It is relevant, but is it useful?

A Reflection on Human-Centred Evaluation in Children Information Retrieval

Hrishita Chakrabarti
Delft University of Technology
Delft, Netherlands
h.chakrabarti@tudelft.nl

Diletta Micol Tobia
Università della Svizzera Italiana
Lugano, Switzerland
diletta.micol.tobia@usi.ch

Garrett Allen
Delft University of Technology
Delft, Netherlands
g.m.allen@tudelft.nl

Monica Landoni
Università della Svizzera Italiana
Lugano, Switzerland
monica.landoni@usi.ch

Maria Soledad Pera
Delft University of Technology
Delft, Netherlands
M.S.Pera@TUDelft.nl

Abstract

The traditional Information Retrieval (IR) evaluation framework—anchored in topical relevance and relevance-based metrics—reflects a system-centred perspective. Yet for specific user groups, relevance alone is insufficient; benchmarking that relies exclusively on conventional metrics overlooks qualities intrinsic to the users IR approaches are meant to serve. Here, we draw attention to Children IR and examine the value of extending traditional evaluation with a human-centred perspective that accounts for how children interpret and evaluate information to more authentically capture performance and better reflect how well an approach truly meets children's needs. Our empirical exploration using a child-focused dataset, multiple ranking strategies, and traditional and extended frameworks reveals not only the limitations of relevance-based assessments but also the advantages of employing frameworks that are tailored to reflect the needs of child users, paving the way for more inclusive and effective evaluation frameworks.

CCS Concepts

• **Social and professional topics** → **Children**; • **General and reference** → *Evaluation*; • **Information systems** → *Information retrieval*.

Keywords

Information Retrieval, Evaluation, Children, Search

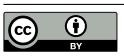
ACM Reference Format:
Hrishita Chakrabarti, Diletta Micol Tobia, Garrett Allen, Monica Landoni, and Maria Soledad Pera. 2026. It is relevant, but is it useful?: A Reflection on Human-Centred Evaluation in Children Information Retrieval. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3774935.3806167

1 Introduction

Evaluation has long been central to Information Retrieval (**IR**), guiding how systems are designed, compared, and optimised. The traditional IR evaluation framework, based on the Cranfield paradigm [7, 49], primarily equates system performance to the relevance of the resources retrieved for a given query [8, 45]. Although widely used for benchmarking, this framework considers topical relevance as a sufficient proxy for usefulness [48, 49]. However, IR systems ultimately serve humans, whose information needs are situated, contextual, and shaped by cognitive, emotional, and experiential factors [18, 20, 46]. From this perspective, a system is effective not merely when it retrieves relevant documents, but when it supports users in satisfying their information needs—an inherently human-centred notion that extends beyond relevance.

Children exemplify a user group for whom a human-centred approach to evaluation is essential. Their developing cognitive, linguistic, and metacognitive skills shape how they judge and interpret resources [22, 25, 39, 51]. Consequently, performance when children are the primary stakeholders depends not only on relevance, but also on whether documents can be used to distil the information they need [25]. In the realm of Children IR [39], evaluation requires considering document utility alongside relevance, which is influenced by factors such as readability, conceptual complexity, and alignment with children's interests and learning goals—dimensions not accounted for in the traditional IR evaluation paradigm.

In this work, we empirically explore the feasibility and value of incorporating a human-centred perspective to offline evaluation for Children IR. Given the broad range of skills and knowledge of the target audience, we scope our analysis to children up to grade 4 (i.e., 10 years old) and four *dimensions* representing their needs when judging resource relevance and utility: (1) *Relevance*, to ensure that retrieved resources relate to inquiry topics, (2) *Comprehension*, as children need to comprehend the content they are exposed to distil knowledge, (3) *Objectivity*, accounting for children's susceptibility to biased information, and (4) *Educational Value*, i.e., whether resource present information appropriate for school–level education, as children's understanding of the world is largely shaped by their school curriculum. We consider three evaluation frameworks to probe the performance of various ranking strategies—from simple first-stage retrievers to mainstream, commercial, and child-tailored state-of-the-art re-rankers—on a child-specific dataset [15]: `Traditional` with common relevance-based metrics, `Multiview` where we add to relevance each of the user-specific dimensions



UMAP '26, Gothenburg, Sweden

ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3806167

separately using metrics that produce independent scores per dimension, and `Composite`, in which we integrate all dimensions into a single score to enable benchmarking.

Performance analysis using the aforementioned frameworks provides insights into how each framework gauges the "fit" of ranking strategies for supporting children's information seeking. This, in turn, exposes the trade-offs inherent in the frameworks' ability to reflect realistic children's perspectives. To enable future work and reproducibility, we share our GitHub repository: https://github.com/SOLandChildren/ChildCenteredEvaluation.

2 Related Work

In the 1960s, the Cranfield paradigm [7] established the use of test collections to evaluate IR systems under controlled laboratory conditions using metrics such as precision and recall. These collections consist of a document corpus, a set of "topics" representing users' information needs, and expert relevance judgments based on topical similarity. The paradigm has been adopted by major IR evaluation campaigns, e.g., TREC [50], CLEF [1], and NTCIR [6], spanning diverse contexts. Still, the underlying objective remains the same: measuring a system's ability to retrieve resources that are topically relevant to a query. However, the broader purpose of an IR system is to support users' information-seeking process, which is shaped by cognitive, emotional, and situational factors [22, 43] that are not captured by topical relevance alone. These factors are particularly important for user groups whose needs and abilities diverge from the mainstream, such as neurodivergent individuals, people with mental health conditions, and children, for whom readability, cognitive load, emotional tone, or presentation style are as critical as topical relevance [14, 25, 35, 39, 53], and yet are not accounted for in the traditional evaluation framework.

Several existing Children IR approaches address the various challenges children face when seeking information [e.g., 4, 10, 11, 31, 34]. Despite targeting children, these approaches are almost exclusively assessed following the traditional relevance-focused evaluation framework [4, 34]. Although effective for benchmarking against baselines and (mainstream) state-of-the-art counterparts, it remains system-centred and overlooks traits inherent to child users. Efforts to assess additional dimensions are limited and are usually treated as complements to traditional evaluation. For instance, Allen et al. [4] use the MRR_B metric—a variant of MRR capturing the position of the first child-inappropriate resource—to gauge the ability of the ranker to prevent children from encountering harmful content. Usability studies provide another complementary lens, enabling examination of whether children prefer a child-tailored system over mainstream ones [e.g., 24, 40]. However, these studies assume that child-tailored systems better meet children's information needs, an assumption grounded in relevance-based evaluation rather than an assessment of the broader factors that shape the extent to which a topically relevant resource can actually facilitate children's information seeking [39]. Hence, no existing evaluation framework jointly quantifies how well an IR approach fares while accounting for children's cognitive abilities, understanding, and engagement when addressing children's information needs. This gap makes it difficult to determine whether current approaches truly achieve their intended purpose: fulfil the information need of searchers with specific profiles, given a task and a context (drawing on the pillars introduced in [26] to ground evaluation in Children IR).

3 Experimental Set Up

Dataset. We use **`kid-friend`** [15], a public TREC-style dataset consisting of 50 adult-formulated queries reflecting children's (up to 4^{th} grade, i.e., 10 years old) personal, political, educational, and entertainment information needs and 2385 web resources (represented by title, snippet, and full text), pooled from the top 10 results of 3 commercial and 3 child-tailored search engines (**SE**) per query. It also includes 2303 query-resource pairs with human annotations, of which 1480 are labelled relevant. All queries and relevance judgments were created by adult proxies, avoiding the risk of exposing children to unsuitable content during data collection.

Judgment Dimensions. Each resource R in `kid-friend` is annotated with ground truth labels across four dimensions central to Children IR, tailored for children up to grade 4: **Relevance**, following established IR evaluation practices, R is judged by its topical similarity with a query using human judgments provided in the dataset. **Comprehension**, to capture whether content is understandable for children, we assign R a readability score (a common indicator for text comprehension [32, 52]) using the Spache-Allen formula, which estimates the minimum school grade level required to read a snippet [3]. Resources with readability level $\leq$ grade 4 are treated as readable. **Objectivity**, as children struggle to distinguish factual content from subjective views [2, 16, 33], we assign R an objectivity score in the $[0, 1]$ range (based on R's snippet) using an `mDeBERTa V3` model fine-tuned for subjectivity detection [29]; score ≥ 0.5 indicates objective content. To gauge how well the content of a resource aligns with children's curriculum-informed knowledge, we compute R's **Educational Value** (based on its snippet) in the range $[0, 1]$[1] using the `fine-web-edu` model [30, 38]. We treat resources with a score ≥ 0.6 as educational.

Rankers. We consider several ranking strategies categorised as: (1) First-stage retrievers: BM25 [19], TF-IDF [44], and Dirichlet LM [56]; (2) State-of-the-art re-rankers: a neural re-ranker, monoT5 [37] and two LLM-based re-rankers, RankVicuna [41] and RankZephyr [42]; (3) Child-tailored re-rankers: KORSCE [34] and REdORank [4]; (4) Commercial SE: Google and Bing. Following common practice [21, 54, 55], the top-100 BM25-retrieved resources serve as input to all re-rankers. For the commercial SE, we use the ranked resources from `kid-friend` (see [15]). As KORSCE and REdORank were designed to promote child-appropriate content during web search [4, 34], we also apply them to re-rank the commercial SE results.

Metrics. We assess ranker performance based on the top 10 resources retrieved for a query. We use common benchmarking metrics: Precision (**P@10**) and **nDCG@10**. To capture other dimensions along with relevance, inspired by the MRR variant introduced in [4], we use $\textbf{MRR}_x$, which measures how early a resource that is *relevant* and $x \in \{\underline{readable}, \underline{objective}, \underline{educational}\}$ is positioned. Similarly, $\textbf{P}_x\textbf{@10}$ is a variation of Precision, which measures the average number of resources retrieved that are *relevant* and x. We integrate all dimensions into a single metric using Ranked Biased

[1]The model outputs values from 0–5, where scores ≥ 3 indicate age-appropriate content [30]; we normalize these to $[0, 1]$ for computation.

Precision (**RBP**), which models user "persistence", i.e., the likelihood of a user examining a document at rank k after viewing the results ranked above it [36] and proven effective in incorporating comprehension alongside relevance for medical IR evaluation [58]. Specifically, we adopt a child-centred variant, **cRBP@10**, defined in Equation 1, which yields a score in $[0, 1]$.

$$cRBP@10 = (1-\rho)\sum_{k=1}^{10}\rho^{k-1} rel(R_k)\cdot comp(R_k)\cdot o(R_k)\cdot e(R_k) \quad (1)$$

where k is the ranking position, $\rho = 0.8$ user persistence, $rel(R_k)$, $o(R_k)$, $e(R_k)$ are the relevance, objectivity and educational scores for R, respectively, and $comp(R_k)$ captures the degree of alignment between R's readability and a child's comprehension, computed using Equation 2 [34], where $RL(R_k)$ is R's readability score and $G = 4$ the target readability level.

$$comp(R_k) = \begin{cases} 1 & RL(R) = G \\ \frac{cos(0.79 * RL(R_k) - (G - 0.21 * G)) + 1}{2} & G < RL(R_k) < G + 4 \\ \frac{cos(0.5236 * RL(R_k)^2 - 0.5236 * G) + 1}{2} & G - 6 < RL(R_k) < G \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

Table 1: Evaluation frameworks used in the study.

Framework	Judgement Dimensions	Metrics	Takeaways
`Traditional`	Relevance	P@10, nDCG@10	- enables benchmarking - provides a limited perspective on ranker effectiveness
`Multiview`	Relevance, Comprehension, Objectivity, Education Value	P_x@10, MRR_x	- accounts for user-specific factors - offers insights into strengths and limitations of a ranking strategy - benchmarking is challenging
`Composite`	=	cRBP@10	- accounts for user-specific factors - enables benchmarking - less transparent on ranker's effectiveness across factors

Evaluation Frameworks. We probe three frameworks, described in Table 1. `Traditional` follows the conventional IR framework; `Multiview` and `Composite` also account for user-specific judgment dimensions. In `Multiview`, we adopt a reductionist approach and evaluate ranker performance by adding to relevance each of the user-specific dimensions separately. In contrast, `Composite` takes a holistic approach, assessing performance based on the aggregated contribution of all considered dimensions.

4 Results and Discussion

We report the performance of all rankers based on different metrics in Table 2. Our goal is not to identify the best-performing ranker for Children IR, but to use these results to distil emergent trends to inform reflection on the benefits and limitations of each framework.

Using `Traditional`, we observe substantial differences across rankers, with the best-to-worst ranker ordering remaining fairly consistent across both relevance-based metrics. Based on the P@10 and nDCG@10 values reported in Table 2, most rankers retrieve 5-7 relevant resources in the top 10, which are also typically ranked near the top, except for KORSCE and REdORank when re-ranking the top 100 BM25 results. Overall, trends from `Traditional` suggest that most strategies are effective at addressing children's inquiries.

Table 2: Ranker performance based on various metrics. Bold and underlined values indicate best and second-best performing rankers per evaluation metric, respectively.

Ranker	Traditional		Multiview						Composite
	P@10	nDCG@10	MRR_{read}	P_{read}@10	MRR_{edu}	P_{edu}@10	MRR_{obj}	P_{obj}@10	cRBP@10
BM25	0.576	0.599	0.493	0.294	0.199	0.080	0.671	0.452	0.065
TF-IDF	0.560	0.587	0.469	0.280	0.199	0.080	0.670	0.448	0.065
Dirichlet LM	0.550	0.583	0.477	0.264	**0.200**	0.074	0.679	0.432	0.064
BM25+MonoT5	0.642	0.643	0.496	0.318	0.162	0.080	0.644	0.492	0.063
BM25+RankVicuna	0.576	0.602	0.509	0.294	0.199	0.080	0.671	0.452	0.065
BM25+RankZephyr	0.618	0.644	0.499	0.304	0.176	**0.096**	0.662	0.490	**0.070**
BM25+KORSCE	0.154	0.164	0.255	0.106	0.075	0.018	0.287	0.114	0.020
BM25+REdORank	0.136	0.134	0.190	0.084	0.003	0.002	0.004	0.004	0.003
Google	**0.748**	**0.793**	0.499	**0.320**	0.194	0.066	**0.829**	**0.608**	0.061
Google+KORSCE	0.722	0.741	0.595	0.312	0.185	0.066	0.755	0.594	0.061
Google+REdORank	0.722	0.755	**0.647**	0.312	0.151	0.066	0.445	0.582	0.055
Bing	0.536	0.557	0.401	0.300	0.168	0.062	0.561	0.422	0.062
Bing+KORSCE	0.530	0.540	0.411	0.300	0.128	0.062	0.525	0.426	0.064
Bing+REdORank	0.524	0.532	0.437	0.276	0.087	0.060	0.210	0.358	0.045

We evaluate rankers using `Multiview` and find that relative ranker performance is no longer consistent across the individual user-specific dimensions. For instance, for Google, on average, 6 results are both relevant and objective (as per its P_{obj}@10), with the first one often ranked at the top (see its MRR_{ojb}). However, a relevant and educational resource seldom appears in the top 10; when it does, it is ranked 5^{th}. Examining the effectiveness of all rankers collectively, we observe that their performance varies widely across dimensions. Among the top 10 results, most rankers have 4-6 resources that are both objective and relevant, with the first such resource typically appearing in the 1^{st} or 2^{nd} position. In contrast, only about 2-3 resources are both relevant and readable, though these resources generally appear at the top. Based on P_{edu}@10 and MRR_{edu} scores, rankers struggle most with prioritizing educational resources: for most queries, rarely a relevant resource that is also educational makes it to the top 10 results, and when it happens, the first resource usually appears in the 5^{th} position. Overall, `Multiview` reveals distinct strengths and limitations of each ranker when accounting for user-specific dimensions, but reaching a consensus on the overall ranker performance is challenging.

`Composite` leads to notable performance differences across rankers; still cRBP scores are relatively small, ranging between 0.003 and 0.070. Although based on the cRBP we can discern the better-performing rankers, the generally low scores suggest that none are particularly effective in prioritising relevant resources that also align with children's cognitive processing levels.

Informed by the trends and insights revealed by each evaluation framework regarding the suitability of the evaluated rankers for Children IR, we reflect on the strengths and limitations of these frameworks (summarised in Table 1). Using `Traditional`, which focuses on relevance alone, the best-performing ranker is easily identifiable due to its notably higher metric score compared to the other counterparts. Still, the only conclusion we can draw from these comparisons is which ranker is better at promoting resources that are topically related to a child's query. This evaluation reflects an implicit assumption that exposing children to relevant resources as early as possible is sufficient for successful search. While prior studies focusing on children's information-seeking practices confirm that they tend to fixate on the first few search results [5, 12],

research also shows that even when directed to relevant resources, children often struggle to extract the information they need [24]. This difficulty often stems from exposure to content that is too complex for young users to read and understand [26, 34]. Therefore, using `Traditional` may risk overstating the "fit" of a ranking strategy—or IR strategy more broadly—in Children IR contexts.

The findings from using `Multiview`, which models user-specific factors influencing children's information seeking, reveal that no single ranker—or even category of rankers—consistently outperforms the others across all three user-specific dimensions. Probing ranker performance by adding to relevance each dimension individually provides a clearer view of each ranker's strengths in accounting for factors that impact children's search success. However, with the reductionist approach adopted under `Multiview`, it is difficult to identify a clear best-performing ranker; instead using the framework highlights trade-offs that have to be accepted when selecting a ranker or addressed, possibly by combining multiple rankers best suited for different user-specific dimensions.

Using `Composite`, the singular metric captures ranker effectiveness in retrieving relevant resources that are also useful for children. This framework, therefore, enables benchmarking that not only acknowledges the fundamental IR requirement of topical relevance but further *tailors* the evaluation to reflect the distinct needs of children, resulting in a more meaningful performance analysis. However, based on the insights drawn from using `Multiview`, we know that no ranker consistently outperforms the others across all three user-specific dimensions. Therefore, although `Composite` can aid in discerning the ranker with the best overall performance across all dimensions, the aggregated score loses the nuance offered in `Multiview` on whether a ranker's effectiveness is due to balanced strength across dimensions or whether its gains in one dimension compensate for limitations in others.

Juxtaposing emergent trends across evaluation frameworks shows that each offers a distinct view on ranker performance. Trends emerging from `Traditional` suggest that most rankers handle children's inquiries well. However, insights from `Multiview` and `Composite` reveal that, when accounting for factors that impact children's search process, several rankers are less effective than they appear. These discrepancies are evident in Figure 1, where we depict Kendall's τ correlations between all performance metrics. Relevance-based metrics (nDCG@10 and P@10) are strongly correlated with MRR_{read}, P_{read}@10, and P_{obj}@10, indicating that rankers good at retrieving relevant resources generally also prioritize readable ones. However, weaker correlations with the other metrics used in `Multiview` suggest that effectiveness in ranking relevant resources does not guarantee prioritization of objective or educational resources. Therefore, relying on `Traditional` may at best identify rankers that meet children's comprehension needs, but leaves unclear whether the ranked resources are also unbiased and aligned with children's understanding of the world.

Access to accurate, age-appropriate information is critical for children's learning and development [9, 27]. Hence, IR strategies ought to be evaluated based on how effectively they support information seeking. Our findings indicate that traditional relevance evaluation offers a myopic—and sometimes misleading—view of system effectiveness in addressing children's needs. This underscores the importance of incorporating a human-centred perspective in Children IR evaluation. `Multiview` and `Composite` demonstrate that human-centred dimensions—accounting for the user, task, and context (i.e., the pillars introduced in [26])—can be easily added to benchmarking with simple extensions, preserving standard community expectations, but providing richer insights.

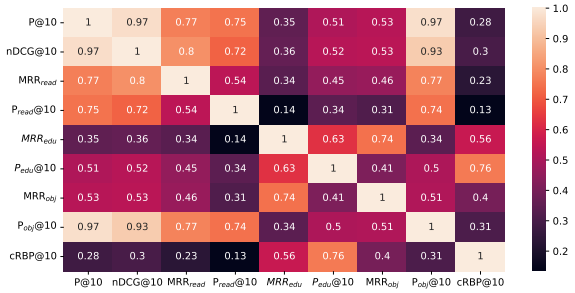


Figure 1: Correlation between evaluation metrics when ranking strategies are ordered by their respective metric score.

In our work, the cRBP metric used in `Composite` weights these judgment dimensions equally. In practice, however, different dimensions may influence children's information-seeking to varying degrees, as also noted in [34]. Existing research shows that topical relevance and comprehension affect adults' search behaviour in distinct ways [e.g., 52, 57], suggesting similar variation for children. Future studies to determine the relative impact of user-specific dimensions on children's search practices should involve not only children but also other stakeholders, such as parents or teachers, who help shape how children appraise online resources [13, 23]. Insights from such studies could inform non-uniform weighting of cRBP dimensions and validate whether the operationalisation of utility factors reflects their actual impact on children's search experience. Moreover, while `Multiview` and `Composite` include several crucial factors affecting children's search success, they are not exhaustive. Further user studies could identify and operationalise additional dimensions, such as the emotional tone of resources [28]. Due to ethical constraints of research involving vulnerable populations, we used `kid-friend`, the only publicly available child-focused dataset at the time. However, its queries were formulated by adult proxies and may not fully reflect children's distinct query formulation habits [17, 47]. Future work should therefore prioritize creating TREC-style datasets grounded on children's real-life search interactions, as explored in [28, 31].

5 Concluding Remarks

In this work, we explored the possibility of assessing whether resources children encounter are relevant, but also *useful* for them. Our analysis of ranker performance using different evaluation frameworks shows that reflecting the user's perspective during offline evaluation—while preserving the spirit and rigour of traditional IR evaluation—is not only feasible but also reveals aspects of ranker behaviour that standard metrics overlook, strengthening benchmarking practices and guiding the development of IR technologies that effectively support young searchers.

Acknowledgments

This work is supported by SNSF Award #[IC00I0-227887 project n.10000973 SOL] and used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-16917. GenAI tools were used in the research and for editing the manuscript for clarity, but the authors are fully accountable for the content.

References

[1] Eneko Agirre, Giorgio Maria Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. 2008. CLEF 2008: Ad hoc track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 15–37.

[2] Safinah Ali, Daniella DiPaola, Irene Lee, Victor Sindato, Grace Kim, Ryan Blumofe, and Cynthia Breazeal. 2021. Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence* 2 (2021), 100040. https://www.sciencedirect.com/science/article/pii/S2666920X21000345

[3] Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. 2022. Supercalifragilisticexpialidocious: Why using the "right" readability formula in children's web search matters. In *European Conference on Information Retrieval*. Springer, 3–18.

[4] Garrett Allen, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. 2023. Multi-Perspective Learning to Rank to Support Children's Information Seeking in the Classroom. In *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 311–317.

[5] Dania Bilal and Joe Kirby. 2002. Differences and similarities in information seeking: children and adults as Web users. *Information processing & management* 38, 5 (2002), 649–670.

[6] Junjie Chen, Haitao Li, Zhumin Chu, Yiqun Liu, and Qingyao Ai. 2025. Overview of the NTCIR-18 Automatic Evaluation of LLMs (AEOLLM) Task. (2025).

[7] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, 173–194.

[8] W Bruce Croft, Donald Metzler, Trevor Strohman, et al. 2010. *Search engines: Information retrieval in practice*. Vol. 520. Addison-Wesley Reading.

[9] Judith H Danovitch. 2019. Growing up with Google: How children's understanding and use of internet-based devices relates to cognitive development. *Human Behavior and Emerging Technologies* 1, 2 (2019), 81–90.

[10] Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright, and Casey Kennington. 2020. Kidspell: A child-oriented, rule-based, phonetic spellchecker. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 6937–6946.

[11] Nevena Dragovic, Ion Madrazo Azpiazu, and Maria Soledad Pera. 2016. " Is Sven Seven?" A Search Intent Module for Children. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 885–888.

[12] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 847–848.

[13] Michael D Ekstrand, Afsaneh Razi, Aleksandra Sarcevic, Maria Soledad Pera, Robin Burke, and Katherine Landau Wright. 2025. Recommending with, not for: Co-designing recommender systems for social good. *ACM transactions on recommender systems* (2025).

[14] Sukru Eraslan, Yeliz Yesilada, Victoria Yaneva, and Le An Ha. 2020. "Keep it simple!": An eye-tracking study for exploring complexity and distinguishability of web pages for people with autism. *Universal Access in the Information Society* 20, 1 (2020), 69–84. doi:10.1007/s10209-020-00708-9

[15] Maik Fröbe, Sophie Charlotte Bartholly, and Matthias Hagen. 2025. How Child-Friendly is Web Search? An Evaluation of Relevance vs. Harm. In *European Conference on Information Retrieval*. Springer, 214–222.

[16] Lauren N Girouard-Hallam, Yu Tong, Fuxing Wang, and Judith H Danovitch. 2023. What can the internet do?: Chinese and American children's attitudes and beliefs about the internet. *Cognitive Development* 66 (2023), 101338.

[17] Jacek Gwizdka and Dania Bilal. 2017. Analysis of children's queries and click behavior on ranked results and their thought processes in google search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 377–380.

[18] Kalervo Jarvelin and Eero Sormunen. 2024. A Blueprint of IR Evaluation Integrating Task and User Characteristics. *ACM Transactions on Information Systems* 42, 6 (2024), 1–38.

[19] K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management* 36, 6 (2000), 809–840.

[20] Gabriella Kazai, Paul Thomas, and Nick Craswell. 2019. The emotion profile of web search. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 1097–1100.

[21] Sarawoot Kongyoung, Craig Macdonald, and Iadh Ounis. 2022. monoQA: Multi-Task Learning of Reranking and Answer Extraction for Open-Retrieval Conversational Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7207–7218. doi:10.18653/v1/2022.emnlp-main.485

[22] Carol Kuhlthau. 1997. Learning in Digital Libraries: An Information Search Process Approach. *Library Trends* 45 (03 1997).

[23] Monica Landoni, Mohammad Aliannejadi, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2021. Right Way, Right Time: Towards a Better Comprehension of Young Students' Needs when Looking for Relevant Search Results. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 256–261. doi:10.1145/3450613.3456843

[24] Monica Landoni, Mohammad Aliannejadi, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2022. Have a clue! the effect of visual cues on children's search behavior in the classroom. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 310–314.

[25] Monica Landoni, Theo Huibers, Emiliana Murgia, Mohammad Aliannejadi, and Maria Soledad Pera. 2021. Somewhere over the rainbow: Exploring the sense for relevance in children. In *Proceedings of the 32nd European Conference on Cognitive Ergonomics*. 1–5.

[26] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2019. Sonny, Cerca! evaluating the impact of using a vocal assistant to search at school. In *International conference of the cross-language evaluation forum for European languages*. Springer, 101–113.

[27] Monica Landoni, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2023. How does Information Pollution Challenge Children's Right to Information Access?. In *3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, ROMCIR 2023*. 17–29.

[28] Monica Landoni, Maria Soledad Pera, Emiliana Murgia, and Theo Huibers. 2020. Inside out: Exploring the emotional side of search engines in the classroom. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*. 136–144.

[29] {Folkert Atze} Leistra and Tommaso Caselli. 2023. Thesis Titan at CheckThat! 2023: Language-Specific Fine-tuning of mDeBERTaV3 for Subjectivity Detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023) (CEUR Workshop Proceedings)*, Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos (Eds.). CEUR Workshop Proceedings (CEUR-WS.org), 351–359. Publisher Copyright: © 2023 Copyright for this paper by its authors.; 24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023 ; Conference date: 18-09-2023 Through 21-09-2023.

[30] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu: the Finest Collection of Educational Content. doi:10.57967/hf/2497

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

[32] Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. In *European Conference on Information Retrieval*. Springer, 33–49.

[33] Miriam J Metzger, Andrew J Flanagin, Alex Markov, Rebekah Grossman, and Monica Bulger. 2015. Believing the unbelievable: Understanding young people's information literacy beliefs and practices in the United States. *Journal of Children and Media* 9, 3 (2015), 325–348.

[34] Ashlee Milton, Oghenemaro Anuya, Lawrence Spear, Katherine Landau Wright, and Maria Soledad Pera. 2020. A ranking strategy to promote resources supporting the classroom environment. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 121–128.

[35] Ashlee Milton and Maria Soledad Pera. 2023. Into the unknown: exploration of search engines' responses to users with depression and anxiety. *ACM Transactions on the Web* 17, 4 (2023), 1–29.

[36] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[37] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*. 708–718.

[38] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems* 37 (2024), 30811–30849.

[39] Maria Soledad Pera, Theo Huibers, Emiliana Murgia, and Monica Landoni. 2025. Some things never change: Overcoming persistent challenges in children IR.

In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3107–3109.

[40] Christine Pinney, Benjamin J Bettencourt, Jerry Alan Fails, Casey Kennington, Katherine Landau Wright, and Maria Soledad Pera. 2024. How Readability Cues Affect Children's Navigation of Search Engine Result Pages. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 62–69.

[41] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088* (2023).

[42] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *arXiv preprint arXiv:2312.02724* (2023).

[43] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[44] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[45] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and trends® in information retrieval* 4, 4 (2010), 247–375.

[46] Tefko Saracevic. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for information science* 26, 6 (1975), 321–343.

[47] Nicholas Vanderschantz and Annika Hinze. 2021. Children's query formulation and search result exploration. *International Journal on Digital Libraries* 22, 4 (2021), 385–410.

[48] Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*. Springer, 355–370.

[49] Ellen M Voorhees. 2019. The evolution of cranfield. In *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Springer, 45–69.

[50] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge.

[51] Virginia A. Walter. 1994. The Information Needs of Children. In *Advances in Librarianship*. Emerald Group Publishing Limited. arXiv:https://www.emerald.com/book/chapter-pdf/9122169/s0065-2830_1994_0000018006.pdf doi:10.1108/S0065-2830(1994)0000018006

[52] Yunjie Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 961–973.

[53] Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2015. Accessible Texts for Autism: An Eye-Tracking Study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 49–57. doi:10.1145/2700648.2809852

[54] Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seungwon Hwang. 2024. ListT5: Listwise Reranking with Fusion-in-Decoder Improves Zero-shot Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 2287–2308. doi:10.18653/v1/2024.acl-long.125

[55] Shi Yu, Chenghao Fan, Chenyan Xiong, David Jin, Zhiyuan Liu, and Zhenghao Liu. 2024. Fusion-in-T5: Unifying Variant Signals for Simple and Effective Document Ranking with Attention Fusion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 7556–7561. https://aclanthology.org/2024.lrec-main.667/

[56] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22, 2 (2004), 179–214.

[57] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 435–444.

[58] Guido Zuccon. 2016. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval*. Springer, 280–292.