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

Atilla Colak, Professor: Sole Pera, Supervisor: Hrishita Chakrabarti

Introduction and Background

Problem:

- For online search, children write queries that are *short*, *misspelled*, and often *underspecified* [1, 2].
- Common search engines are not built with children in mind [3].
- Thus, non-optimal child queries lead to child-inappropriate results (web results of advanced or unsafe language) [3].

Similar Existing Efforts: Reformulating children's search queries so that the retrieved results are more child-friendly.

Some previously explored strategies:

- spelling and grammar correction [4].
- number word expansion ("w8" = "wait")[4].
- "for kids" keyword expansion [5].

Gap:

- single-perspective approach to reformulation (missing out on the benefit of other perspectives).
- limited "multi-perspective" reformulation research [4, 6].

Motivated by this gap: **multi-step query reformulation using LLM (*Gemini 2.5-Flash Model*)** [7].

Research Question

To what extent can a multi-step query reformulation using LLM impact the readability and content-safety of retrieved results for a given child query?

Methodology

Reformulation Method:

- multiple strategies are chained instead of passing all in a single big prompt (to reduce hallucination risk) [8].
- chosen reformulation strategies were shown to be promising [4, 5].
- chose reformulation strategies with a lower risk of semantic meaning change.
- system constraint to minimize the risk of hallucination.
- model temperature of 0 to make LLM outputs more deterministic.

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.

Figure 1. Rules (r) and output constraints (c) for the LLM

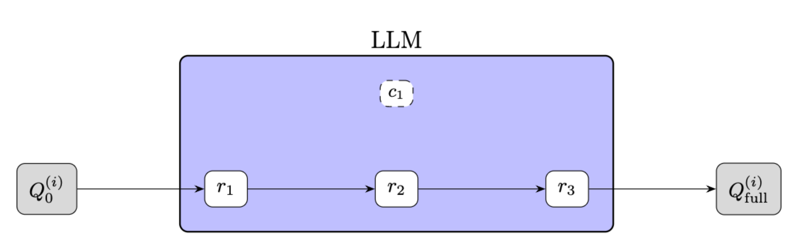


Figure 2. Multi-step query reformulation pipeline.

Experimental Setup

We use the `Children-Queries` dataset comprising 301 English queries typed by children aged 6-13.

Experiment Pipeline:

- run original, fully-reformulated, and single-rule reformulated queries (ablations) on Brave Search API.
- compute seven metrics for each query (based on retrieved top-10 web result snippets).
- test how reformulated results contrast with original query results and the impact of individual reformulation rules through ablations.

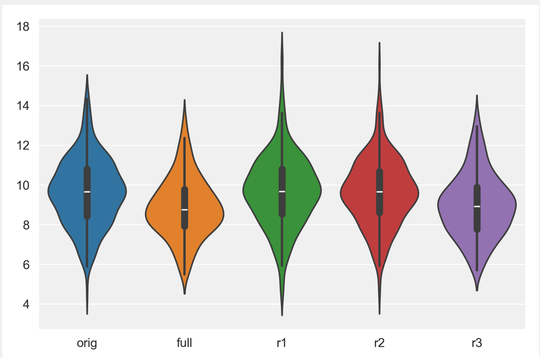
Evaluation Metrics:

- **Readability:** Flesch-Kincaid Grade Level (**FKGL**), Coleman-Liau, and Dale-Chall scores estimate how easy text is to read — lower is simpler.
- FKGL focuses on sentence and word length; Dale-Chall highlights hard vocabulary; Coleman-Liau uses character-level stats.
- **Content Safety:** Uses Perspective API to detect TOXICITY, PROFANITY, THREAT, and INSULT — each as a 0–1 risk score.
- Safety scores model nuanced harm beyond profanity (e.g., insults vs. threats), enabling fine-grained child-safe assessments.

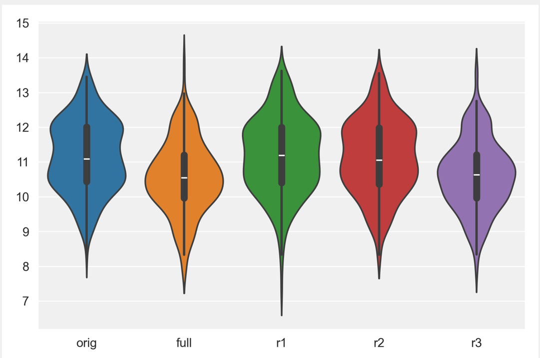
Results

Summary of the results:

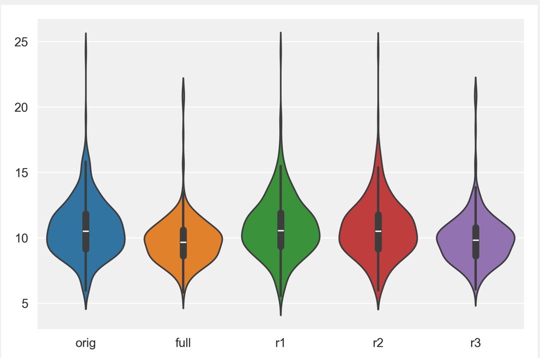
- Full multi-step reformulation significantly improves readability across all metrics (avg. 0.5–0.7 grade levels).
- The "for kids" rule (r_3) gives the biggest individual boost, but combining all three rules works best.
- r_1 and r_2 show no readability gains on their own.
- Slight increase in content risks (e.g., +0.003 in toxicity), but impact is negligible and still very low overall (most extreme outlier < 0.35).
- Reformulated results are easier to understand without compromising safety.



(a) FKGL



(b) Dale-Chall



(c) Coleman-Liau

Figure 3. Readability distributions across query variants. Lower = easier reading; White dots = medians; thick bars = IQR.

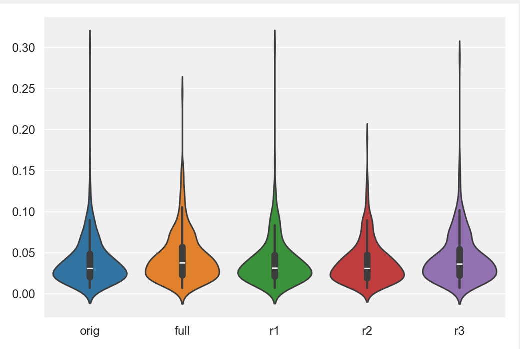


Figure 4. Toxicity attribute distribution across query variants. Lower = less likely toxic; White dots = medians; thick bars = IQR.

Responsible Research

We followed key ethical and reproducibility practices throughout our study:

1. The dataset contains no personal or identifiable information and is IRB-approved.
2. All code, prompts, and intermediate results (reformulation outputs, collected web results, and metrics) are documented and made openly available for transparency and reproducibility.

Conclusion and Future Work

Our results show that chaining spelling/grammar correction, synonym substitution, and the "for kids" expansion inside an LLM chain reduces the reading grade of top-10 search snippets on average by 0.5-0.7 levels.

Potential Design Implications for Info Access Systems:

- **Client-side deployment:** The pipeline can run as a browser extension or school proxy, avoiding the need for a standalone search engine.
- **Multilingual potential:** LLM prompts can be adapted with language tags to support non-English queries with minimal tuning.

Limitations & Future Work:

- Results may not generalize across search engines beyond Brave.
- Safety scores rely on a single API call; averaging or smoothing may improve robustness.
- Relevance of reformulated results was not evaluated—future work can include relevance metrics.
- Exploring adaptive, query-specific reformulation chains is a promising next step [9].

Reflection on RQ:

- Our work answers the research question affirmatively: multi-step reformulation with LLMs can enhance both readability and (minimally impact) safety for children's search queries.

References

[1] Sergio Duarte Torres and Ingmar Weber. What and how children search on the web. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 393–402, 2011.

[2] Dania Bilal and Jacek Gwizdka. Children's query types and reformulations in google search. *Information Processing & Management*, 54 (6):1022–1041, 2018.

[3] Dania 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*, 63(9):1879–1896, 2012.

[4] Maarten van Kalsbeek, Joost de Wit, Rudolf Berend Trieschnigg, PE van der Vet, Theo WC Huibers, and Djoerd Hiemstra. Automatic reformulation of children's search queries. 2010.

[5] Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation in the information domain of children. *Journal of the Association for Information Science and Technology*, 65(7):1368–1384, 2014.

[6] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad 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*, pages 92–101, 2018.

[7] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. Enhancing conversational search: Large language model-aided informative query rewriting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5985–6006. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.398. URL https://doi.org/10.18653/v1/2023.findings-emnlp.398.

[8] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023. doi: 10.48550/ARXIV.2311.07911. URL https://doi.org/10.48550/arXiv.2311.07911.

[9] Zihan Zhang, Meng Fang, and Ling Chen. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6963–6975. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.415. URL https://doi.org/10.18653/v1/2024.findings-acl.415.

A.A.Colak-1@student.tudelft.nl