruct	0.3903	0.2185	44.01%
PopQA	RTT	gemma-2-9b-it	0.2275	0.1272	44.06%
PopQA	Typos	Llama-3.1-8B-Instruct	0.3779	0.3289	12.98%
PopQA	Typos	Qwen2.5-7B-Instruct	0.3798	0.3266	14.02%
PopQA	Typos	gemma-2-9b-it	0.2255	0.2075	8.02%
PopQA	Formality	Llama-3.1-8B-Instruct	0.4943	0.3861	21.89%
PopQA	Formality	Qwen2.5-7B-Instruct	0.5529	0.4344	21.45%
PopQA	Formality	gemma-2-9b-it	0.2870	0.2733	4.77%
PopQA	politeness	Llama-3.1-8B-Instruct	0.3719	0.3324	10.61%
PopQA	politeness	Qwen2.5-7B-Instruct	0.4301	0.4136	3.83%
PopQA	politeness	gemma-2-9b-it	0.2553	0.2230	12.65%
PopQA	Readability	Llama-3.1-8B-Instruct	0.3974	0.2690	32.29%
PopQA	Readability	Qwen2.5-7B-Instruct	0.4523	0.3272	27.66%
PopQA	Readability	gemma-2-9b-it	0.2557	0.2179	14.79%

Table 27: RAG experiment results with query expansion and ModernBERT as retrieval model on F1 scores.

H Full Reranking Experiment Results

Dataset	Linguistics	R@5	R@10	R@20	R@100
MS MARCO	RTT	0.4620	0.5208	0.5618	0.6102
MS MARCO	Typos	0.4620	0.5208	0.5618	0.6102
MS MARCO	Readability	0.4058	0.4688	0.5188	0.5718
MS MARCO	Formality	0.4070	0.4714	0.5154	0.5598
MS MARCO	Politeness	0.4078	0.4706	0.5146	0.5664
PopQA	RTT	0.6640	0.7230	0.7674	0.8192
PopQA	Typos	0.6640	0.7230	0.7674	0.8192
PopQA	Readability	0.6864	0.7506	0.7870	0.8344
PopQA	Formality	0.7620	0.8100	0.8404	0.8760
PopQA	Politeness	0.7050	0.7722	0.8096	0.8534

Table 28: Contriever retrieval results with RAG+Reranking for original queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
MS MARCO	RTT	0.4144	0.4674	0.5158	0.5754
MS MARCO	Typos	0.4106	0.4598	0.4932	0.5440
MS MARCO	Readability	0.3061	0.3773	0.4420	0.5235
MS MARCO	Formality	0.2755	0.3341	0.3803	0.4365
MS MARCO	Politeness	0.3665	0.4304	0.4751	0.5448
PopQA	RTT	0.4900	0.5510	0.5946	0.6654
PopQA	Typos	0.5420	0.5888	0.6238	0.6724
PopQA	Readability	0.5839	0.6421	0.6835	0.7332
PopQA	Formality	0.6837	0.7266	0.7543	0.7856
PopQA	Politeness	0.6655	0.7300	0.7701	0.8081

Table 29: Contriever retrieval results with RAG+Reranking for rewritten queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
PopQA	RTT	0.6836	0.7462	0.7934	0.8344
PopQA	Typos	0.6836	0.7462	0.7934	0.8344
PopQA	Readability	0.7012	0.7620	0.8024	0.8432
PopQA	Formality	0.7684	0.8162	0.8502	0.8832
PopQA	Politeness	0.7282	0.7878	0.8310	0.8688
MS MARCO	RTT	0.3628	0.4302	0.4860	0.5680
MS MARCO	Typos	0.3628	0.4302	0.4860	0.5680
MS MARCO	Readability	0.4428	0.5194	0.5876	0.6746
MS MARCO	Formality	0.4482	0.5284	0.5932	0.6720
MS MARCO	Politeness	0.4478	0.5222	0.5802	0.6638

Table 30: ModernBERT retrieval results with RAG+Reranking for original queries

Dataset	Linguistics	R@5	R@10	R@20	R@100
PopQA	RTT	0.5096	0.5744	0.6204	0.6856
PopQA	Typos	0.5940	0.6496	0.6886	0.7406
PopQA	Readability	0.6237	0.6836	0.7289	0.7796
PopQA	Formality	0.7337	0.7809	0.8113	0.8493
PopQA	Politeness	0.6888	0.7487	0.7860	0.8317
MS MARCO	RTT	0.2940	0.3558	0.4160	0.5146
MS MARCO	Typos	0.3248	0.3902	0.4548	0.5396
MS MARCO	Readability	0.3279	0.4179	0.4994	0.6461
MS MARCO	Formality	0.3479	0.4392	0.5224	0.6479
MS MARCO	Politeness	0.4060	0.4882	0.5553	0.6579

Table 31: ModernBERT retrieval results with RAG+Reranking for rewritten queries

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.1428	0.1302	8.82%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.1416	0.1290	8.90%
MS MARCO	RTT	gemma-2-9b-it	0.1292	0.1202	6.97%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.1428	0.1118	21.71%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.1416	0.1120	20.90%
MS MARCO	Typos	gemma-2-9b-it	0.1292	0.1066	17.49%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.1552	0.1127	27.36%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.1540	0.1100	28.57%
MS MARCO	Formality	gemma-2-9b-it	0.1430	0.0979	31.56%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.1876	0.1586	15.46%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.1820	0.1573	13.55%
MS MARCO	Politeness	gemma-2-9b-it	0.1730	0.1489	13.95%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.1820	0.1528	16.04%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.1780	0.1517	14.76%
MS MARCO	Readability	gemma-2-9b-it	0.1630	0.1449	11.12%
PopQA	RTT	Llama-3.1-8B-Instruct	0.5184	0.3366	35.07%
PopQA	RTT	Qwen2.5-7B-Instruct	0.4924	0.3350	31.97%
PopQA	RTT	gemma-2-9b-it	0.5086	0.3504	31.10%
PopQA	Typos	Llama-3.1-8B-Instruct	0.5184	0.3490	32.68%
PopQA	Typos	Qwen2.5-7B-Instruct	0.4924	0.3352	31.93%
PopQA	Typos	gemma-2-9b-it	0.5086	0.3660	28.04%
PopQA	Formality	Llama-3.1-8B-Instruct	0.6554	0.5099	22.20%
PopQA	Formality	Qwen2.5-7B-Instruct	0.6260	0.4861	22.35%
PopQA	Formality	gemma-2-9b-it	0.6294	0.4999	20.58%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.5144	0.4811	6.48%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.4818	0.4540	5.77%
PopQA	Politeness	gemma-2-9b-it	0.4978	0.4685	5.88%
PopQA	Readability	Llama-3.1-8B-Instruct	0.5142	0.4158	19.14%
PopQA	Readability	Qwen2.5-7B-Instruct	0.4870	0.3874	20.45%
PopQA	Readability	gemma-2-9b-it	0.5028	0.4091	18.63%

Table 32: Generation results with Contriever retrieval and reranking on AM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.0288	0.0214	25.69%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.0336	0.0302	10.12%
MS MARCO	RTT	gemma-2-9b-it	0.0388	0.0288	25.77%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.0288	0.0174	39.58%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.0336	0.0254	24.40%
MS MARCO	Typos	gemma-2-9b-it	0.0388	0.0308	20.62%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.0232	0.0059	74.43%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.0362	0.0149	58.93%
MS MARCO	Formality	gemma-2-9b-it	0.0344	0.0090	73.84%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.0248	0.0079	68.28%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.0384	0.0265	30.90%
MS MARCO	Politeness	gemma-2-9b-it	0.0512	0.0153	70.18%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.0166	0.0076	54.22%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.0330	0.0441	33.74%
MS MARCO	Readability	gemma-2-9b-it	0.0452	0.0521	15.34%
PopQA	RTT	Llama-3.1-8B-Instruct	0.2616	0.1454	44.42%
PopQA	RTT	Qwen2.5-7B-Instruct	0.2646	0.1646	37.79%
PopQA	RTT	gemma-2-9b-it	0.0092	0.0090	2.17%
PopQA	Typos	Llama-3.1-8B-Instruct	0.2616	0.1660	36.54%
PopQA	Typos	Qwen2.5-7B-Instruct	0.2646	0.1536	41.95%
PopQA	Typos	gemma-2-9b-it	0.0092	0.0094	2.17%
PopQA	Formality	Llama-3.1-8B-Instruct	0.3452	0.1375	60.18%
PopQA	Formality	Qwen2.5-7B-Instruct	0.4498	0.1851	58.86%
PopQA	Formality	gemma-2-9b-it	0.0218	0.0068	68.81%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.2016	0.1649	18.22%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.2734	0.2725	0.32%
PopQA	Politeness	gemma-2-9b-it	0.0166	0.0063	62.25%
PopQA	Readability	Llama-3.1-8B-Instruct	0.2460	0.1547	37.13%
PopQA	Readability	Qwen2.5-7B-Instruct	0.3212	0.1873	41.70%
PopQA	Readability	gemma-2-9b-it	0.0134	0.0383	185.57%

Table 33: Generation results with Contriever retrieval and reranking on EM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.2310	0.2146	7.12%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.2402	0.2272	5.42%
MS MARCO	RTT	gemma-2-9b-it	0.2394	0.2211	7.64%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.2310	0.2015	12.76%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.2402	0.2090	12.99%
MS MARCO	Typos	gemma-2-9b-it	0.2394	0.2124	11.26%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.2493	0.1913	23.26%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.2604	0.1984	23.81%
MS MARCO	Formality	gemma-2-9b-it	0.2621	0.1990	24.06%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.2494	0.2142	14.11%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.2605	0.2303	11.58%
MS MARCO	Politeness	gemma-2-9b-it	0.2729	0.2272	16.74%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2303	0.1816	21.15%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2491	0.2216	11.05%
MS MARCO	Readability	gemma-2-9b-it	0.2564	0.2255	12.04%
PopQA	RTT	Llama-3.1-8B-Instruct	0.3757	0.2295	38.90%
PopQA	RTT	Qwen2.5-7B-Instruct	0.3730	0.2440	34.57%
PopQA	RTT	gemma-2-9b-it	0.2291	0.1488	35.02%
PopQA	Typos	Llama-3.1-8B-Instruct	0.3757	0.2501	33.43%
PopQA	Typos	Qwen2.5-7B-Instruct	0.3730	0.2395	35.80%
PopQA	Typos	gemma-2-9b-it	0.2291	0.1606	29.88%
PopQA	Formality	Llama-3.1-8B-Instruct	0.4826	0.2859	40.75%
PopQA	Formality	Qwen2.5-7B-Instruct	0.5492	0.3202	41.69%
PopQA	Formality	gemma-2-9b-it	0.2936	0.2278	22.42%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.3298	0.2906	11.90%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.3804	0.3633	4.50%
PopQA	Politeness	gemma-2-9b-it	0.2345	0.2032	13.38%
PopQA	Readability	Llama-3.1-8B-Instruct	0.3691	0.2563	30.56%
PopQA	Readability	Qwen2.5-7B-Instruct	0.4162	0.2883	30.74%
PopQA	Readability	gemma-2-9b-it	0.2392	0.1888	21.06%

Table 34: Generation results with Contriever retrieval and reranking on F1 scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.1964	0.1390	29.23%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.1882	0.1354	28.06%
MS MARCO	RTT	gemma-2-9b-it	0.1690	0.1246	26.27%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.1964	0.1772	9.78%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.1882	0.1678	10.84%
MS MARCO	Typos	gemma-2-9b-it	0.1690	0.1618	4.26%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.2118	0.1655	21.84%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.2012	0.1531	23.89%
MS MARCO	Formality	gemma-2-9b-it	0.1770	0.1388	21.58%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.2394	0.2134	10.86%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.2316	0.2067	10.77%
MS MARCO	Politeness	gemma-2-9b-it	0.2160	0.1943	10.06%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2324	0.1791	22.92%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2254	0.1853	17.78%
MS MARCO	Readability	gemma-2-9b-it	0.2082	0.1702	18.25%
PopQA	RTT	Llama-3.1-8B-Instruct	0.5340	0.3540	33.71%
PopQA	RTT	Qwen2.5-7B-Instruct	0.5130	0.3476	32.24%
PopQA	RTT	gemma-2-9b-it	0.5264	0.3640	30.85%
PopQA	Typos	Llama-3.1-8B-Instruct	0.5340	0.4546	14.87%
PopQA	Typos	Qwen2.5-7B-Instruct	0.5130	0.4348	15.24%
PopQA	Typos	gemma-2-9b-it	0.5264	0.4648	11.70%
PopQA	Formality	Llama-3.1-8B-Instruct	0.6572	0.6163	6.22%
PopQA	Formality	Qwen2.5-7B-Instruct	0.6238	0.5969	4.32%
PopQA	Formality	gemma-2-9b-it	0.6336	0.5979	5.63%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.5686	0.5427	4.55%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.5400	0.5093	5.68%
PopQA	Politeness	gemma-2-9b-it	0.5484	0.5187	5.41%
PopQA	Readability	Llama-3.1-8B-Instruct	0.5694	0.4929	13.44%
PopQA	Readability	Qwen2.5-7B-Instruct	0.5390	0.4661	13.53%
PopQA	Readability	gemma-2-9b-it	0.5532	0.4800	13.23%

Table 35: Generation results with ModernBERT retrieval and reranking on AM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.0446	0.0248	44.39%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.0472	0.0370	21.61%
MS MARCO	RTT	gemma-2-9b-it	0.0428	0.0326	23.83%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.0446	0.0342	23.32%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.0472	0.0376	20.34%
MS MARCO	Typos	gemma-2-9b-it	0.0428	0.0352	17.76%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.0332	0.0105	68.27%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.0496	0.0240	51.61%
MS MARCO	Formality	gemma-2-9b-it	0.0366	0.0087	76.14%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.0348	0.0131	62.45%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.0572	0.0371	35.20%
MS MARCO	Politeness	gemma-2-9b-it	0.0550	0.0141	74.30%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.0214	0.0091	57.63%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.0456	0.0566	24.12%
MS MARCO	Readability	gemma-2-9b-it	0.0492	0.0559	13.69%
PopQA	RTT	Llama-3.1-8B-Instruct	0.2626	0.1582	39.76%
PopQA	RTT	Qwen2.5-7B-Instruct	0.2794	0.1710	38.80%
PopQA	RTT	gemma-2-9b-it	0.0120	0.0102	15.00%
PopQA	Typos	Llama-3.1-8B-Instruct	0.2626	0.2084	20.64%
PopQA	Typos	Qwen2.5-7B-Instruct	0.2794	0.2166	22.48%
PopQA	Typos	gemma-2-9b-it	0.0120	0.0122	1.67%
PopQA	Formality	Llama-3.1-8B-Instruct	0.3422	0.1481	56.71%
PopQA	Formality	Qwen2.5-7B-Instruct	0.4472	0.2155	51.82%
PopQA	Formality	gemma-2-9b-it	0.0244	0.0073	69.95%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.2006	0.1761	12.23%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.3020	0.3095	2.49%
PopQA	Politeness	gemma-2-9b-it	0.0216	0.0083	61.73%
PopQA	Readability	Llama-3.1-8B-Instruct	0.2526	0.1645	34.89%
PopQA	Readability	Qwen2.5-7B-Instruct	0.3580	0.2152	39.89%
PopQA	Readability	gemma-2-9b-it	0.0182	0.0479	163.37%

Table 36: Generation results with ModernBERT retrieval and reranking on EM scores.

Dataset	Linguistics	LLMs	Original	Rewritten	Delta
MS MARCO	RTT	Llama-3.1-8B-Instruct	0.2837	0.2203	22.35%
MS MARCO	RTT	Qwen2.5-7B-Instruct	0.2866	0.2349	18.04%
MS MARCO	RTT	gemma-2-9b-it	0.2797	0.2308	17.49%
MS MARCO	Typos	Llama-3.1-8B-Instruct	0.2837	0.2622	7.56%
MS MARCO	Typos	Qwen2.5-7B-Instruct	0.2866	0.2672	6.78%
MS MARCO	Typos	gemma-2-9b-it	0.2797	0.2638	5.68%
MS MARCO	Formality	Llama-3.1-8B-Instruct	0.3004	0.2395	20.25%
MS MARCO	Formality	Qwen2.5-7B-Instruct	0.3094	0.2477	19.96%
MS MARCO	Formality	gemma-2-9b-it	0.3011	0.2440	18.95%
MS MARCO	Politeness	Llama-3.1-8B-Instruct	0.3011	0.2602	13.58%
MS MARCO	Politeness	Qwen2.5-7B-Instruct	0.3139	0.2778	11.51%
MS MARCO	Politeness	gemma-2-9b-it	0.3173	0.2664	16.02%
MS MARCO	Readability	Llama-3.1-8B-Instruct	0.2736	0.2099	23.27%
MS MARCO	Readability	Qwen2.5-7B-Instruct	0.2925	0.2595	11.28%
MS MARCO	Readability	gemma-2-9b-it	0.2935	0.2581	12.04%
PopQA	RTT	Llama-3.1-8B-Instruct	0.3850	0.2438	36.67%
PopQA	RTT	Qwen2.5-7B-Instruct	0.3903	0.2531	35.15%
PopQA	RTT	gemma-2-9b-it	0.2396	0.1546	35.49%
PopQA	Typos	Llama-3.1-8B-Instruct	0.3850	0.3173	17.57%
PopQA	Typos	Qwen2.5-7B-Instruct	0.3903	0.3175	18.65%
PopQA	Typos	gemma-2-9b-it	0.2396	0.2090	12.76%
PopQA	Formality	Llama-3.1-8B-Instruct	0.4808	0.3310	31.16%
PopQA	Formality	Qwen2.5-7B-Instruct	0.5468	0.3868	29.27%
PopQA	Formality	gemma-2-9b-it	0.2973	0.2773	6.73%
PopQA	Politeness	Llama-3.1-8B-Instruct	0.3466	0.3192	7.91%
PopQA	Politeness	Qwen2.5-7B-Instruct	0.4185	0.4083	2.44%
PopQA	Politeness	gemma-2-9b-it	0.2617	0.2258	13.69%
PopQA	Readability	Llama-3.1-8B-Instruct	0.3906	0.2886	26.12%
PopQA	Readability	Qwen2.5-7B-Instruct	0.4571	0.3353	26.64%
PopQA	Readability	gemma-2-9b-it	0.2671	0.2240	16.13%

Table 37: Generation results with ModernBERT retrieval and reranking on F1 scores.

I Rewriting Prompts

I.1 Formality

Prompt 1:

You are an AI assistant skilled at transforming formal queries into casual, everyday language. Rewrite the following query so that it sounds very informal. Experiment with different colloquial openings, varied sentence constructions, and a mix of slang, idioms, and casual expressions throughout the sentence. Avoid using the same phrase repeatedly (e.g., "hey, so like") and ensure the meaning remains unchanged.

Prompt 2:

Your task is to convert the given query into an informal version that feels natural and conversational. Instead of a uniform introductory phrase, use a range of informal expressions (such as interjections, casual questions, or slang) at different parts of the sentence. Mix up the structure—sometimes start with an interjection, other times rephrase the sentence completely—while keeping the original meaning intact.

Prompt 3:

Task: Transform Formal to Extremely Informal Language

Convert the following formal sentence into an extremely informal, messy, and natural version. The output should sound like authentic, real-world casual speech—as if spoken in an informal chat, online conversation, or street talk.

Critical Rules: - Diversity is key: No two sentences should follow the same pattern. - Break formal sentence structures: Chop up long phrases, reorder words, or make them flow casually. - Use a wide range of informality techniques—DO NOT rely on just contractions or slang. - Avoid starting every sentence with the same phrase or discourse marker. - Embrace randomness: Let the outputs sound wild, real, and unpredictable. - Do not start your sentences with "Yo", "Hey so like", etc.

Final Execution Instruction: Generate an informal version of the following sentence that: - Uses multiple different informality techniques. - Avoids repetitive sentence structures or patterns. - Sounds raw, conversational, and unpredictable.

Original Query:

I.2 Readablility

Prompt 1:

1. Task Definition:

You are rewriting a query to make it significantly less readable while preserving the original semantic meaning as closely as possible.

2. Constraints & Goals:

- Flesch Reading Ease Score: The rewritten text must have a Flesch score below 60 (preferably below 50).

- Semantic Similarity: The rewritten text must have SBERT similarity > 0.7 compared with the original query.

- Length: The rewritten text must remain approximately the same length as the original query (±10%).

- Preserve Domain Terminology: Do not remove or drastically change domain-specific words, abbreviations, or technical terms (e.g., "IRS," "distance," etc.).

- Abbreviation: Do not expand abbreviations unless the original query already used the expanded form.

- No New Information: You must not add additional details beyond what the original query states.

- Question Format: Retain the form of a question if the original is posed as a question.

3. How to Increase Complexity:

- Lexical Changes: Use advanced or academic synonyms only for common words. For domain or key terms (e.g., "distance," "IRS," "tax"), keep the original term or use a very close synonym if necessary to maintain meaning.

- Syntactic Complexity: Introduce passive voice, nominalizations, embedded clauses, and parenthetical or subordinate phrases. Ensure the sentence flow is more formal and convoluted without changing the core meaning.

- Redundancy & Formality: Employ circumlocution and excessively formal expressions (e.g., "due to the fact that" instead of "because") while avoiding any semantic drift.

- Dense, Indirect Construction: Favor longer phrases, indirect references, and wordiness. Avoid direct or simple phrasing.

Original Query: ...

Less Readable Query:

Prompt 2:

Task Description:

Transform a given query into a significantly less readable version while preserving its original semantic meaning as closely as possible.

Constraints & Goals:

- **Readability:** The rewritten text must have a **Flesch Reading Ease Score below 60**, preferably below 50.

- **Semantic Similarity:** The rewritten text must achieve an **SBERT similarity score > 0.7** with the original query.

- **Length Consistency:** The modified text should be **within ±10% of the original length**.

- **Preserve Key Terminology:** **Do not alter domain-specific words, abbreviations, or technical jargon** (e.g., "IRS," "distance").

- **Abbreviation Handling:** **Do not expand abbreviations** unless they are already expanded in the original query.

- **Maintain Original Intent:** Do not add, remove, or alter the factual content of the query.

- **Retain Question Structure:** If the input is a question, the output must also be a question.

Techniques to Decrease Readability:

1. **Lexical Complexity:** Replace common words with **advanced, academic, or formal synonyms**, while keeping domain-specific terms unchanged.

2. **Syntactic Complexity:** Introduce **passive voice, nominalizations, embedded clauses, or subordinate structures** to increase sentence density.

3. **Redundancy & Formality:** Use **circumlocution, excessive formality, and indirect phrasing** (e.g., "in light of the fact that" instead of "because").

4. **Dense Sentence Structure:** Prefer **wordy, indirect, and convoluted constructions** over direct phrasing.

Original Query: ...

Less Readable Query:

Prompt 3:

Objective:

You are tasked with **rewriting a given query to make it significantly less readable** while preserving its original semantic meaning with high fidelity.

Guiding Principles:

- **Readability Constraint:** The rewritten text must have a **Flesch Reading Ease Score of <= 60**, preferably <= 50.

- **Semantic Integrity:** Ensure an **SBERT similarity score of at least 0.7** between the original and rewritten text.

- **Length Tolerance:** Maintain an **approximate length deviation of no more than ±10%** from the original.

- **Terminology Preservation:** Domain-specific terms (e.g., "IRS," "distance") **must remain intact** or be substituted only with **near-synonymous equivalents**.

- **Abbreviation Handling:** If an abbreviation exists, **retain it as is** unless the original query explicitly expands it.

- **Strict Content Preservation:** Do **not introduce any new information** or omit existing details.

- **Question Retention:** If the input is a question, the reformulated output **must remain a question**.

Techniques for Readability Reduction:

- **Lexical Sophistication:** Replace commonplace words with **more complex, formal, or technical alternatives** while maintaining clarity of meaning.

- **Structural Density:** Employ **passive constructions, embedded clauses, and nominalized phrases** to increase syntactic complexity.

- **Circumlocution & Wordiness:** Favor **verbose, indirect expressions** over concise phrasing (e.g., "with regard to" instead of "about").

- **Elaborate Phrasing:** Use **multi-clause structures and intricate sentence formations** to reduce direct readability.

Original Query: ...

Less Readable Query:

I.3 Politeness

Prompt 1:

Task: Rewrite Queries to Sound More Polite and Courteous

Rephrase the given query into a more polite, respectful, and considerate version while preserving its original intent. The output should reflect a natural, well-mannered tone suitable for professional or friendly interactions. The generated query should be a single sentence.

Critical Rules:

- Use a variety of politeness techniques, including warm greetings, indirect requests, and expressions of gratitude.

- Avoid robotic or overly formal constructions—make it sound naturally courteous, warm and friendly.

- Do not always start your sentence with 'Could you please tell'. Use emotional undertones and specific attempts at politeness.

- Maintain the original meaning without unnecessary embellishment.

- Do not start the generated query with 'I hope you are ...' or end with a single 'Thank you' sentence. Generate only a single polite query sentence.

Original Query: ...

Polite Query:

Prompt 2:

Task: Enhance the Courtesy of a Given Query

Transform the provided query into a more respectful, friendly, and warm version, ensuring it conveys respect and warmth while keeping the original intent intact. The reworded request should sound engaging, professional, and well-mannered. The generated query should be a single sentence.

Key Considerations:

- Use a mix of politeness techniques, including indirect phrasing, friendly introductions, and appreciative language.

- Keep the tone natural—avoid overly rigid or formal wording that feels robotic.

- Vary sentence structures instead of defaulting to "Could you please...".

Use emotional undertones and specific attempts at politeness.

- Maintain the original meaning while subtly enhancing the request's politeness and friendliness.

- Avoid beginning the generated query with 'I hope you are...' or concluding it with a separate 'Thank you.' sentence. Generate only one polite query sentence.

Original Query: ...

Polite Query:

Prompt 3:

Task: Refining Queries for Politeness and Warmth

Transform a given query into a more courteous, engaging, and warm request while ensuring it retains the original intent. The revised version should sound friendly, professional, and respectful. The generated query should be a single sentence.

Guidelines:

- Incorporate politeness techniques such as indirect requests, warm introductions, and appreciative language.

- Ensure the tone is natural—avoid excessive formality that feels robotic.

- Diversify sentence structures rather than defaulting to "Could you please...". Use emotional undertones and specific attempts at politeness.

- Subtly enhance warmth and professionalism while preserving clarity and intent.

- Avoid beginning the generated query with 'I hope you are ...' or concluding it with a standalone 'Thank you' sentence. Generate only one polite query sentence.

Original Query: ...

Polite Query:

J LLMs Prompts

J.1 Few-shot Prompts

Few-shot examples:

PopQA:

- readability:

Question: What genre is Golden?

Answer: rock music

Question: In which specific genre does the work titled "Golden" find its classification?

Answer: rock music

- politeness:

Question: What genre is Golden?

Answer: rock music

Question: Would you be so kind as to share with me what genre Golden falls under?

Answer: rock music

- formality:

Question: What genre is Golden?

Answer: rock music

Question: Hey, so like, do you know what genre Golden is?

Answer: rock music

- round-trip translation:

Question: What genre is Golden?

Answer: rock music

Question: What genre of Golden?

Answer: rock music

- typos:

Question: What genre is Golden?

Answer: rock music

Question: What genra is Golden?

Answer: rock music

EntityQuestions:

- readability

Question: Where was Michael Jack born?

Answer: Folkestone

Question: In what geographical locale did the individual known as Michael Jackson enter into existence?

Answer: Folkestone

- politeness:

Question: Where was Michael Jack born?

Answer: Folkestone

Question: Would you be so kind as to share the birthplace of Michael Jack?

Answer: Folkestone

- formality:

Question: Where was Michael Jack born?

Answer: Folkestone

Question: Hey, so like, do you know where Michael Jack was born?

Answer: Folkestone

- round-trip translation:

Question: Where was Michael Jack born?

Answer: Folkestone

Question: Where was Michael Jacques born?

Answer: Folkestone

- typos:

Question: Where was Michael Jack born?

Answer: Folkestone

Question: Where was Michael Jack born?

Answer: Folkestone

MS MARCO:

- readability:

Question: how long can chicken stay good in the fridge

Answer: 1 to 2 days

Question: What is the time span within which chicken can sustain its quality for consumption when preserved in a refrigerated setting?

Answer: 1 to 2 days

- politeness:

Question: how long can chicken stay good in the fridge

Answer: 1 to 2 days

Question: Would you be so kind as to share how long chicken remains fresh in the refrigerator?

Answer: 1 to 2 days

- formality:

Question: how long can chicken stay good in the fridge

Answer: 1 to 2 days

Question: Hey, so like, do you know how long chicken can last in the fridge?

Answer: 1 to 2 days

- round-trip translation:

Question: how long can chicken stay good in the fridge

Answer: 1 to 2 days

Question: How long will chicken stay fresh in the refrigerator

Answer: 1 to 2 days

- typos:

Question: how long can chicken stay good in the fridge

Answer: 1 to 2 days

Question: how leng can chickon stay good in the fridge

Answer: 1 to 2 days

Natural Questions:

- readability:

Question: how many pieces in a terry’s chocolate orange

Answer: six

Question: What is the total quantity of individual segments contained within a Terry’s chocolate orange confectionery item?

Answer: six

- politeness:

Question: how many pieces in a terry’s chocolate orange

Answer: six

Question: Would you be so kind as to share the number of segments typically found in a Terry's chocolate orange?

Answer: six

- formality:

Question: how many pieces in a terry's chocolate orange

Answer: six

Question: Hey, so like, do you know a terry's chocolate orange contains how many pieces

Answer: six

- round-trip translation:

Question: how many pieces in a terry's chocolate orange

Answer: six

Question: How many pieces of Terry's Chocolate Orange

Answer: six

- typos:

Question: how many pieces in a terry's chocolate orange

Answer: six

Question: how meny pieces in a tarry's chocolate orange

Answer: six

Prompt:

You are a professional question-answer task assistant. Use the following pieces of retrieved context to answer the question briefly.

Context:

contexts

Below are examples of questions and answers:

few_shot_examples

Now, it's your turn to answer the question below. The answer should contain ONLY one sentence and DO NOT explain reasons.

K Rewriting Examples

Category	Type	Query
RTT	Original	What type of music does The Eruption of Mount St. Helens! play?
	Rewritten	What music the eruption of Mount St Helens! play?
RTT	Original	Who is Hilde Coppi married to?
	Rewritten	With whom was Hilde Coppi married?
RTT	Original	Which company is HMS Blankney produced by?
	Rewritten	What company is producing HMS Blankey?
RTT	Original	Where is Flemington Racecourse located?
	Rewritten	Where is Flemington Racecourse?
Typos	Original	Where was R. Kent Greenawalt born?
	Rewritten	Wher was R. Kent Greenawalt born?
Typos	Original	What kind of work does M. Ramanathan do?
	Rewritten	What kind ofh work does M. Ramanathan do?
Typos	Original	What type of music does El Cantor del circo play?
	Rewritten	What type of music does El Cantorh del circo pay?
Typos	Original	What is Rembrandt famous for?
	Rewritten	What is Rembrandt faumous for?
Formality	Original	Which company is Galaxy Camera produced by?
	Rewritten	Hey, quick question! Which company actually makes the Galaxy Camera?
Formality	Original	Who is the author of Intensity?
	Rewritten	Yo, do you know who wrote Intensity?
Formality	Original	Which country is Parchliny located in?
	Rewritten	Hey, just curious, do you know what country Parchliny is in?
Formality	Original	Who is Liu Bei's child?
	Rewritten	Hey, so, do you know who Liu Bei's kid is? I'm super curious about it!
Readability	Original	What type of music does Anbe Sivam play?
	Rewritten	What genre of musical compositions is performed by Anbe Sivam?
Readability	Original	Where was FC Utrecht founded?
	Rewritten	In what location was the establishment of FC Utrecht initiated?
Readability	Original	Where was John Ernle educated?
	Rewritten	At which institution did John Ernle receive his education?
Readability	Original	Where was The Shiru Group founded?
	Rewritten	In which geographical location did The Shiru Group originate?
Politeness	Original	What music label is Time in Place represented by?
	Rewritten	May I kindly inquire which music label represents Time in Place?
Politeness	Original	Which country was The Border Blasters created in?
	Rewritten	Would you be so kind as to share which country The Border Blasters originated from?
Politeness	Original	Which country is Oleksin, Otwock County located in?
	Rewritten	Could you kindly share which country Oleksin, Otwock County is situated in?
Politeness	Original	Where did Wolfe Tone die?
	Rewritten	Would you be so kind as to share the location where Wolfe Tone passed away?

Table 38: Rewriting examples across all linguistic variations from the EntityQuestions dataset, with queries split across rows for readability.

L Computational Resources

Our experimental setup utilized models of varying scales: Gemma-2 (2B, 9B, 27B parameters), Llama-3.1 (8B, 70B parameters), and Qwen-2.5 (3B, 7B, 32B, 72B parameters). For retrieval, we employed ModernBERT Embed (149M parameters) and Contriever. We conducted comprehensive evaluations across 5 linguistic dimensions (4 dimensions plus 2 grammatical correctness subtypes), 4 datasets, 9 language models, and 2 retrieval systems, totaling 360 experimental configurations. Each model inference run required approximately 1.5 hours, resulting in 540 GPU hours on 16 L40S GPUs distributed across different model configurations. Retrieval evaluation required an additional 40 GPU hours. Data rewriting was performed using GPT-4o-mini, requiring 40 hours of API usage. Total computational cost comprised 620 GPU hours on L40S hardware plus commercial API usage for data preprocessing.