How Child-Friendly is Web Search? An Evaluation of Relevance vs. Harm

Maik Fröbe, Sophie Charlotte Bartholly, and Matthias Hagen

Friedrich-Schiller-Universität Jena

Abstract. Today's children grow up with easy access to the Web and respective search engines. However, when children try to inform themselves about current conflicts, crimes, etc., general web search engines might show relevant but inappropriate content. Instead, search engines specifically aimed at children as searchers try to filter out inappropriate content or even operate on much smaller, manually curated indexes of appropriate documents only. To understand whether this is effective, we compare three general and three child-oriented web search engines based on a new German evaluation corpus of 50 queries spanning personal, political, educational, and entertainment information needs of children. For each query, we annotate the search engines' top-10 results with respect to relevance and potential harm—a child-friendly result should be relevant but harmless. Our comparison shows that the child-oriented search engines effectively remove potentially harmful documents, while the general web search engines return more relevant documents at the expense of significantly more potentially harmful content.

Keywords: Child-Oriented Search; Retrieval Evaluation

1 Introduction

Around 70 % of children between 6 and 13 years in Germany use the web [17], with web search among the most popular applications [27]. Many parents use protections so that children are not exposed to unsuitable content [27] and search engines developed specifically for children as searchers aim to provide a safe environment of only non-harmful content [16,21]. Still, search engines like Google are more popular among children [3], motivating us to conduct a Cranfield-style comparison [11,12] of three general and three child-oriented search engines to understand the current status of web search's "child-friendliness".

Children search differently than adults [18], having more difficulties formulating their queries [28], easily losing focus and getting overstimulated by too much content [18], and being even more highly biased to only examining the top-ranked results [29]. Different to standard IR evaluation, child-oriented search results must not only be relevant, but also be appropriate for children [15]. We thus create a new evaluation corpus of 50 queries covering personal, political, educational, and entertainment information needs of children who recently started school [17]. The queries purposely cover information needs that might be difficult to appropriately present to children, e.g., on wars, climate change, sexuality, etc. We annotate the top-10 results of three commercial and three child-oriented web search engines with respect to relevance and potential harmfulness for children. Our experiments are in German and we publish our queries, judgments, and the collected search results as the Kid-FRIEND dataset, including DeepL-based translations to English (manually spot-checked to ensure a reasonable quality).[1]

An analysis of the judgments in our Kid-FRIEND dataset shows that the commercial search engines provide more relevant results but they also more frequently display content that is potentially harmful for children. In contrast, child-oriented search engines indeed provide safer results but these are also often less relevant—which exemplifies that there currently seems to be a trade-off between relevance and child-friendliness. With our work, we aim to follow recent position papers calling for a more comprehensive examination of the impact of retrieval and recommendation systems on children [21,32]. Our analyses and resources are meant to enable further research in this direction.

2 Related Work

We briefly review research on childrens' search behavior and information needs, on child-appropriateness, and on existing search engines for children.

Search Behavior of Children. Children usually learn to read and write at age five to six [27] but search engines and how to properly use them is rarely part of childrens' education [26]. Search queries of children often contain spelling errors [30,26], do not mention some important keywords [30,26] even though children often formulate long queries in natural language [29]. Children often formulate many queries in a session but they rarely use query suggestions [31], probably as typing requires high concentration for children so that they usually look at the keyboard [18]. Children are less focused on the search engine result pages, resulting in more "random" clicks and more clicks on ads [18,26,29].

Information Needs of Children. Broder [7] categorized information needs as (1) informational, (2) navigational, or (3) transactional. Children of age 5 to 11 years mostly submit transactional queries on games, while teenagers between 12 to 15 years mostly search for social media [31]. Many children also search to conduct schoolwork or to support buying decisions [17].

Criteria for Appropriate Content for Children. Coleman [13] identified four classes of content not appropriate and potentially harmful for children: (1) aggression (e.g., violence, harassment), (2) explicit sexual content (e.g., pornography, sexting), (3) values (e.g., hate speech), and (4) commercial content (e.g., advertisements). Exposure to such inappropriate content might unsettle and frighten children [23,27]. A recent survey found that 6 % of children had observed depictions of violence, 4 % problematic advertising, and 3 % pornography [17].

[1] Code and data: github.com/webis-de/ECIR-25 and zenodo.org/records/14684724

Table 1. Overview of our 50 topics in the 4 domains and the potential harm types (classes by Coleman [13]) they might expose (✓) or not expose (✗) to children.

Domain	**#Topics**	**Aggression**	**Sexual**	**Values**	**Commercials**
Education	15	✗	✓	✓	✓
Entertainment	11	✓	✓	✗	✓
Personal	14	✗	✓	✗	✓
Political	10	✓	✗	✓	✗

Child-Oriented Search Engines. There are multiple search engines for children as searchers, with Yahoo! Kids being the first public one from 1996 [6]. Yahooligans was a search engine for children that was frequently studied [5,6] and the PuppyIR project provided a research framework on building retrieval interfaces for children [22,14]. Still, most children use Google (91 %), even while many (48 %) are aware of child-oriented alternatives [3].

3 An Evaluation Dataset for Web Search Kid-Friendliness

The Kid-FRIEND dataset is made up of 50 topics and pools the top-10 results of six search engines annotated for harm and relevance in German language including DeepL-based translations to English. We describe the dataset creation including the construction of topics and their domains, the document corpus, and the assessment of relevance and harm.

Topic Creation. We manually create 50 topics in German language in personal, political, entertainment, and educational domains and formulate them so that all potential harms defined by Coleman [13] are covered. Our topics are in the TREC-format [19], where each topic consists of a title (what searchers would submit to the search engine), a description (what searchers actually mean), and a narrative (describing what makes documents relevant and harmful). We built our topics to mimic the interests of a child up to grade 4. Gaming and social media queries are among the most popular searches of children [31], and we include such information needs into our entertainment category. As their bodies change while growing up, children also specifically search for those changes [27] that we include into our personal category. Current political topics, e.g., conflicts, wars, and crimes, may expose children unintentionally to inappropriate content [27]; we include such information needs in our political category. Additionally, children frequently use search engines for school work [17]; we include such topics in the school category. Table 1 provides an overview of our topics across the four domains and the potential harms that our topics might expose. For instance, the 14 topics from the personal domain that we formulate might expose inappropriate sexual or commercial content. Please note that the types of exposed harms are specific to our formulated topics, e.g., we do not have a political topic for which we think there is a risk of exposure to inappropriate sexual content, as we did not include such topics (e.g., an information need on abortion).

Document Corpus. We submit the titles of all 50 topics to three commercial search engines (Bing, DuckDuckGo, and Google) and three search engines for children (FragFinn, Helles-Koepfchen, and Seitenstark)[2] and crawl all top-10 documents as document corpus. We selected the commercial search engines as they have the biggest market share in Germany (Google has 91 %, Bing has 4.8 %, DuckDuckGo has 0.8 %; we had to exclude Ecosia with 0.9 % as their Cloudflare setup blocked automated access [1]). All three child-safe search engines operate small, manually curated indexes (i.e., domain-experts assess which documents to include, one search engine manually creates keyphrases for every indexed document). We use all search engines in their default setting, i.e., for the commercial search we do not activate potentially available flags for child-safe search to explicitly include documents that might be harmful for children into our corpus. We crawl the search engine result pages and the documents with Scriptor[3] [20] that aims to build reproducible web corpora. Scriptor uses a headless browser to render pages (loading javascript, images, etc.) and scrolls to the bottom of the page waiting until the page is fully loaded before making a snapshot that includes the rendered HTML and a screenshot of the page. We configure a threshold of one minute to render the web pages. A few pages could not be downloaded within this time limit (e.g., because of infinite scroll) or were not available anymore (e.g., because the index of some child-safe search engines were outdated), in those cases, we manually downloaded the pages with a browser using the save page functionality to download a complete page, using the Wayback machine[4] in case a page was not available anymore. We include the HTML in WARC files and the main content of documents extracted with resiliparse [4] and create an English version from deepl translations, yielding two versions of our document corpus.

Pooling for Relevance and Harm. To analyze if search engines allow for child-safe search, we annotate each of the retrieved top-10 documents for relevance and potential harm according to the dimensions of Coleman [13]. We use doccano [25] for our relevance annotators with two in-house native speakers as annotators. We annotate relevance with labels from 0 (not relevant) over 1 (the document is on the topic of the information need but does not directly answer it) to 2 (the information need is directly answered). For children, however, relevance also depends on the comprehensible communication of the information [33]. As it is particularly difficult for children to distinguish between credible and untrustworthy sources, information in the results should come from trustworthy sources, as children tend to trust blindly [26]. Whether something is suitable for children can be determined by two factors: Is it addressed to children? And is it safe for children or not harmful to them [15]. The simpler of the two questions, whether it is aimed at children, can be answered by distinguishing between "children's media" and "adult media" [15]. For annotating harm, we also provide labels from

[2] https://www.fragfinn.de/, https://helles-koepfchen.de/, https://seitenstark.de

[3] https://github.com/webis-de/scriptor

[4] https://web.archive.org/

0 (appropriate for children) over 1 (not appropriate and not potentially harmful) to 2 (harmful). For instance, documents labeled with a harm of 1 are too complex to be appropriate for children but do not contain content of any of the four harmful categories defined by Coleman [13], whereas any content from the harmful Coleman categories is labeled with a harm of 2. Our harm-labeling schema is inspired by the health misinformation tracks [2,8,9] that also assigned non-harmful content an harm of 0 so that smaller harm scores in the evaluation are better (contrary to relevance where higher scores are better, which is suitable to combine relevance and harm into combined measures [10]).

4 Evaluation

We compare the six search engines using our relevance and harm labels. An ideal child-safe search engine would show highly relevant documents (i.e., relevance score of 2) that have no potential harm for children (i.e., a harm score of 0).

Experimental Setup. We do not create own retrieval systems and only compare six search engines as black boxes of which we do not know about their internal workings. All search engine result pages were crawled from the same server from the country in the language of the information needs. The crawling was done without parallelism in around one week. We conduct significance tests contrasting the three child-safe search engines with Google as baseline for relevance and harm using Students t-test at a p-value of 0.05 with Bonferroni correction.

Evaluation Measures. We report precision at three and ten together with the reciprocal rank (RR) and the Scaled Discounted Cumulative Gain (SDCG). SDCG [24] is a variant of nDCG that assumes that a perfect ranking (i.e., all documents labeled with a score of 2) is possible, even when those documents are not labeled. Therefore, SDCG fits our scenario well, as we can assume that for every of our topics enough highly relevant documents exist on the web, even when they are not included in our corpus, identical for harm, where we can assume that for each topic also harmful documents exist on the web.

Results. Table 2 shows the results of our evaluation. All three child-safe search engines retrieve fewer relevant results then the commercial search engines. However, the commercial search engines also show substantially more results that are potentially harmful. Bing achieves the highest relevance scores with an SDCG of 0.781, but also has the highest harm scores of 0.393 in SDCG. The child-safe search engines show a diverse picture, providing a wide range of possible trade-offs, e.g., the effectiveness in relevance ranges from 0.121 in SDCG for Seitenstark to 0.381 for FragFinn. Following the same trend, Seitenstark has the fewest results that are not appropriate for children (SDCG of 0.005 for harm) whereas FragFinn contains more non-appropriate results (SDCG of 0.133 for harm). Importantly, the harm scores of the child-safe search engines originate from documents that were labeled with a harm score of 1 that indicates that

Table 2. Effectiveness comparison: an ideal child-friendly search result has high relevance scores and low harmfulness scores (best scores shown in bold). A † indicates a significant difference (Bonferroni corrected) of a child-oriented search engine to Google.

Search Engine		**Relevance**				**Harm**			
Type	**Name**	**P@3**	**P@10**	**RR**	**SDCG**	**P@3**	**P@10**	**RR**	**SDCG**
Commercial	Bing	.876	.773	**.931**	**.781**	.688	.651	.859	.393
	DuckDuckGo	.669	.648	.758	.560	.787	.810	.867	.385
	Google	**.957**	**.835**	.924	.672	.671	.603	.806	.324
Child	FragFinn	.531†	.458†	.691†	.381†	.276†	.257†	.418†	.133†
	Helles-Koepfchen	.257†	.229†	.436†	.174†	.129†	.119†	.209†	.060†
	Seitenstark	.261†	.170†	.376†	.121†	**.011**†	**.009**†	**.028**†	**.005**†

documents are not really appropriate for children, e.g., because they are too complex, whereas commercial search engines also retrieved documents with a harm score of 2. In all cases, the three child-safe search engines have significantly lower relevance scores than Google but show significantly fewer results that are not appropriate. This is expected, as the child-safe search engines index only a small subset of the web. However, it remains an interesting open question if more relevant documents would exist in the child-safe indexes, and if they would be retrievable via better queries.

5 Conclusion and Future Work

We created an evaluation corpus to compare six search engines on how suitable they are for children as searchers by labeling results for relevance and potential harm. Our results show that commercial search engines provide significantly more relevant search results compared to child-safe search engines. However, commercial search engines also show substantially more inappropriate content that might be harmful for children. For future work, it would be interesting to expand the dataset to more queries and more languages, and as soon as child-safe search engines are effective on our Cranfield-style datasets, expanding them to user experiments with children would also be interesting. Furthermore, large language models could yield as an interesting way to scale the corpus creation, especially when large language models would be able to label if documents are potentially harmful for children. This way, a large dataset could be created and expensive human judgment effort could be focused to a small part that promises the most potential for improvements. Another interesting line for future work would be how to improve the effectiveness of child-safe search engines. For this, it could be interesting to use the responses of commercial search engines with potentially harmful results as relevance feedback to retrieve from smaller indices operated by child-safe search engines that contain only non-harmful documents.

References

1. Die verbreitetsten Suchmaschinen: 20 Google-Alternativen (2023), `https://www.loewenstark.com/wissen/suchmaschinen`
2. Abualsaud, M., Smucker, M.D., Lioma, C., Maistro, M., Zuccon, G.: Overview of the TREC 2019 Decision track. In: Proceedings of TREC (2019)
3. Behrensand, P., Rathgeb, T.: Kinder und Medien, Computer und Internet Basisuntersuchung zum Medienumgng von 6- bis 13-Jährigen in Deutschland. pp. 30–43. Medienpädagogischer Forschungsverbund Südwest (2011)
4. Bevendorff, J., Potthast, M., Stein, B.: FastWARC: Optimizing large-scale web archive analytics. In: Wagner, A., Guetl, C., Granitzer, M., Voigt, S. (eds.) 3rd International Symposium on Open Search Technology (OSSYM 2021). International Open Search Symposium (Oct 2021)
5. Bilal, D.: Web search engines for children: A comparative study and performance evaluation of "Yahooligans!," "AskJeeves for Kids," and "Super Snooper.". In: Proceedings of the ASIS annual meeting. vol. 36, pp. 70–83. ERIC (1999)
6. Bilal, D.: Children's use of the Yahooligans! Web search engine. III. Cognitive and physical behaviors on fully self-generated search tasks. J. Assoc. Inf. Sci. Technol. **53**(13), 1170–1183 (2002)
7. Broder, A.Z.: A taxonomy of web search. SIGIR Forum **36**(2), 3–10 (2002). `https://doi.org/10.1145/792550.792552`
8. Clarke, C.L.A., Maistro, M., Smucker, M.D.: Overview of the TREC 2021 Health Misinformation track. In: Proceedings of TREC (2021)
9. Clarke, C.L.A., Rizvi, S., Smucker, M.D., Maistro, M., Zuccon, G.: Overview of the TREC 2020 Health Misinformation track. In: Proceedings of TREC (2020)
10. Clarke, C.L.A., Vtyurina, A., Smucker, M.D.: Offline evaluation without gain. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR 2020: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14–17, 2020. pp. 185–192. ACM (2020). `https://doi.org/10.1145/3409256.3409816`, `https://doi.org/10.1145/3409256.3409816`
11. Cleverdon, C.: The Cranfield tests on index language devices. In: ASLIB Proceedings. pp. 173–192. MCB UP Ltd. (Reprinted in Readings in Information Retrieval, Karen Sparck-Jones and Peter Willett, editors, Morgan Kaufmann, 1997) (1967)
12. Cleverdon, C.W.: The significance of the Cranfield tests on index languages. In: Bookstein, A., Chiaramella, Y., Salton, G., Raghavan, V.V. (eds.) Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum). pp. 3–12. ACM (1991)
13. Coleman, T.: Media and the well-being of children and adolescents. Journal of Youth and Adolescence **43**, 2083–2087 (2014)
14. Dowie, D., Azzopardi, L.: Re-leashed! The PuppyIR framework for developing information services for children, adults and dogs. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S.M., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24–27, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7814, pp. 824–827. Springer (2013). `https://doi.org/10.1007/978-3-642-36973-5_94`

15. Eickhoff, C., Serdyukov, P., de Vries, A.P.: A combined topical/non-topical approach to identifying web sites for children. In: King, I., Nejdl, W., Li, H. (eds.) Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9–12, 2011. pp. 505–514. ACM (2011). https://doi.org/10.1145/1935826.1935900, https://doi.org/10.1145/1935826.1935900
16. Feierabend, S., Plankenhorn, T., Rathgeb, T.: KIM-Studie 2014 : Kindheit, Internet, Medien. Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger. pp. 35–39. Medienpädagogischer Forschungsverbund Südwest (2014)
17. Feierabend, S., Rathgeb, T., Kheredmand, H., Glöckler, S.: KIM-Studie 2022 : Kindheit, Internet, Medien. Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger. pp. 28–33. Medienpädagogischer Forschungsverbund Südwest (2022)
18. Gossen, T., Höbel, J., Nürnberger, A.: A comparative study about children's and adults' perception of targeted web search engines. In: Jones, M., Palanque, P.A., Schmidt, A., Grossman, T. (eds.) CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 – May 01, 2014. pp. 1821–1824. ACM (2014)
19. Harman, D.: TREC-style evaluations. In: Agosti, M., Ferro, N., Forner, P., Müller, H., Santucci, G. (eds.) Information Retrieval Meets Information Visualization - PROMISE Winter School 2012, Zinal, Switzerland, January 23–27, 2012, Revised Tutorial Lectures. Lecture Notes in Computer Science, vol. 7757, pp. 97–115. Springer (2012). https://doi.org/10.1007/978-3-642-36415-0_7
20. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., Potthast, M.: Reproducible web corpora: Interactive archiving with automatic quality assessment. Journal of Data and Information Quality (JDIQ) **10**(4), 17:1–17:25 (Oct 2018). https://doi.org/10.1145/3239574
21. Landoni, M., Huibers, T., Murgia, E., Pera, M.S.: Good for children, good for all? In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 14611, pp. 302–313. Springer (2024). https://doi.org/10.1007/978-3-031-56066-8_24
22. Lingnau, A., Ruthven, I., Landoni, M., van der Sluis, F.: Interactive search interfaces for young children – The PuppyIR approach. In: ICALT 2010, 10th IEEE International Conference on Advanced Learning Technologies, Sousse, Tunisia, 5–7 July 2010. pp. 389–390. IEEE Computer Society (2010). https://doi.org/10.1109/ICALT.2010.111
23. Livingstone, S., Haddon, L., Goerzig, A., Ólafsson, K.: Risks and safety on the Internet: The perspective of European children. (2011)
24. Moffat, A.: Computing maximized effectiveness distance for recall-based metrics. IEEE Trans. Knowl. Data Eng. **30**(1), 198–203 (2018). https://doi.org/10.1109/TKDE.2017.2754371
25. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text annotation tool for human (2018), https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano
26. van der Sluis, F., van Dijk, B.: A closer look at children's information retrieval usage. In: Proceedings of the Workshop on Accessible Search Systems, 33st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). Glasgow: University of Strathclyde. pp. 3–10. ACM (2010)

27. Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., Livingstone, S., Hasebrink, U.: EU Kids Online 2020: Survey results from 19 countries (2020)
28. Torres, S.D., Hiemstra, D., Serdyukov, P.: An analysis of queries intended to search information for children. In: Belkin, N.J., Kelly, D. (eds.) Information Interaction in Context Symposium, IIiX 2010, New Brunswick, NJ, USA, August 18–21, 2010. pp. 235–244. ACM (2010)
29. Torres, S.D., Hiemstra, D., Serdyukov, P.: Query log analysis in the context of information retrieval for children. In: Crestani, F., Marchand-Maillet, S., Chen, H., Efthimiadis, E.N., Savoy, J. (eds.) Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010. pp. 847–848. ACM (2010)
30. Torres, S.D., Hiemstra, D., Weber, I., Serdyukov, P.: Query recommendation in the information domain of children. J. Assoc. Inf. Sci. Technol. **65**(7), 1368–1384 (2014)
31. Torres, S.D., Weber, I.: What and how children search on the Web. In: Macdonald, C., Ounis, I., Ruthven, I. (eds.) Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24–28, 2011. pp. 393–402. ACM (2011)
32. Ungruh, R., Pera, M.S.: Ah, that's the great puzzle: On the quest of a holistic understanding of the harms of recommender systems on children. In: Designing for Children's Digital Well-Being: A Research, Policy and Practice Agenda (DCDW '24), co-located with ACM IDC 2024 (2024). https://doi.org/10.48550/ARXIV.2405.02050
33. Wentzel, S.D.: Evaluating information retrieval systems for children in an educational context (2019), http://essay.utwente.nl/78835/