QuickFix:

A Multi-step Query Reformulation Method For Children's Online Search Queries

Atilla Colak
Supervisors: Sole Pera, Hrishita Chakrabarti
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Atilla Colak
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Hrishita Chakrabarti, Catholijn Jonker

Abstract

Children often struggle to retrieve age-appropriate information when seeking information online. One big reason for this is that their search queries are short, misspelled, or vague. As a solution to this problem, previous research investigated query reformulation, where the input query is manipulated in a way that the retrieved web results are more child-appropriate. This was measured by various metrics and scores, such as readability and content safety of the retrieved web search results. The problem with present query reformulation strategies, however, is that each tackles this problem from one perspective, missing out on the potential benefits of other perspectives. For instance, expanding the query with the "for kids" cue has shown to be a good way to target a specific audience and helps retrieve more child-appropriate content; however, on top of this considering substitutes for uncommon words with simpler synonyms might further enhance the child-appropriateness of the retrieved results as it tackles the reformulation from a different perspective than "for kids" audience cue expansion.

Motivated by this, we propose a multi-step query reformulation strategy that combines multiple reformulation strategies and applies them to the given input child query in a multi-step manner using a Large Language Model (LLM). We use LLM to apply the reformulation strategies to the input query in a chain-of-calls (where each call is prompted to apply a different reformulation strategy). This proposed method captures the perspective of multiple reformulation strategies, rather than a single one, unlike existing reformulation strategies. The results of our experiments, which include a baseline comparison (of the retrieved search results from the reformulated query against the original query) and an ablation study, provide insight into the performance of our strategy.

With this work, we aim to demonstrate the potential of combining multiple reformulation strategies and their impact on improving the readability and content safety of retrieved web search results when applied to children's search queries. Our findings reveal a significant improvement in the readability of retrieved results after using the proposed reformulation method. Ultimately, this work contributes to the development of next-generation, child-centric search systems that deliver clearer, safer results for children.

1 Introduction

Web search is an integral part of children's day-to-day lives. In the United States, more than 70% of young people aged 6 to 13 years use the Internet in various contexts, such as at school or home [1]. However, many studies show that, when searching online, children write queries that are *short*, *misspelled*, and often *underspecified* [2, 3]. Furthermore, because standard search engines are fine-tuned for adults, suboptimal child queries in these engines retrieve materials and web pages that exceed the child's reading level or contain unsafe language [4]. This, in return, results in problems for children. For instance, for a child who uses online search for learning, when the top results are not suitable for the child's comprehension and language level, the learning process becomes slower [5].

Prior approaches and their limitations. To address this issue, researchers have explored various methods along the information retrieval (IR) process, including query reformulations (where the input child query is reformulated in a way that aims to improve the child suitability of retrieved web results from that query). These investigated methods include a spelling and grammar correction reformulation [6], a word substitution reformulation where uncommon words are replaced with more common synonyms [6], and a keyword expansion reformulation that adds the phrase "for kids" to the input query [7]. Beyond these, further

strategies have also been proposed. For instance, [6] also implemented phonetic reformulation (to correct verbal spelling errors) and acronym resolution (to address the abbreviated writing styles found in some children's queries).

Yet, these existing reformulation strategies address the problem from only one perspective. Hence, even though several produced promising results (for example, [7], where a few explored strategies improved the ordering of child friendly results, and [6], where some reformulation methods consistently achieved relevant reformulations for the input queries), they miss out on further improved results that may come from accounting for multiple perspectives at the same time. Thus, what is missing is a reformulation method that captures multiple perspectives simultaneously.

Contributions. Towards this gap, we propose a multi-step query reformulation method using an LLM and pose the following research question (**RQ**): *To what extent can a multi-step query reformulation using LLM impact the readability and content-safety of retrieved results for a given children query?* With this research, we aim to contribute to future research on query reformulation methods by introducing a new angle for investigation.

Approach. To address this research question, we develop a multi-step query reformulation strategy that applies three existing reformulation techniques in sequence: (i) grammar and spelling correction, (ii) synonym simplification where uncommon or advanced words are substituted with a simpler synonym, and (iii) audience cue expansion that appends "for kids" to the end of the query. This design is grounded in the premise that LLMs perform well in similar linguistic and IR tasks [8–10]. These reformulation steps are implemented as a chain of LLM prompts, with each prompt focusing on one specific reformulation.[1]

Experimental Design. To evaluate the effectiveness of our method, we conduct two empirical experiments using a dataset comprising 301 English queries authored by children aged 6-13, released by [11]. In our first experiment (baseline comparison), we compare the original queries against their fully reformulated counterparts in terms of readability and content safety of retrieved web results, to understand the overall performance of our proposed reformulation method. In the second experiment (ablation study), we isolate and apply each reformulation step independently to assess the individual contribution of each reformulation step. These experiments and evaluations allow us to examine both the overall and step-wise effectiveness of our proposed multi-step reformulation strategy.

Outline of the results. Our experiments demonstrate a statistically significant increase in the readability of the retrieved results (as measured by the web result descriptions) when a query is reformulated using the proposed method. Moreover, from the ablation study, we observe that full multi-step reformulation outperforms every individual reformulation step tested. There is also a statistically significant decrease in content safety, but the decline in amount is negligibly slight. Based on these results, we empirically show that our proposed multi-step reformulation method delivers a clear benefit, which is retrieving web results that are easier to understand for children. Finally, we also make the empirical observation that considering the multiple perspectives leads to better results than any single perspective used in our experiment in isolation.

Paper outline. In **Section 2 (Related Work)**, we delve further into the existing reformulation research efforts and look into where our proposed method is positioned in this space. **Section 3 (Methodology)** explores our implementation and LLM instructions and details. In **Section 4 (Experimental Setup)**, we explain our experimental procedure,

[1]The complete project code is available at `https://github.com/AtillaColak/QuickFix`

metrics, dataset, and the method and tools we use to test and statistically validate our proposed method. In **Section 5 (results)**, we report our quantifiable findings. This is followed by contextualization of the results and their interpretation in the broader context of child-friendly web search in **Section 6 (Discussion)**. We address the ethical considerations of our study and its reproducibility in **Section 7 (Responsible Research)**. Finally, **Section 8 (Conclusions, Limitations, and Future Work)** summarizes the key findings of our work and outlines the limitations as well as potential next steps of future research from it.

2 Related Work

When searching for information online, children were found to struggle with query formulation [12–14]. The implications of this struggle are exacerbated by the fact that common search engines are not primarily designed for child users [15, 16] and offer little to no help with children's query formulation process [11]. This leads to the retrieval of web results that are not appropriate for children [4].

Many previous works approached this problem from the angle of query reformulation and suggestion. A promising method proposed by [7] utilizes social media tags to suggest queries that are more relevant to the child's topic, such as including the "for kids" tag in the recommended query. A complementary study proposes a completion-style suggestion module that builds candidate phrases from large corpora written by or for children and then ranks them on seven "kid-friendliness" features [17].

One common shortcoming of existing query reformulation and suggestion methods is that they approach the problem from a single perspective. Furthermore, although some explorations have approached the solution from multiple perspectives, these efforts are limited and not recent. For instance, [11] introduced the *ReQuIK* system that generates and evaluates query suggestions based on four perspectives before them: (i) classifies the query's search intent, (ii) scores each suggestion on child-centric lexical and topical features (kid-friendliness), (iii) filters the suggestions based on the readability level of potential retrieved documents, and (iv) removes near-identical queries to make sure there is topical diversity. Although this proposed method appears to improve recommendation quality and is multi-perspective in its decision making for which suggestions to keep, it is still one-dimensional in the underlying query suggestion generation, as all candidate suggestions are produced by a single auto-complete-style generator rather than combining distinct reformulation strategies.

By contrast, one work that explores reformulating queries in a multi-perspective manner is by [6], where, as a small part of their investigation, they combine different reformulation methods. Although valuable as an early exploration towards this angle, it leaves two key gaps that motivate our work.

1. The study is **dutch-specific**: it is implemented and evaluated only for Dutch queries, with hand-built Dutch phonetic and *WordNet* resources[2]. Therefore, its findings or implementation do not easily generalize to English or other languages.

2. Its **evaluation endpoint is relevance of reformulation**: it judges merely whether added query expansion terms are "relevant" to the original query based on manual evaluation of "with how many relevant words was the query expanded". It does not

[2]The Dutch Wordnet used in this is available at `https://www.fon.hum.uva.nl/paul/papers/1999-uva-VossenBloksmaBoersma.pdf`

have any analysis of how these expanded queries would perform when searched on the web and whether they would impact readability and content safety of these retrieved web search results (as these metrics are important to understand how suitable the web search results would be for the children).

These limitations highlight the need for a multi-step query reformulation method and exploration of how it impacts the "child-friendliness" of the search results; thus, it presents an opportunity for our proposed method. By combining multiple reformulation methods and applying them to a given child query using an LLM, we can capture the perspective of each reformulation method used in our pipeline. We use LLM for our reformulation tasks, given the tested use of LLMs in similar IR tasks, such as query reformulation and rewriting [8–10].

3 Methodology

This section provides an in-depth analysis of our LLM details, reformulation pipeline, and the key decisions made during our implementation.

LLM model details. We use the Gemini 2.5-Flash LLM model[3]. We chose this specific LLM model because of (i) how well it performs in *reasoning tasks*[4] against many other top LLM models, and (ii) the big context window of $1M$-tokens. These factors enable us to safely embed reformulation rules and output constraints alongside the noisy child query input.

Reformulation Strategy. For a given arbitrary original query instance $Q_0^{(i)}$, we produce its reformulated counterpart $Q_{\text{full}}^{(i)}$ using the Gemini 2.5-Flash LLM. The three reformulation steps described in Table 1 are sequentially applied in that order to the given query input. We also use an LLM chain where each reformulation instruction is applied in a separate LLM call and the result is passed onto the next reformulation step as input, rather than bundling all instructions into one. We do this to reduce hallucination risk and improve adherence to every guideline and reformulation instruction, an approach supported by evidence that LLM models can overlook individual requirements when they are presented simultaneously [18, 19]. Furthermore, we chose these reformulation steps not only for their exploration in the previous literature and promising results [6, 7], but also to minimize the potential change on the semantic meaning of the input query (as we aim to focus our findings and analysis primarily on child-friendliness and we do not investigate the semantic similarity of original queries and their reformulated counterparts.)

For the constraint c_1 in Table 1, we enforced this word limit to make sure that LLM has minimal room to hallucinate and add unintended additional clauses to the output. We also found that the longest query in our experiment dataset was 19 words long. After fixing the grammatical errors and appending "For Kids" to it, the query becomes 20 words long. Therefore, we decided that the designated word limit was reasonable even for the longest query in our dataset.

[3]The model details are available at https://deepmind.google/models/gemini/flash/

[4]*Humanity's Last Exam (HLE)* is a benchmark comprising roughly 12000 graduate-level questions that span various domains, crafted to probe advanced chain-of-thought reasoning and broad factual knowledge.

ID	Description
r_1	Fix grammatical and spelling errors.
r_2	Replace uncommon or advanced words with simpler synonyms, preserving original meaning and not altering proper nouns or titles.
r_3	Append "for kids" to the end of the query.
c_1	Keep it under 21 words. Do not add new subject matter, opinions, or links.

Table 1: Rules (r) and the output constraints (c) for the LLM

LLM Prompt. The *system instruction* (see Listing 1) declares the assistant's role, injects exactly one rule $r_j \in \{r_1, r_2, r_3\}$, and re-states the global constraint c_1.

The *user message* then supplies the original text and ends with the marker *"Rewritten query:"*. Gemini is instructed to reply with *only* the rewritten string without any explanations or extra tokens. As an additional step of caution, we set the temperature to 0 (to make the LLM outputs more deterministic). This prompting strategy is reused for r_1, r_2, r_3, allowing us to chain the three reformulation steps as also shown in Figure 1.

```
system_instruction = (
    "You are a query-rewriting assistant for children ages 6-13.\n"
    f"{rule_key}: {RULES[rule_key]}\n"
    f"Constraints: {CONSTRAINTS}\n"
    "Output only the rewritten query with no extra text."
)

prompt = ( # this is the user message
    f"Apply this rule: {Rules[rule_key]}\n"
    f"Original query: {query}\n"
    "Rewritten query:"
)
```

Listing 1: LLM prompts. "rule key" is the reformulation rule of the current step.

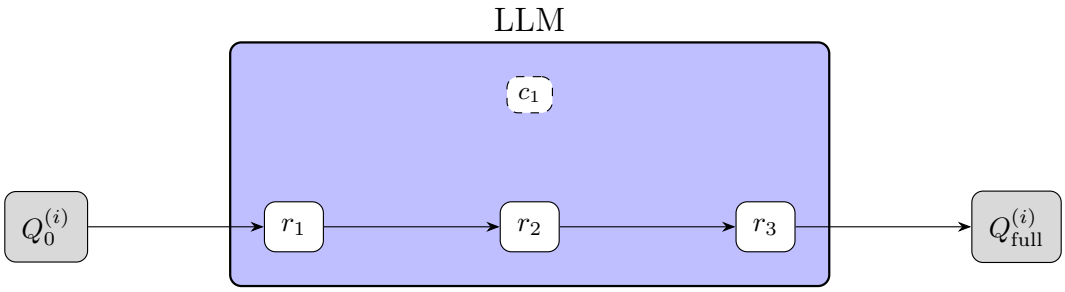


Figure 1: Per-instance multi-step query-reformulation pipeline. The dashed box shows the system-level guideline c_1; the blue rectangle groups the LLM steps.

4 Experimental Setup

This section explains our dataset, the scores we use for our evaluations, and the rationale for each analysis. Three main parts of our experiments are summarized below (as also shown in Figure 2).

- **Web-results collection**: run original queries ($\mathcal{Q}_0$), their fully-reformulated counterparts ($\mathcal{Q}_{\text{full}}$), and their single-step reformulated versions (notated by $\mathcal{Q}_{r_1}$, $\mathcal{Q}_{r_2}$, and $\mathcal{Q}_{r_3}$ respectively for each reformulation rule provided in Table 1) on Brave Search API and collect the top-10 retrieved web result snippets for every query (§4.1).
- **Evaluation scores computation**: for each query, compute seven numeric scores for every snippet (three readability scores and four safety probability scores) in its top-10 list and then average those seven numeric scores across the ten snippets (§4.2).
- **Performance evaluation**: test whether the reformulated queries outperform the originals and diagnose the impact of each of the rules in the overall result (§4.3).

Our overall goal in these experiments is therefore to (i - Baseline comparison) quantify the effect of the full multi-step reformulation and (ii - Ablation study) explain those effects by re-running the pipeline with each reformulation step in isolation. For our experiments, we use the `Children-Queries` dataset[5], which contains $N = 301$ English queries typed by children aged 6-13, released by [11].

4.1 Web Results Collection

We execute all queries ($\mathcal{Q}_0$, $\mathcal{Q}_{\text{full}}$, $\mathcal{Q}_{r_1}$, $\mathcal{Q}_{r_2}$, and $\mathcal{Q}_{r_3}$) on the *Brave Search API*[6] which returns the top 10 snippets for each query. In this work, by "snippets" we refer to the web result descriptions. The cut-off at ten mirrors prior IR studies [4]. Children also favor high-ranking hits [2, 20]. Moreover, evaluating *snippets* rather than full pages is a practice followed because (i) it allows for faster and still relevant evaluation of the web results [21] and (ii) it has the potential to largely impact the clickthrough behaviors [22].

For convenience, we introduce $\mathcal{R}_v$ for the set of all top-10 snippet lists retrieved for $\mathcal{Q}_v$, $v \in \{0, \text{full}, \text{r}_1, \text{r}_2, \text{r}_3\}$.

4.2 Evaluation Scores

For every snippet s in the top-10 retrieved snippets of a query, we compute three **readability** scores and four **content-safety** probabilities, and then average each metric across those 10 snippets. By doing so, we obtain seven numerical scores per query.

[5]The dataset we used for our experiment is available at `https://scholarworks.boisestate.edu/cs_scripts/5/`

[6]Chosen for its generous monthly quota and JSON snippet returns.

Readability Scores. The three readability scores we use are computed with `textstat`[7] and shown below with their formulas:

1. **Flesch-Kincaid Grade Level (FKGL):** $0.39\left(\frac{\#\text{ words}}{\#\text{ sentences}}\right) + 11.8\left(\frac{\#\text{ syllables}}{\#\text{ words}}\right) - 15.59.$
2. **Coleman-Liau Index (coleman):** $0.0588\,L - 0.296\,S - 15.8,$
   where L is the average number of letters per 100 words and S the average number of sentences per 100 words.
3. **Dale-Chall (dale):** $\text{Raw} = 0.1579\,(\%\text{difficult words}) + 0.0496\left(\frac{\#\text{ words}}{\#\text{ sentences}}\right);$

   $$\text{Score} = \begin{cases} \text{Raw} & (\% < 5) \\ \text{Raw} + 3.6365 & (\% \geq 5). \end{cases}$$

   "Difficult" words are those *not* in the Dale-Chall familiar-word list.

For all three of these metrics, a lower score indicates an easier-to-understand text. We chose FKGL because of its wide adoption in the literature [23]. Dale-Chall complements FKGL by focusing on *lexical familiarity* rather than word length, flagging short but technical terms (e.g., *ion*, *URL*). The Coleman-Liau index relies on characters-per-word and sentences-per-100-words, making it robust for short web snippets where sentence segmentation may be unreliable; hence, using a character-based measure adds another length-agnostic view of difficulty [24].

Content Safety Scores. We query the Google Perspective API[8] for the attributes `TOXICITY`, `PROFANITY`, `THREAT`, and `INSULT`. Perspective uses a multilingual Transformer fine-tuned on millions of crowd-labeled comments for each of the attributes; each attribute returns a probability between 0.0-1.0 interpreted as the likelihood a reader would perceive the snippet as containing that attribute.

We use Perspective because it offers a *nuanced* view of online harm: beyond simple profanity matching, it distinguishes multiple kinds of toxic speech (such as threat, insult, and general toxicity). This lets us gauge content-safety for children along several safety dimensions rather than a single coarse profanity count.

Let $\mathcal{M}_{k,v}$ be the set of *per-query* mean scores for metric $k \in \{\text{FKGL, dale, coleman, toxicity, profanity, threat, insult}\}$, computed for the queries of variant $v \in \{0, \text{full}, \text{r}_1, \text{r}_2, \text{r}_3\}$. Each per-query mean is obtained via the evaluation procedure described above in this subsection.

4.3 Performance Evaluation

After collecting the web results and evaluation scores, we conduct two complementary analyses:

(i) Baseline comparison. This comparison helps answer the question "Does the complete, three-step reformulation improve the readability and content safety of retrieved web results over original child queries?" The full multi-step reformulated queries $\mathcal{Q}_{\text{full}}$ are contrasted with the original child queries $\mathcal{Q}_0$. For every metric k, we compute the paired differences $\Delta_{k,\text{full}} = \mathcal{M}_{k,\text{full}} - \mathcal{M}_{k,0}$. Normality of each $\Delta_{k,\text{full}}$ distribution is checked with Shapiro-Wilk (α=0.05) and visual analysis. If approximately normal, we apply a two-tailed paired t-test; otherwise, the Wilcoxon signed-rank test. Then, we report the two-tailed p-value and a sample median value of the differences ($\Delta_{k,\text{full}}$).

(ii) Ablation study. This evaluation isolates the *contribution* of each reformulation step. We repeat the above statistical procedure (in section (i)) three times, contrasting $\mathcal{Q}_{r_1}$, $\mathcal{Q}_{r_2}$, and $\mathcal{Q}_{r_3}$ *individually* against $\mathcal{Q}_0$. This reveals the individual impact of each of the reformulation methods used (r_1, r_2, and r_3).

[7] `https://pypi.org/project/textstat/`

[8] `developers.perspectiveapi.com`

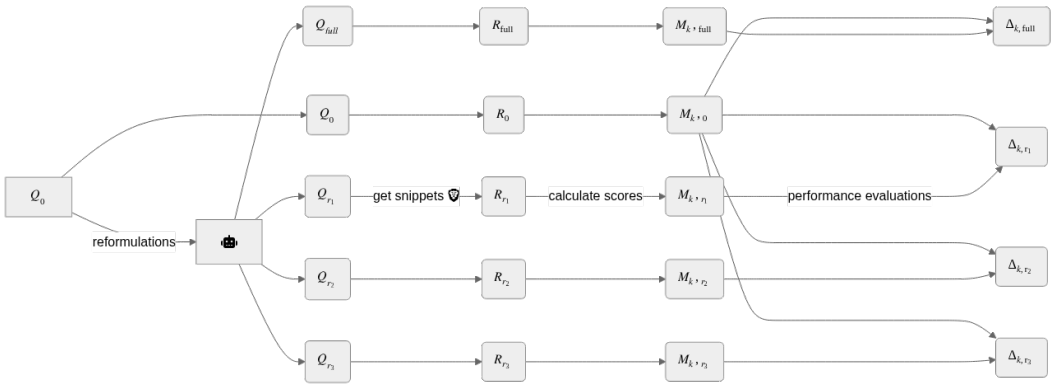


Figure 2: End-to-end experiment pipeline. This pipeline is executed for each of the seven evaluation metrics. $k \in$ {FKGL, dale, coleman, toxicity, profanity, threat, insult}

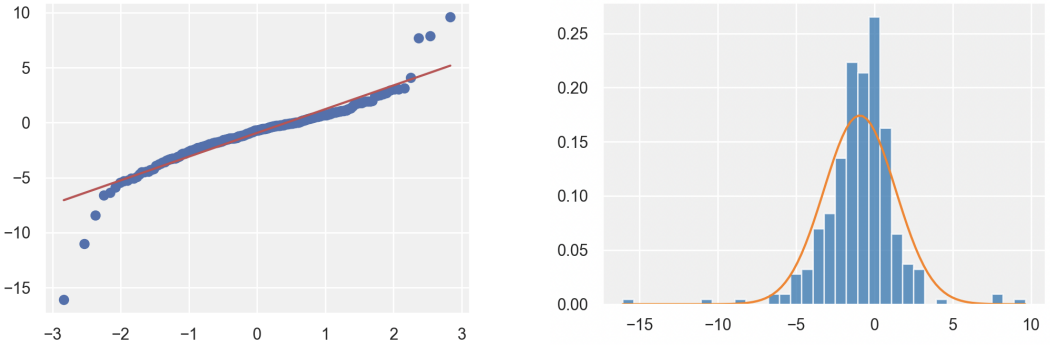
4.4 Significance Analysis

The Shapiro-Wilk statistics for all 28 metric-query-variants combinations (k-v) of $\Delta_{k,v}$ ($\mathcal{M}_{k,v} - \mathcal{M}_{k,0}$) were significant ($p < 0.05$), rejecting normality. However, this statistic becomes more sensitive to even mild variations from normal as the sample size grows. Therefore, we complemented the analysis with visual checks. Supporting these normality results is Figure 3 demonstrating heavier tails, outliers, and also mild skew for the distribution of $\Delta_{\text{coleman,full}}$. Similar trends appear for the rest of the metric-query-variant combinations (complete set of graphs shown in Appendix A).

Although the paired t-test is generally robust enough for moderate normality violations with a large sample size of 301 values ($n > 30$) [25], our distributions don't demonstrate enough symmetry to justify the use of the paired t-test either. Therefore, for our **results** section, we continue with the Wilcoxon signed-rank test (with $\alpha = 0.05$) with the following hypotheses:

$\mathbf{H}_0$ (null): For the given metric, the population medians of the scores calculated for the reformulated and the original queries are equal; hence, the reformulation has no significant effect.

$\mathbf{H}_\text{A}$ (alternative): For the same metric, the population medians of the scores calculated for the reformulated and original queries differ; hence, the reformulation has a significant effect.



(a) Q-Q plot

(b) Histogram

Figure 3: Differences distribution for the Paired Coleman-Liau scores of full reformulation and original ($\Delta_{\text{coleman,full}}$).

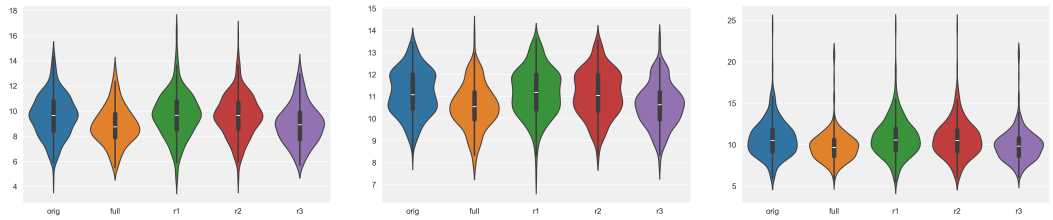
5 Results

This section reports the quantitative findings of and observations made from our experiment results.

5.1 Readability

We've observed that the proposed multi-step reformulation method reduced the readability grade level on average (using the median) by 0.5-0.7 grade levels. However, the "for kids" expansion rule (r_3) also reduced the readability levels on average by 0.4-0.6 levels. The other two reformulation rules, r_1 and r_2, when applied in isolation, showed a zero median improvement in the readability levels. The reason for the zero median improvements was that more than 50% of the web results retrieved for each of these ablations (57.1% for r_1 and 73.1% for r_2) were the same as their original counterparts. As an extra step of confirmation, for r_1 and r_2, we looked at the results again after removing instances where no reformulation was applied to the original query instance. This left us with 51.8% of the original result set for r_1 and with 30.6% for r_2. Out of these remaining instances for r_1, only in 27.6% of the cases, the retrieved snippets were identical to those of the original query. For r_2, this percentage was 14.1%. Moreover, even after analysing for these instances where the queries were indeed changed, there was still no significant readability improvement for either r_1 or r_2 ablation. Hence, we observe that **the fully reformulated queries achieve the best readability improvement when compared with any ablations** (a pattern also shown in Table 2).

Moreover, Figure 4 reveals that the outlier readability scores of Coleman-Liau are more extreme than the outliers of the other two readability scores.



(a) FKGL (b) Dale-Chall (c) Coleman-Liau

Figure 4: Distribution of the readability scores for the results of each query variant collection (lower = easier to understand). White dots mark the medians. Thick bars show the inter-quartile range.

Metric	Variant vs. *orig*	p-value	Sample Median Difference
Dale-Chall	*full*	$<.001$	-0.55
	r1	0.44	0.00
	r2	0.31	0.00
	r3	$<.001$	-0.43
FKGL	*full*	$<.001$	-0.67
	r1	0.50	0.00
	r2	0.75	0.00
	r3	$<.001$	-0.52
Coleman-Liau	*full*	$<.001$	-0.70
	r1	0.31	0.00
	r2	0.13	0.00
	r3	$<.001$	-0.60

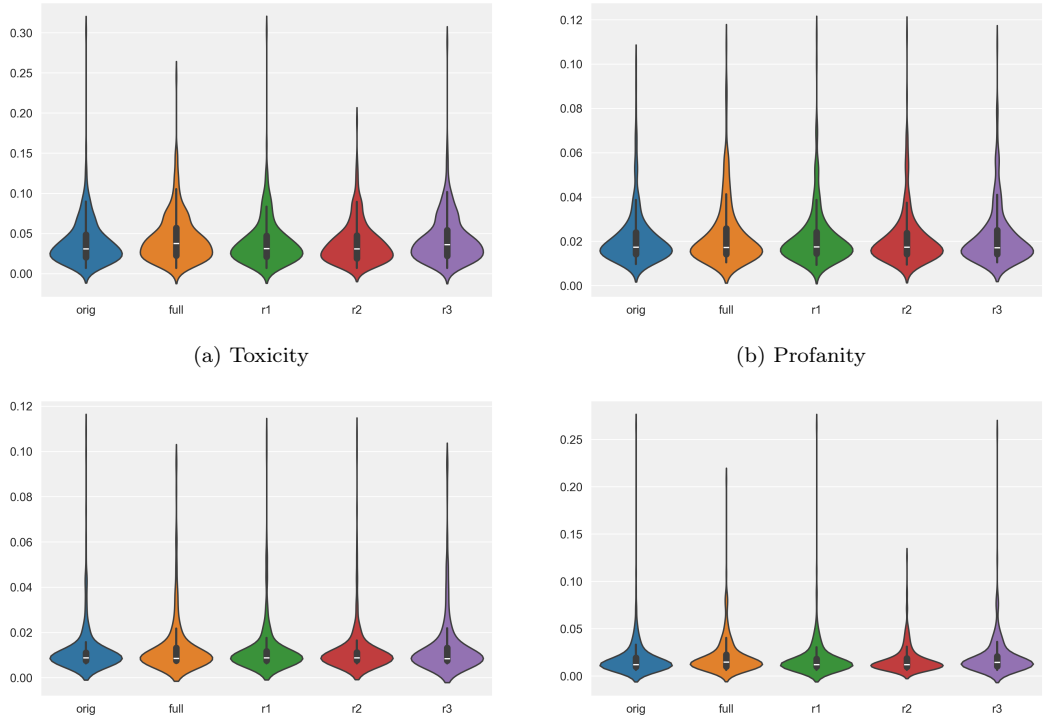
Table 2: Reformulation results in terms of each readability score. Negative median difference indicates that, on average, results retrieved from the respective query variant were easier to understand than their original query counterpart.

5.2 Content safety

We've observed no considerable improvements in content safety after reformulation. Although the content safety after reformulation is lower (measured by the positive sample median difference for Toxicity, Profanity, and Insult), the size of the impact appears negligibly small (also shown in Table 3). The largest of these observed decreases in content safety was based on the Toxicity measurement. For this, the 0.0031 sample median increase indicates that results retrieved from a given reformulated query, on average, are 0.31% more likely to be classified as containing Toxic language than their original query counterpart.

Although a similar "increase in toxicity and profanity" pattern follows for r_3 ablation, the impact size seems to be even smaller than that of full reformulation. Furthermore, the content safety impact of r_1 and r_2 ablations in isolation seems to be extremely small.

Across all variants, however, the distributions and observed Perspective probabilities (as also seen in Figure 5) remain well below Perspective's suggested research decision threshold (to classify a text as containing that attribute or not) of 0.70[9]. This confirms that the practical impact is negligible, as even the most extreme outlier observed in any one of these attributes or distributions is below 0.35.



(a) Toxicity

(b) Profanity

(c) Threat

(d) Insult

Figure 5: Violin plots for the four Perspective attributes (lower = less likely to contain that unsafe language). White dots mark the medians. Thick bars show the inter-quartile range.

[9]The full details of intended various purpose thresholds are available at `https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US`

Attribute	Variant vs. *orig*	p-value	Sample Median Difference
Toxicity	*full*	<0.01	.0031
	r1	0.39	$<.0001$
	r2	0.29	$<.0001$
	r3	<0.01	.0018
Profanity	*full*	0.09	.0002
	r1	0.80	$<.0001$
	r2	0.02	$<.0001$
	r3	0.69	-.0001
Threat	*full*	0.40	.0000
	r1	0.98	$<.0001$
	r2	0.61	$<.0001$
	r3	0.82	-.0000
Insult	*full*	<0.01	.0010
	r1	0.42	$<.0001$
	r2	0.22	$<.0001$
	r3	<0.01	.0008

Table 3: Wilcoxon signed-rank test results for each Perspective content safety attribute. Negative median difference indicates likely safer average text (of that attribute) retrieved from the respective query than the average text retrieved from the original query.

5.3 Overall

Our results establish that the multi-step reformulation indeed effectively improves the readability of the retrieved web search results on all three readability scores (FKGL, Dale-Chall, Coleman-Liau) while having a negligibly small negative impact on the safety scores.

The single-step ablations reinforce this finding: only the "for kids" expansion (r_3) produces a smaller but still noticeable readability gain. Although spelling and grammar correction (r_1) and synonym substitution (r_2) in isolation do not show prominent readability improvements, when they are combined with r_3 in the full multi-step reformulation pipeline, the readability improvement is greater than when r_3 ablation is applied in isolation (indicating their added contribution when combined). Out of all three ablations, across four perspective attributes, none of them has a considerable impact (sample median difference $<= 0.01$ on the 0.0-1.0 scale).

The interpretation of our findings in this chapter is discussed in the following **Discussion** section.

6 Discussion

Our results show that chaining spelling/grammar correction, synonym substitution, and the "for kids" expansion inside an LLM chain reduces the average reading grade of the top-10 search snippets on average by 0.5-0.7 grade levels, without considerably worsening content safety scores (on average $\Delta < 0.01$ on the 0.0-1.0 scale). The improvement is statistically significant on all three readability scores, and it exceeds the improvement achieved by any of the three reformulation steps used alone. Our finding mirrors previous research, where the expansion "for kids" appears to improve the results [7], but extends it by demonstrating the additional gain achieved by combining reformulation strategies.

The ablation results show that appending "for kids" (r_3) accounts for the bulk of the improvement, while grammar and spelling correction (r_1) and synonym substitution (r_2) matter only in combination. One likely reason is that many child-focused sites include explicit audience cues, such as the terms "kids" or "for kids", in their URLs, titles, tags, or anchor text; appending "for kids" therefore may be matching these cues directly and promotes such pages into the top 10 [7]. On the

other hand, correcting a single misspelling or substituting synonyms might be treated as roughly equivalent by the internal ranking system. This is further supplemented by the result that even after looking only at the results where r_1 and r_2 reformulated the original query, there was still no significant change in readability scores.

Thus, considering our experiment results, our proposed method successfully addresses the readability objective of our research question while negatively but minimally impacting the content safety objective of it. More importantly, this success is a concrete step towards covering our identified research gap: the lack of multi-perspective children search query reformulation methods.

Moreover, these findings situate our work within a growing body of research that leverages LLMs to mediate between user queries and retrieval systems [9, 10]. Recent studies on generative query rewriting for conversational search [10] and ensemble prompt strategies [9] confirm that LLMs can be used to successfully reformulate user queries (adding clarifying words, paraphrasing, or appending intent cues) without the labor-intensive step of hand-crafting linguistic rules. Our child-centric results complement those efforts by showing that an audience-aware query reformulation LLM pipeline can deliver measurable accessibility gains for a vulnerable user group that is children, whose needs are often overlooked by mainstream search technologies [26].

6.1 Design implications for information access systems

Our reformulation method offers numerous real-life implications and use cases to explore.

1. Reformulation as a *client-side* service. Because our pipeline operates entirely at query time and needs only one public endpoint (for LLM inference calls), it could run in a browser extension or a school proxy server. This avoids the regulatory overhead of hosting a dedicated "children's search engine" while still delivering the child-oriented results. Moreover, given the lightweight nature of such a reformulation layer, it makes it possible to adjust it to various similar use cases (by, for instance, extending it to new reformulation strategies, or different search engines and information retrieval systems).

2. Development of a multi-language reformulation system. LLMs already carry cross-lingual knowledge and, depending on how the model is further trained, can perform cross-lingual tasks [27]. Therefore, a language tag (`[LANG=NL]`) prepended to the system prompt could allow the same reformulation system to serve children writing non-English queries, with minimal additional tuning required. This would be an economically attractive option for low-resource markets (such as international schools).

7 Responsible Research

Throughout our paper, we considered and aimed to adhere to high ethical standards.

- **Dataset:** In our experiments and dataset, we made sure there is no affiliation to any individual who took part in the formation of the dataset and that it is an anonymous set of queries written by children. Furthermore, the dataset was also approved by Boise State University's Institutional Review Board **(IRB)**, which makes sure that human data adheres to strong ethical guidelines[10].
- **Reproducibility:** In our implementation and analysis, we tried to minimize randomness so that our work is easily reproducible. Whenever we used pseudo-randomly generated values, we specified the seed so that other researchers could use the same values. Furthermore, we documented and made publicly available our code for the implementation and statistical

[10]More details regarding Boise State University's IRB guidelines and ethical requirements available at `https://www.boisestate.edu/research-compliance/irb/guidance/`

analysis, the datasets we used, and the accumulated final as well as intermediate results of our experiments in the project repository.

- **Use of AI:** In this paper and experimentation, we used ChatGPT for two purposes: (i) generating boilerplate Python statistical analysis code and (ii) generating LaTeX tables and a `tikz` diagram (the list of prompts we used is provided in Appendix B). Nevertheless, whenever we used ChatGPT, we manually verified the code it generated and made sure of the validity of the output and did not rely on it as a drop-in solution. For the statistical analysis Python code, we validated the code line by line, fixed mistakes, and made sure that the statistical methods were appropriately used. For the LaTeX code, we did not use the values AI auto-filled in. Instead, we copied the code for the table or diagram, deleted the values, and inputted the values ourselves to make sure there is no mistake.

8 Conclusions, Limitations, and Future Work

This work demonstrated that a three-step Gemini 2.5-Flash reformulation pipeline (composed of grammar & spelling correction, synonym simplification, and a "for kids" audience cue expansion) reduces the average reading grade of top-10 snippets by roughly 0.5-0.7 levels while minimally impacting the content safety. An ablation study revealed that while "for kids" expansion drives most of the readability gain, combining it with the two reformulation steps prior yields the best overall improvement, establishing that multi-perspective reformulation is superior to any individual step investigated in this paper.

However, these positive findings have some **limitations and need for future exploration**. Results may be Brave-specific because ranking and snippet generation differ across engines; repeating the experiment on Google or Bing would test generalizability [28]. Our handling of Brave's auto-generated placeholder snippets (roughly 1% of the snippets collected) could also influence outcomes and deserves a sensitivity analysis: because children would also see these auto-generated snippets, we included them in our analysis.

Moreover, as a starting point, we focused our experiments and LLM prompts on English queries. This decision stemmed from the fact that LLMs have been shown to perform the best in English, given the predominantly English corpus they are pre-trained on [29]. Nevertheless, the explored cross-lingual performance of LLMs [30, 31] offers a potential to extend our implementation to handle non-English queries as well.

Another limitation is that we measured readability and safety, but not topical relevance after reformulation, so future work should collect graded relevance judgements or simulate clicks to measure any relevance-readability trade-off. In addition, our content safety calculations rely on single calls to the Perspective API, whose probabilistic scoring can show slight variations across two requests for the same given text. Although in our test we observed these fluctuations to be small ($\approx 10^{-4}$), future work should average multiple calls or round each score to the nearest thousandth (to three decimal places) by considering further decimal places as noise.

Finally, this work leaves room to investigate a new perspective: **adaptive reformulation**. The demonstrated performance of our proposed reformulation method using LLM comes from applying the same fixed reformulation chain to every input query. However, developing an adaptive controller that inspects the input query and decides which reformulation to apply may offer a more robust and versatile reformulation method. Given the non-determinism and randomness involved in such an adaptive reformulation method, it would require thorough research to minimize potential side effects. However, LLMs have indeed been used for adaptive decision-making tasks in other domains before [32, 33], suggesting that a similar approach could offer value here. Pursuing this research angle would not only refine the robustness of child-centric information retrieval (IR), but also deepen our understanding of how LLMs can be used for multi-perspective IR workflows.

References

[1] D. Bilal and L.-M. Huang, "Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing," *Aslib Journal of Information Management*, vol. 71, no. 2, pp. 241–259, 2019.

[2] S. Duarte Torres and I. Weber, "What and how children search on the web," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 393–402.

[3] D. Bilal and J. Gwizdka, "Children's query types and reformulations in google search," *Information Processing & Management*, vol. 54, no. 6, pp. 1022–1041, 2018.

[4] D. Bilal, "Ranking, relevance judgment, and precision of information retrieval on children's queries: Evaluation of google, y ahoo!, b ing, y ahoo! k ids, and ask k ids," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 9, pp. 1879–1896, 2012.

[5] I. M. Azpiazu, N. Dragovic, M. S. Pera, and J. A. Fails, "Online searching and learning: Yum and other search tools for children and teachers," *Information Retrieval Journal*, vol. 20, pp. 524–545, 2017.

[6] M. van Kalsbeek, J. de Wit, R. B. Trieschnigg, P. van der Vet, T. W. Huibers, and D. Hiemstra, "Automatic reformulation of children's search queries," 2010.

[7] S. D. Torres, D. Hiemstra, I. Weber, and P. Serdyukov, "Query recommendation in the information domain of children," *Journal of the Association for Information Science and Technology*, vol. 65, no. 7, pp. 1368–1384, 2014.

[8] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, and J.-R. Wen, "Large language models for information retrieval: A survey," *arXiv preprint arXiv:2308.07107*, 2023.

[9] K. D. Dhole and E. Agichtein, "Genqrensemble: Zero-shot llm ensemble prompting for generative query reformulation," in *European Conference on Information Retrieval*. Springer, 2024, pp. 326–335.

[10] F. Ye, M. Fang, S. Li, and E. Yilmaz, "Enhancing conversational search: Large language model-aided informative query rewriting," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 5985–6006. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.398

[11] I. Madrazo Azpiazu, N. Dragovic, O. Anuyah, and M. S. Pera, "Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children," in *Proceedings of the 2018 conference on human information interaction & retrieval*, 2018, pp. 92–101.

[12] D. Bilal, "Children's use of the yahooligans! web search engine. iii. cognitive and physical behaviors on fully self-generated search tasks," *Journal of the American Society for information science and technology*, vol. 53, no. 13, pp. 1170–1183, 2002.

[13] A. Druin, E. Foss, L. Hatley, E. Golub, M. L. Guha, J. Fails, and H. Hutchinson, "How children search the internet with keyword interfaces," in *Proceedings of the 8th International conference on interaction design and children*, 2009, pp. 89–96.

[14] E. Foss, A. Druin, R. Brewer, P. Lo, L. Sanchez, E. Golub, and H. Hutchinson, "Children's search roles at home: Implications for designers, researchers, educators, and parents," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 3, pp. 558–573, 2012.

[15] O. Anuyah, J. A. Fails, and M. S. Pera, "Investigating query formulation assistance for children," in *Proceedings of the 17th ACM conference on interaction design and children*, 2018, pp. 581–586.

[16] B. J. Bettencourt, M. S. Pera, C. Kennington, K. L. Wright, and J. A. Fails, "Kid query: Co-designing an application to scaffold query formulation," in *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, 2024, pp. 828–833.

[17] M. S. Pera and Y.-K. Ng, "Using online data sources to make query suggestions for children," in *Web Intelligence*, vol. 15, no. 4. SAGE Publications Sage UK: London, England, 2017, pp. 303–323.

[18] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," *CoRR*, vol. abs/2311.07911, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.07911

[19] S. Sun, R. Yuan, Z. Cao, W. Li, and P. Liu, "Prompt chaining or stepwise prompt? refinement in text summarization," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 7551–7558.

[20] D. Bilal, "Children's use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks," *Journal of the American Society for information Science*, vol. 51, no. 7, pp. 646–665, 2000.

[21] J. He, P. Duboue, and J.-Y. Nie, "Bridging the gap between intrinsic and perceived relevance in snippet generation," in *Proceedings of COLING 2012*, 2012, pp. 1129–1146.

[22] K. Collins-Thompson, P. N. Bennett, R. W. White, S. De La Chica, and D. Sontag, "Personalizing web search results by reading level," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 403–412.

[23] M. S. Pera, E. Murgia, M. Landoni, T. Huibers, and M. Aliannejadi, "Where a little change makes a big difference: a preliminary exploration of children's queries," in *European Conference on Information Retrieval*. Springer, 2023, pp. 522–533.

[24] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring." *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

[25] M. J. Jennings, B. D. Zumbo, and J. F. Joula, "The robustness of validity and efficiency of the related samples t-test in the presence of outliers," *Psicologica*, vol. 23, no. 2, 2002.

[26] M. T. Shaikh, M. S. Pera, and Y.-K. Ng, "Suggesting simple and comprehensive queries to elementary-grade children," in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 1. IEEE, 2015, pp. 252–259.

[27] F. Yuan, S. Yuan, Z. Wu, and L. Li, "How vocabulary sharing facilitates multilingualism in llama?" *arXiv preprint arXiv:2311.09071*, 2023.

[28] D. Bilal and R. Ellis, "Evaluating leading web search engines on children's queries," in *Human-Computer Interaction. Users and Applications: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part IV 14*. Springer, 2011, pp. 549–558.

[29] Z. Li, Y. Shi, Z. Liu, F. Yang, N. Liu, and M. Du, "Quantifying multilingual performance of large language models across languages," *CoRR*, vol. abs/2404.11553, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.11553

[30] M. Adeyemi, A. Oladipo, R. Pradeep, and J. Lin, "Zero-shot cross-lingual reranking with large language models for low-resource languages," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 650–656. [Online]. Available: https://doi.org/10.18653/v1/2024.acl-short.59

[31] N. Chirkova and V. Nikoulina, "Zero-shot cross-lingual transfer in instruction tuning of large language models," in *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024, Tokyo, Japan, September 23 - 27, 2024*, S. Mahamood, M. L. Nguyen, and D. Ippolito, Eds. Association for Computational Linguistics, 2024, pp. 695–708. [Online]. Available: https://doi.org/10.18653/v1/2024.inlg-main.53

[32] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68 539–68 551, 2023.

[33] Z. Zhang, M. Fang, and L. Chen, "Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 6963–6975. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-acl.415

A Extra Visualizations

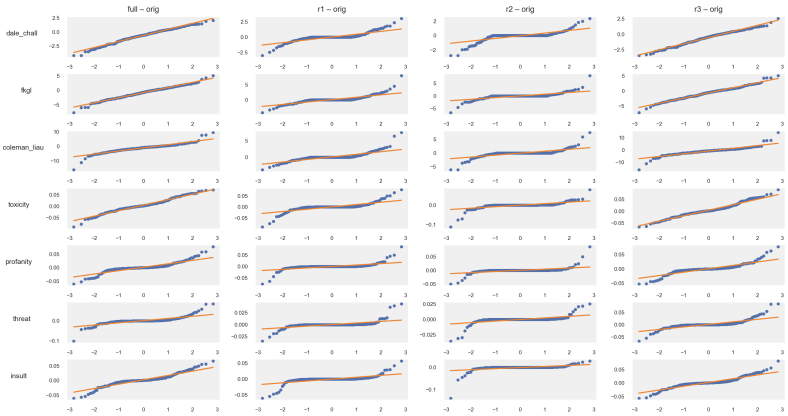


Figure 6: Q-Q plots of difference distributions for every metric (rows) and reformulation variant (columns). Deviations from the diagonal line indicate variations from normality.

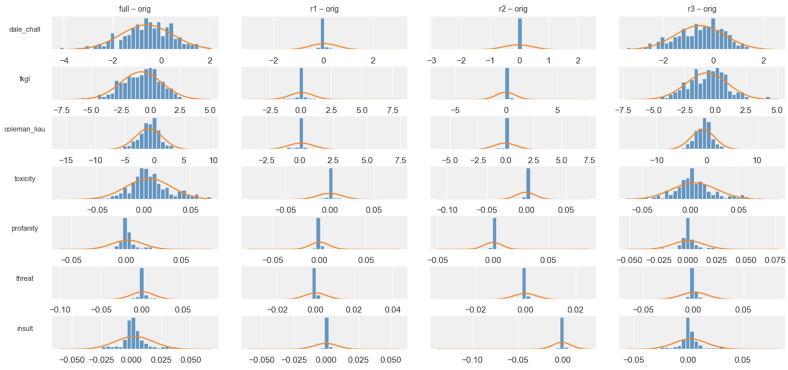


Figure 7: Histograms of the difference distributions for every metric (rows) and reformulation variant (columns) with over-laid fitted normal densities (orange).

B Prompts Used for AI

We listed below the prompts we used for AI.

Generating boilerplate Python statistical analysis code: We used it twice in this regard, (i) once for the significance tests and visualization generations (**The prompt we used**: "write me a Python statistical analysis component that given my attached dataset and project plan (and the description of planned statistical procedure found in that project plan), it analyses statistical significance of our model (full vs original) and impact of each ablation (r1,r2,r3 vs orig). "), and (ii) once for a more detailed analysis towards the r_1 and r_2 ablations (**The prompt we used**: "given my attached already existing statistical analysis code and dataset, give me the python code for the following analysis: for the queries that were changed after applying reformulation (r1

and r2 separately and respectively), what percentage of the results (here I don't mean for a query instance, but for all these filtered queries, what percentage of them have changed result snippet sets) were still the same and what was the sample median difference for these instances as well as the p value for that.").

Generating LaTeX tables and a `tikz` diagram: We also used it to generate the two result tables (Table 2 and Table 3). **The prompt we used for Table 2:** "given the attached results, give me latex code to generate a table with these information (columns would be "Metric", "query variant vs original", "p_val", and "sample median difference". use multirows for each comparison belonging to a metric (4 comparisons for a given metric)". **The follow-up prompt we used for Table 3:** "now similarly generate me a different table with using the results I newly attached". Additionally, we used it to generate the LLM reformulation pipeline diagram (Figure 1). While the code it generated did not work as a drop-in solution, it was a starting point for the diagram that we then fixed and improved on. **The prompt we used for Figure 1:** "given my attached project plan, generate me a tikz diagram showing the LLM reformulation pipeline."