Enhancing Children's Web Searches through Age-Specific Vocabulary Reformulation

An emperical study assessing the effects on Readability and Education Relevance

Rembrandt Hazeleger
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 21, 2025

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

Abstract

Children increasingly rely on web search engines to support their learning and exploration. However, conventional search systems are not optimised for their developmental stage, often returning information that is linguistically complex or educationally irrelevant. The retrieved results are often written at a higher reading grade level than children can easily comprehend, resulting in poor engagement and learning outcomes.

This research investigates whether substituting simpler vocabulary into search queries can improve the educational relevance and readability of retrieved web content. We develop a reformulation pipeline consisting of: (1) a rule-based method that substitutes key query terms with synonyms ranked by Age of Acquisition (AoA) scores, and (2) a computational intelligence approach that uses a Large Language Model (LLM) to generate child-friendly rephrasings. The results retrieved from the queries are evaluated across two dimensions: readability and educational relevance.

Our results show that rule-based reformulations improve readability, but retrieved results stayed consistent in terms of educational relevance. LLM-based reformulations enhance educational relevance; however, they don't improve readability. This trade-off highlights the complementary strengths of both methods and underlines the potential of direct query reformulation to make web search more accessible and educationally effective for children.

1 Introduction

In an increasingly digital world, the ability to retrieve valuable and age-appropriate online information is critical for supporting children's learning and autonomy. However, current web search engines are not designed with young users in mind, creating barriers to educational access and digital literacy. These difficulties arise from developmental differences in children's reading abilities, attention span, and language use [1]. While adults tend to use precise keyword queries, children are more likely to phrase their information needs in natural-language form [2, 3], This variability can reduce retrieval accuracy [4].

Although previous work has investigated interface adaptations [5] and suggestion mechanisms [6, 7], relatively little attention has been paid to the formulation of the query itself. However, subtle changes in query language can substantially influence search outcomes [4]. This makes the query itself an interesting intervention point, as even small modifications in query phrasing may help surface content that is more accessible, addressing the challenge that children's queries often yield results written above their comprehension level [8]. To support more effective and inclusive web search for children, this work explores how query reformulation can be used to retrieve results that are both easier to understand and educationally relevant.

Our approach aims to retrieve child-appropriate content by substituting keywords with simpler synonyms. We hypothesise that embedding simpler synonyms into queries will enhance both the readability and the educational relevance of retrieved web content.

This leads us to our central research question:

> *To what extent does incorporating age-specific vocabulary into search queries improve the retrieval effectiveness of age-appropriate educational web content?*

To explore this question, we design a query reformulation pipeline. The first method is rule-based and replaces high-complexity terms using synonyms ranked by Age of Acquisition (AoA) scores [9]. This technique prioritises transparency and allows direct control over word choice within the query. The second method uses a Large Language Model (LLM, GPT-3.5-turbo) to rephrase queries using simpler vocabulary. This LLM-based approach offers broader contextual awareness and expressive variation [10], making it well-suited for capturing nuances in natural language and generating rewordings that preserve meaning while improving accessibility.

We evaluate both reformulation strategies using two metrics: **readability**, assessed via Flesch-Kincaid and Spache formulas, and **educational relevance**, which is defined by the degree to which a web resource's content reflects the knowledge and topics found within established K-12 educational standards [11]. This is scored by BiGBERT [12], a transformer-based classifier trained on K-12 curricula. The Brave Search API provides the top-5 results for each query to reflect the reality that children tend to only examine the first few search outcomes [2].

Our findings reveal a trade-off: rule-based reformulations enhance readability, while LLM-based reformulations increase educational relevance. These insights suggest that child-specific query rewriting can significantly improve the readability and educational relevance of retrieved search results.

2 Related Work

In recent years, children have become increasingly reliant on web search engines for educational purposes [13]. However, existing systems are not well adapted to their specific needs. A study analyzing 300 search results for queries submitted by children revealed that over 90% of retrieved pages exceed age-appropriate readability levels, based on Flesch-Kincaid and Flesch Reading Ease scores [8]. This discrepancy is compounded by children's limited ability to formulate effective queries, often due to their limited vocabulary and frequent spelling mistakes [14]. These issues suggest an urgent need to better align search technology with the cognitive development of young users.

The impact of query formulation on search results has been highlighted in recent work by Pera et al. [4], who examined 345 queries produced by children aged 9 to 11. Their findings show that even small variations in queries can lead to significant differences in the top-ranked results, affecting not only relevance but also readability and emotional tone. This underscores the importance of directly refining queries to enhance the accessibility of retrieved content.

Existing work has addressed this challenge through solutions that aid the user externally, such as interface adaptations [5] and query suggestion systems [6, 7]. However, relatively little research has explored direct query reformulation as a means of improving result quality.

An effective approach is the ReQuIK system by Madrazo Azpiazu et al. [7], which generates alternative query formulations and ranks them according to child-friendly criteria such as vocabulary frequency and readability. Although effective, ReQuIK operates externally to the user's input and does not directly alter the original query. In contrast, our approach focuses on internal query reformulation to ensure better lexical alignment with child readers. This internal method allows the query itself to become a site of intervention, helping search engines retrieve content that is more immediately suitable without relying on post-hoc reranking. Such direct intervention is especially beneficial in child-facing systems, where transparency and immediate feedback are critical [15].

This work addresses the intersection of direct query reformulation and child-centred information retrieval by investigating how query-level modifications can improve the readability and educational relevance of web search results for children.

3 Methodology

This chapter details the approach used to develop and evaluate our query reformulation pipeline for children's search queries. It is structured in two parts. First, we describe the design of the reformulation strategy, including the resources and methods used to simplify queries. Then, we outline the experimental setup used to assess the effectiveness of the proposed strategies.

3.1 Query Reformulation Strategy

To address the research question effectively, we implement two complementary reformulation strategies: a transparent, rule-based method and a context-aware [16] LLM-based method (see Figure 1).

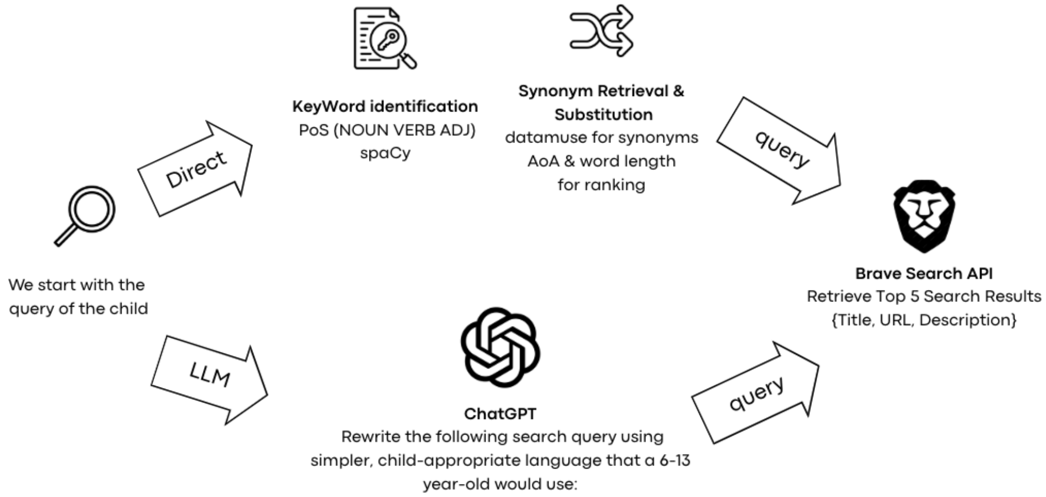
We start with the query of the child
Direct
KeyWord identification
PoS (NOUN VERB ADJ)
spaCy
Synonym Retrieval & Substitution
datamuse for synonyms
AoA & word length
for ranking
query
Brave Search API
Retrieve Top 5 Search Results
{Title, URL, Description}
LLM
ChatGPT
Rewrite the following search query using simpler, child-appropriate language that a 6-13 year-old would use:
query

Figure 1: Overview of the dual-method query reformulation pipeline

3.1.1 Age of Acquisition Ratings

To guide lexical simplification, we employ the Age of Acquisition (AoA) Ratings (2024) dataset, which includes 51 715 English words annotated with the average age (in years) at which native speakers acquire each term [9]. The ratings are based on multiple-choice testing of U.S. schoolchildren, parental reports of toddler vocabulary, and adult retrospective judgments. We use AoA scores to rank synonyms by simplicity.

3.1.2 Rule-Based Simplification

The rule-based method focuses on transparent lexical simplification by substituting complex terms with simpler synonyms. This rule-based method offers strong explainability, making it suitable for applications where interpretability is a requirement [17].

Preprocessing Queries are normalised using standard / well preforming text cleaning techniques:

- Conversion to lowercase [18]
- Removal of punctuation [19]
- Elimination of redundant whitespace [20]

To ensure that synonyms can be reliably identified, we first correct spelling mistakes using the `pyspellchecker` Python package.

Keyword Identification Following Wang et al. [21], we apply a rule-based keyword extraction approach designed for short queries. Each query is processed using the `spaCy` library [22] to obtain part-of-speech (PoS) tags.

We retain tokens that meet the following criteria:

- PoS tag is `NOUN`, `VERB`, or `ADJ`
- Token is not a stopword

Wang et al. [21] demonstrate that part-of-speech information is a strong indicator of keyword relevance in natural language queries. In particular, their rule-based and machine-learned selection mechanisms focus on identifying key phrase-level structures where nouns, verbs, and adjectives frequently convey the core intention of the query. Therefore, we specifically retain `NOUN`, `VERB`, and `ADJ` tokens, which are most likely to encapsulate the user's intent.

Furthermore, stopwords are excluded because they are high-frequency function words that typically do not contribute meaningful content to the query (e.g., "the", "is", "how"). As noted by Wang et al., stopwords often introduce noise in keyword extraction from short queries, even though they can be nouns, verbs, and adjectives.

For example, the query *"How do volcanoes erupt?"* yields the keywords: `["volcanoes", "erupt"]`.

Synonym Substitution For each keyword:

1. Retrieve synonyms via the Datamuse API [23].
2. Rank candidates by lowest AoA score (tie-break: shorter length).
3. Substitute the top synonym.

Oshika et al. [24] show that replacing late-acquired words with earlier-acquired synonyms significantly improves the readability of translated texts without compromising meaning. This motivates our use of Age of Acquisition (AoA) for synonym ranking. In case of ties, we prefer shorter synonyms, as word length has been shown to be a strong predictor of readability [25].

3.1.3 LLM-Based Rewriting

To complement the deterministic approach, we use a computational intelligence reformulation method based on ChatGPT (GPT-3.5-turbo), which has demonstrated state-of-the-art performance in rewriting tasks [26].

This method is particularly suited for rephrasing full sentence queries. A pattern commonly observed in how children express information needs [2, 3].

Prompt Engineering Each original query is embedded into a prompt instructing the model to rewrite it for a child audience (ages 6-13). The prompt used for this research is:

```
"Rewrite the following search query using simpler, child-appropriate
language that a 6-13 year old would use:  'sample query'"
```

Execution Procedure Queries are submitted using the official ChatGPT Library [27]. We use the `GPT-3.5-turbo` model with the following parameters:

- **model (gpt-3.5-turbo)**: Model ID used to generate the response.
- **input ("Rewrite the following search query using simpler, child-appropriate language that a 6-13 year old would use: query")**: Text, image, or file inputs to the model, used to generate a response.

The outputs are stored for downstream analysis and comparison.

3.1.4 Retrieval Procedure

Each original and reformulated query (rule-based and LLM-based) is submitted to the Brave Search API using Python `requests`. We specify the following parameters:

- **q (URL-encoded query)**: The user's search query term
- **safesearch (off)**: Filters search results for adult content. off = No filtering is done
- **result_filter (web)**: A comma delimited string of result types to include in the search response.

We are retrieving the top five results for each query. This is motivated by research showing children's strong preference for the first few results [2]. For each query, we parse the titles, URLs, and descriptions of the top five web results.

3.2 Experimental Design

This section outlines the procedure used to evaluate the impact of query reformulation on readability and educational relevance of search results.

3.2.1 Child Query Dataset

We begin with the Madrazo-Azpiazu et al. dataset (2018) [28], comprising 301 search queries submitted by 97 children aged 6-13 years (CC-BY-NC-ND). These queries form the basis of our empirical evaluation.

3.2.2 Assessment Metrics

To quantify the effectiveness of the reformulations, we compute the readability and educational relevance of the retrieved web results

Readability

- *Flesch-Kincaid Grade Level (FKGL)* [29] - Estimates the U.S. school grade level required to understand a given text. It is computed as:

  $$\text{Grade Level} = 0.39\left(\frac{\text{words}}{\text{sentences}}\right) + 11.8\left(\frac{\text{syllables}}{\text{words}}\right) - 15.59$$

  This formula considers both syntactic complexity (words per sentence) and lexical difficulty (syllables per word) to evaluate how challenging a text is to read.

- *Spache Readability Formula* [30] - Also provides an estimate of the U.S. grade level, but is made for texts written for children up to fourth grade. Its formula is:

  $$\text{Grade Level} = 0.141\left(\frac{\text{words}}{\text{sentences}}\right) + 0.086 \times (\text{percentage of unfamiliar words}) + 0.839$$

  The Spache formula specifically considers whether words fall outside a predefined list of familiar vocabulary, making it more sensitive to language unfamiliar to younger readers.

These metrics were chosen for their complementary strengths. FKGL is a general-purpose readability formula, useful for benchmarking across a wide range of texts. Spache is developed for younger readers and reflects the vocabulary and sentence structures typical for children aged around 10. Using both allows us to evaluate readability from both a general and child-specific perspective.

Educational Relevance To assess whether the retrieved results are suitable for an educational context, we apply *BiGBERT* [12], a BERT-based classifier fine-tuned to identify alignment with K-12 educational standards. The model evaluates web resources and determines whether the content is pedagogically appropriate. The motivation for this metric stems from the growing role of web search in children's learning journeys. As children increasingly rely on search engines for schoolwork and self-guided learning [13], it becomes critical that the content they retrieve is not only readable, but also educationally relevant.

3.2.3 Experimental Design

Our empirical comparisons proceed as follows:

1. Parse titles, URLs, and descriptions of the top five web results.

2. Compute average FKGL and Spache scores on result descriptions.

3. Compute average BiGBERT educational relevance based on the URL and description of the results.

4. Compare the performance of the reformulation methods by conducting paired t-tests on the average scores (FKGL, Spache, and BiGBERT) calculated over the top-5 search results retrieved via the Brave Search API for each query.

4 Results

We evaluate the efficacy of the proposed query reformulation strategies across two dimensions: **readability** and **educational relevance**. Three conditions were compared: (1) original queries, (2) Rule-Based Reformulation using synonym substitution, and (3) LLM-based Reformulation. For all statistical comparisons, a significance level of ($p < 0.05$) was chosen. Statistical significance was assessed via pairwise t-tests, with results visualized.

4.1 Readability Evaluation

Figure 2 presents the readability score distributions for original, Rule-Based, and LLM-based queries, evaluated using both the Flesch-Kincaid Grade Level (FKGL) and Spache indices.

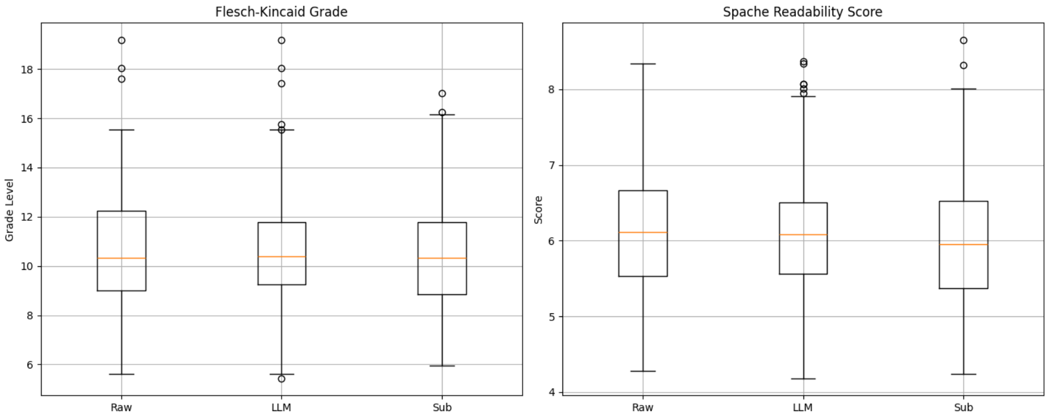

Figure 2: Distribution of Readability Scores across FKGL and Spache metrics

Spache Readability revealed a statistically significant improvement for Rule-Based over Raw search results, suggesting that synonym substitution improves linguistic simplicity. A marginal difference was found between Rule-Based and LLM-based outputs, with Rule-Based result descriptions tending to yield slightly lower grade levels.

No significant differences were observed in FKGL scores across the results retrieved by the queries.

Notably, the LLM-based reformulations exhibited a narrower score distribution, indicating reduced variance in lexical complexity.

Table 1 reports the pairwise comparisons for the readability metrics.

Metric	**Comparison**	**t-statistic**	**p-value**
FKGL	Raw vs LLM	0.066	0.948
FKGL	Raw vs Rule-Based	0.606	0.545
FKGL	Rule-Based vs LLM	-0.540	0.590
Spache	Raw vs LLM	0.496	0.620
Spache	**Raw vs Rule-Based**	**2.249**	**0.025**
Spache	Rule-Based vs LLM	-1.805	0.072

Table 1: Pairwise t-tests on Readability Metrics

4.2 Educational Relevance

To assess content alignment with educational objectives, we scored the top-five search results retrieved by each query variant using the BiGBERT classifier. Results are shown in Figure 3.

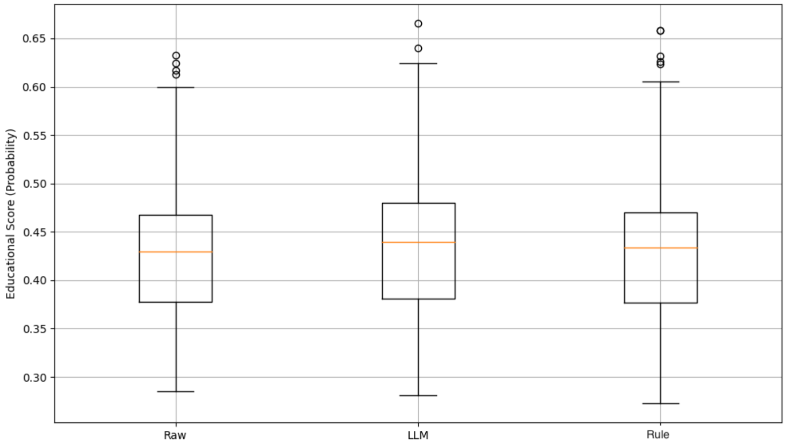

Figure 3: Predicted Educational Relevance of Query Results

The LLM-based reformulations retrieved results with significantly higher educational relevance compared to the original (raw) queries. This improvement likely stems from the model's ability to rephrase queries in a way that more effectively conveys the underlying educational intent, using phrasing patterns that retrieval systems associate with pedagogical content. However, there were no statistically significant differences in educational relevance between the rule-based and raw queries, nor between the rule-based and LLM-based queries, indicating that the LLM's advantage was most pronounced when compared directly to unreformulated input.

Table 2 shows the statistical comparison of educational relevance scores across conditions.

Comparison	t-statistic	p-value
Raw vs Rule-Based	-0.395	0.693
Raw vs LLM	**-2.559**	**0.011**
Rule-Based vs LLM	-1.327	0.186

Table 2: Pairwise *t*-tests on Educational Relevance Scores

5 Responsible Research

In conducting this project, careful attention was given not only to technical effectiveness but also to ethical integrity. Ensuring that the methods and outcomes align with responsible research practices was a guiding priority throughout. To promote transparency and reproducibility, all relevant code and configuration files have been made publicly available.[1] The following sections outline the ethical safeguards taken and the reproducibility measures embedded in the system's design.

[1] https://github.com/rembrandtking/child-query-reformulation

5.1 Ethical Considerations

This research aligns with the five principles of the Netherlands Code of Conduct for Research Integrity [31]: **honesty**, **scrupulousness**, **transparency**, **independence**, and **responsibility**. These values guided both the methodological and ethical dimensions of this work.

The data set used, which contains search queries submitted by children, is publicly available and fully anonymised, thus respecting the boundaries of privacy and informed consent. No personally identifiable information was collected, stored, or processed during the project. The reformulated queries do not simulate or represent individual users but rather abstract linguistic modifications.

The analysis refrains from cherry-picking results or excluding negative outcomes, and all retrieval evaluations were performed automatically using pre-established metrics. No manual editing was applied to either LLM outputs or educational relevance scores, ensuring unbiased reporting.

All external sources, APIs, and datasets were appropriately credited, and every claim made in this report is grounded in reproducible experimental evidence.

5.2 Reproducibility

The reformulation system was explicitly designed with reproducibility in mind. Each stage of the pipeline can be independently validated:

- **Query preprocessing:** implemented in Python with version-controlled scripts.
- **Keyword extraction:** based on deterministic rules using `spaCy` PoS tagging.
- **Synonym substitution:** integrates Datamuse API, filtered by Age of Acquisition (AoA) scores and word length.
- **LLM-based reformulation:** queries submitted to ChatGPT (GPT-3.5-turbo) via batch API calls.
- **Retrieval and evaluation:** performed using Brave Search API and BiGBERT model.

All intermediate outputs (normalised queries, extracted keywords, reformulated variants) were logged to enable full traceability. Although complete query-result mappings are subject to search engine API terms, a sanitised version of the reformulation outputs and evaluation scores is available on the GitHub repository.

6 Discussion

This section interprets the outcomes of our reformulation strategies and situates them within the broader landscape of child-oriented information retrieval (IR). We reflect on the design implications for future search systems, examine ethical and societal consequences, and acknowledge the limitations of our study.

6.1 Interpreting Reformulation Outcomes

The results highlight a complementary relationship between the two reformulation strategies. The rule-based method produced statistically significant gains in readability, affirming prior work by Torres et al. [2], who found that aligning queries with children's vocabulary improves retrieval. Our use of AoA-based substitutions appears to follow a similar pattern, leveraging lexical simplification to bridge the gap between user input and age-appropriate content.

Conversely, the LLM-based method did not yield improved readability but significantly enhanced the retrieval of educationally relevant results. This mirrors observations from Madrazo Azpiazu et al. [7], whose ReQuIK system achieved comparable relevance gains through neural ranking models.

A more granular analysis of the results reveals nuanced characteristics that further distinguish the two approaches. The Spache metric detected the readability improvements introduced by rule-based reformulation, in contrast to FKGL, which remained largely unchanged. One possible explanation is that the FKGL formula was designed initially with adult technical readers in mind, particularly military personnel [32], making it less sensitive to vocabulary simplifications targeted at younger readers. In this context, Spache proved more appropriate for detecting age-relevant vocabulary shifts, strengthening the case for metric selection closely connected to the target audience.

Additionally, the LLM-based outputs demonstrated a notably narrower score distribution, indicating a more consistent lexical profile. This uniformity, while not necessarily leading to statistically superior average scores, may offer value in educational scenarios that benefit from predictable reading levels. It points to the LLM's inherent ability to regulate linguistic variance [33].

Although rule-based reformulations improved vocabulary simplicity, their limited contextual awareness may lead to awkward phrasings or unintended shifts in meaning [34]. This might account for the limited improvement in education-focused retrieval observed in our results. A potential explanation is that replacing keywords solely based on acquisition age does not always preserve the intended informational focus of the original query. Further investigation would be needed to confirm this effect.

The LLM-based method, by leveraging contextual understanding [35], produces reformulations that more accurately capture the user's intent, resulting in improved retrieval alignment with educational objectives. However, this advantage is accompanied by a potential trade-off in transparency and controllability, particularly relevant in child-focused applications where predictability and oversight are critical.

Together, these findings suggest that neither method singularly satisfies all performance criteria. Instead, they address distinct, complementary dimensions of child-oriented query reformulation. Integrating both techniques, using the rule-based method for controlled simplification and the LLM for semantic alignment, can yield a balanced and adaptable reformulation pipeline. Such a hybrid system could dynamically adjust reformulation strategies based on user age, educational intent, or domain context, thereby enhancing both readability and relevance.

6.2 Broader Societal Impact

Beyond technical considerations, these findings underscore a broader societal issue: current search engines, while universally accessible, do not serve all users equally. Children, especially early-stage readers, may be confronted with content that exceeds their comprehension abilities, not because of a lack of internet access, but due to lexical barriers embedded in query formulation and result ranking. Our study demonstrates that these barriers can be mitigated through lightweight interventions. In this way, reformulation techniques contribute not just to technical performance but also to digital equity in information access.

Nevertheless, care must be taken not to over-automate linguistic support. If input is too simplified or lacks lexical variety, it may fail to foster productive vocabulary development [36]. Future systems should therefore offer graduated levels of reformulation, adapting dynamically to the user's linguistic progress and learning context.

Finally, while large language models like ChatGPT show strong performance in generating age-appropriate reformulations, their use in educational contexts raises practical concerns. These include limited transparency, potential prompt sensitivity, and lack of explicit control over readability levels. Future work should explore mechanisms for guiding LLM reformulations through more structured prompts or controlled decoding to better meet specific developmental targets.

6.3 Limitations and Considerations

While the results demonstrate potential, several limitations have to be acknowledged. First, our use of the Brave Search API limits comparability with mainstream engines such as Google or Bing. These platforms employ other ranking algorithms that may respond differently to reformulated input.

Second, our evaluation of educational relevance is based on predictions from BiGBERT, which, while effective, does not assess subjective factors such as user engagement and perceived comprehension. These aspects are particularly important in educational contexts, where understanding and usability from the learner's perspective are central to effectiveness. Without human raters, the system's alignment with actual user needs remains uncertain.

Finally, while the AoA-based rule system offers transparency, it does not account for polysemy. A synonym selected solely based on age of acquisition may not reflect the intended meaning of the original term, especially for words with multiple senses. This semantic mismatch can result in reformulations that, although simpler and more readable, diverge from the user's original search intent. As our study primarily focused on readability and educational relevance, we did not explicitly assess how well reformulated queries preserved the informational goals of the original input. This leaves open the possibility that certain reformulations, despite scoring well on our metrics, may have reduced topical relevance.

7 Conclusions and Future Work

This research set out to answer the question: *To what extent does incorporating age-specific vocabulary into search queries improve the retrieval of age-appropriate educational web content?*

To investigate this, we implemented two complementary strategies: a rule-based approach that simplified language through synonym substitution based on Age of Acquisition data, and a neural approach that leveraged large language models to rewrite queries into child-directed phrasing.

Our findings indicate that both reformulation methods contribute positively, albeit in different ways. The rule-based strategy consistently simplified the language of retrieved content, improving its readability for younger audiences. This supports the broader notion that simplifying query vocabulary with easier terms can help bridge the gap between search intent and result comprehension. In parallel, the LLM-based approach yielded results that were more educationally aligned, suggesting that semantically richer, context-aware reformulations better capture the educational intent behind child-oriented information needs.

Interestingly, the outputs generated by LLMs also tended to exhibit more consistent language levels, potentially supporting more uniform learning experiences.

Taken together, these insights demonstrate that child-sensitive query reformulation can enhance search outcomes not just through simplification but also through improved semantic alignment. Rather than treating these strategies as mutually exclusive, their complementary strengths point to the value of hybrid approaches that combine lexical clarity with contextual intelligence. Collectively, the findings suggest that effective child-oriented search experiences depend not only on linguistic simplicity but also on sensitivity to educational context, an insight with broader design implications for intelligent retrieval systems.

These results point to several promising directions for future work. First, introducing user-centred evaluations would allow us to assess the perceived usefulness and clarity of reformulated queries in real-world educational settings. Involving children and educators in this process could reveal insights not captured by automated metrics alone, including aspects of engagement, comprehension, and trust.

Second, the development of a hybrid reformulation pipeline presents an opportunity to combine the interpretability of rule-based systems with the adaptive potential of LLMs. A layered design, where simple vocabulary replacements precede context-sensitive reformulation, might offer more reliable control over language complexity while still benefiting from the nuanced understanding LLMs bring to query intent.

In addition, future versions of the LLM-based approach could benefit from more targeted prompt design. While the current formulation aimed for general simplification, experimenting with prompts made for specific reading levels, learning objectives, or subject areas may yield more nuanced and pedagogically aligned reformulations.

Lastly, the modular architecture of the current pipeline lends itself well to adaptation. Future iterations could explore integration with various search engines, the use of more advanced language models, or other sets of queries formulated by children. These developments would support the continued evolution of search systems that cater to the informational needs of young users.

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*, CIKM '11, pages 393–402, New York, NY, USA, 2011. Association for Computing Machinery.

[2] Sergio Duarte Torres. *Information Retrieval for Children: Search Behavior and Solutions*. Phd thesis - research ut, graduation ut, University of Twente, Netherlands, February 2014. The PhD thesis was awarded with a cum laude degree.

[3] Dania Bilal and Jacek Gwizdka. Children's query types and reformulations in google search. *Information Processing & Management*, 54(6):1023–1041, 2018.

[4] Maria Soledad Pera, Emiliana Murgia, Monica Landoni, Theo Huibers, and Mohammad Aliannejadi. Where a little change makes a big difference: A preliminary exploration of children's queries. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Annalina Caputo, and Udo Kruschwitz, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Proceedings*, Lecture Notes in Com-

puter Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 522–533. Springer, 2023. Green Open Access added to TU Delft Institutional Repository 'You share, we take care!' - Taverne project https://www.openaccess.nl/en/you-share-we-take-care Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.; 45th European Conference on Information Retrieval, ECIR 2023 ; Conference date: 02-04-2023 Through 06-04-2023.

[5] T. Gossen. *Search engines for children: Search user interfaces and information-seeking behaviour*. 01 2016.

[6] Meher T. Shaikh, Maria Soledad Pera, and Yiu-Kai Ng. Suggesting simple and comprehensive queries to elementary-grade children. In *Proceedings of the 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Workshops*, pages 252–259, 2015.

[7] 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 *CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, pages 92–101, United States, February 2018. ACM. 3rd ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2018 ; Conference date: 11-03-2018 Through 15-03-2018.

[8] Dania Bilal. . comparing google's readability of search results to the flesch readability formulae: A preliminary analysis on children's search queries. *Proceedings of the 76th American Society for Information Science and Technology (ASIS&T) Annual Meeting*, 01 2013.

[9] Marc Brysbaert. English age of acquisition (aoa) ratings (kuperman et al., 2012). `https://osf.io/d7x6q`, 2024. Retrieved November 25, 2024.

[10] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Generative query reformulation for effective adhoc search. 08 2023.

[11] American International Accreditation Association for Schools & Colleges (AIAASC). Standards & Indicators: AIAASC K-12 Accreditation. Technical report, American International Accreditation Association for Schools & Colleges, 2021. Retrieved from AIAASC website.

[12] Garrett Allen, Brody Downs, Aprajita Shukla, Casey Kennington, Jerry Alan Fails, Katherine Landau Wright, and Maria Soledad Pera. Bigbert: Classifying educational web resources for kindergarten-12$^{\text{th}}$ grades. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, pages 176–184, Cham, 2021. Springer International Publishing.

[13] Carsten Eickhoff, Pieter Dekker, and Arjen P. de Vries. Supporting children's web search in school environments. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 129–137, New York, NY, USA, 2012. Association for Computing Machinery.

[14] Frans Sluis and E.M.A.G. Dijk. A closer look at children's information retrieval usage: Towards child-centered relevance. 01 2010.

[15] Eirini Mougiakou, Spyros Papadimitriou, and Maria Virvou. Intelligent tutoring systems and transparency: The case of children and adolescents. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8, 2018.

[16] Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. Can large language models understand context? In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[17] Zheng Tang and Mihai Surdeanu. Interpretability rules: Jointly bootstrapping a neural relation extractorwith an explanation decoder. In Yada Pruksachatkun, Anil Ramakrishna, Kai-Wei Chang, Satyapriya Krishna, Jwala Dhamala, Tanaya Guha, and Xiang Ren, editors, *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 1–7, Online, June 2021. Association for Computational Linguistics.

[18] Sannidhi Shetty, Pratiksha Narasimha Nayak G, Pranamya Mady, Vaishnavi K Bhustali, and Chetana Hegde. English to kannada translation using bert model. In *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, pages 1–6, 2023.

[19] Harjit Singh and Ashish Oberoi. Pre-processing phase to develop an interface to query relational databases in punjabi language: Query normalization. 07 2018.

[20] Robert Miner and Rajesh Munavalli. An approach to mathematical search through query formulation and data normalization. In Manuel Kauers, Manfred Kerber, Robert Miner, and Wolfgang Windsteiger, editors, *Towards Mechanized Mathematical Assistants*, pages 342–355, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[21] David X. Wang, Xiaoying Gao, and Peter Andreae. Automatic keyword extraction from single-sentence natural language queries. In Patricia Anthony, Mitsuru Ishizuka, and Dickson Lukose, editors, *PRICAI 2012: Trends in Artificial Intelligence*, pages 637–648, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[22] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in python. `https://spacy.io`, 2020. Version 3.0+.

[23] Datamuse. Datamuse api. `https://www.datamuse.com/api/`, 2024. Accessed: 2025-05-08.

[24] Masashi Oshika, Makoto Morishita, Tsutomu Hirao, Ryohei Sasano, and Koichi Takeda. Simplifying translations for children: Iterative simplification considering age of acquisition with LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8567–8577, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[25] Juan Yu. A readability study and its relevance to simplification on translations of lun yu. *Studies in Literature and Language*, 9(3):47–57, 2014.

[26] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980, 2024.

[27] OpenAI. Openai python api library. `https://github.com/openai/openai-python`, 2024. Accessed: 2025-06-10.

[28] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. Dataset for looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. `https://scholarworks.boisestate.edu/cs_scripts/5/`, 2018. Computer Science Faculty Scripts and Data, Boise State University. DOI: `https://doi.org/10.18122/B2WQ5T`.

[29] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Barry S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, Naval Air Station Memphis, Research Branch Report 8-75, Millington, TN, 1975.

[30] George Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 1953.

[31] Netherlands Federation of University Medical Centres, Association of Universities in the Netherlands, Royal Netherlands Academy of Arts and Sciences, Netherlands Organisation for Scientific Research, and TO2 Federation. Netherlands code of conduct for research integrity. `https://www.nwo.nl/en/netherlands-code-conduct-research-integrity`, 2018. This is the English translation of the original Dutch document.

[32] Glenda M. McClure. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15, 1987.

[33] Ariel Rosenfeld and Teddy Lazebnik. Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard, 2024.

[34] Islam Nassar, Michelle Ananda-Rajah, and Gholamreza Haffari. Neural versus non-neural text simplification: A case study. In Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay, editors, *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 172–177, Sydney, Australia, 4–6 December 2019. Australasian Language Technology Association.

[35] Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. Llms' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements, 2024.

[36] Eva Thue Vold. Development of lexical richness among beginning learners of french as a foreign language. *Nordic Journal of Language Teaching and Learning*, 10:182–211, 01 2023.