Inside Out 2: Make Room for New Emotions & LLM: A Reproducibility Study of the Emotional Side of Search in the Classroom

Chakrabarti, Hrishita; Tobia, Diletta Micol; Landoni, Monica; Pera, Maria Soledad

DOI
10.1145/3726302.3730315

Publication date
2025

Document Version
Final published version

Published in
SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval

Inside Out 2: Make Room for New Emotions & LLM

A Reproducibility Study of the Emotional Side of Search in the Classroom

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

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

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

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

Abstract

In an existing study, the `InsideOut Framework` is used to produce and explore the emotional profiles of search engines (SE) in response to queries formulated by children aged 9 to 11 in the classroom context, revealing the emotional diversity of SE responses. Since then, there have been significant technological advances in emotion detection and information access. In this work, we conduct a comprehensive reproducibility study where we probe today's emotional profile of SE using both a lexicon-based and a language-model based approach tailored to the Italian language, thus addressing an acknowledged limitation of the original study. Additionally, considering the prevalence of agents based on Large Language Models (LLM) as information access systems among children, we extend the analysis to capture the emotional undertones of LLM responses and juxtapose them to those of SE. Our findings emphasize the importance of leveraging the appropriate emotion detection technique to produce and explore emotional profiles and lead us to reflect on the interplay of emotions on children's search-as-learning experience.

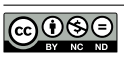
CCS Concepts

• **Social and professional topics** → **Children**; • **Information systems** → **Information retrieval**; **Sentiment analysis**.

Keywords

Children, Search, LLM, Emotions, Information Access Systems

ACM Reference Format:
Hrishita Chakrabarti, Diletta Micol Tobia, Monica Landoni, and Maria Soledad Pera. 2025. Inside Out 2: Make Room for New Emotions & LLM: A Reproducibility Study of the Emotional Side of Search in the Classroom. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730315



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730315

1 Introduction

Emotions play an integral role during information seeking, as "they affect a searcher's attention, memory, performance, and judgments" [28]. Numerous studies have shown that they influence decision-making processes, altering how individuals process information and select options [14]. Emotions impact search behaviour through changes in physical actions, cognitive processes, and affective outcomes [34]. In this context, Kazai et al. [21] specifically investigated the relationship between emotions and Search Engines **(SE)** by examining "the **emotion profile** of web search" through sentiment and emotion profiles they assigned to the retrieved and clicked results on Search Engine Result Pages (**SERP**) generated in response to users' search queries. Their study, based on query logs from a commercial SE addressing general inquiries from English-speaking adults, revealed significant differences in the emotional profiles of clicked versus non-clicked results. This emphasized how emotions influence searchers' attention, perception of results, and ultimately clicks. It also highlighted how the emotions expressed in a user's query can shape the tone of the results portrayed in SERP, underscoring the need for understanding how these emotional nuances can impact the user's overall search experience.

Inspired by this work, Milton and Pera [30] explored how the emotions and terminology of content retrieved by SE differed when queries were formulated by English-speaking adults experiencing depression and anxiety. They found that negative emotions were exacerbated in the SE responses to queries from their users, compared to those from the general population. These findings raise concerns about the potential impact of emotionally-charged search results on users who are vulnerable to emotions.

Another rapidly growing demographic susceptible to emotional influence is **children**. They increasingly rely on the Web to obtain information [19, 24, 39], but their skills and search habits significantly differ from those of adult searchers [6]. This prompted Landoni et al. [26] to investigate the emotional profile of SE in response to children's inquiries. The study, structured in two phases, analyzed search behaviours and SE responses when children undertook curriculum-related information discovery tasks to complete a class assignment with the help of Microsoft Bing. The findings emerging from Phase 1, based on the emotional profile of searches related to 'general' curriculum topics, revealed a consistent correlation between the emotions portrayed in the retrieved results and the children's clicks. The findings from Phase 2, focused on

children explicitly engaging with emotionally-charged search tasks in the classroom, demonstrated that emotionally intense queries significantly influence the emotional content of the results children clicked on. These outcomes highlighted the important role of emotions on the search behaviour of this group.

Besides showcasing the interplay among children, SE, and emotions, Landoni et al. [26] introduce the **`InsideOut Framework`**, which crafts emotional profiles of SE by capturing the valence of sentiments and emotions generally present in SE responses, e.g., SERP. In the original work, the authors leveraged a lexicon-based strategy—a key component of the framework—to capture the emotional undertones of text samples (e.g., queries or snippets) in English and Italian. For the latter, the authors used a Google-translated version of an English lexicon that was state-of-the-art at the time, which they acknowledged as a limitation of their original work.

Since the study's publication in early 2020, there have been notable advances in Natural Language Processing **(NLP)** technologies, including those used for inferring the sentiment and emotions expressed in texts written in English and beyond [e.g., 12, 22]. It cannot be denied that during this period, the Information Retrieval (**IR**) community has also seen considerable research progress, particularly in improving the resources retrieved and displayed on a SERP in response to user queries [e.g., 3, 18]. Given these developments, we argue that the emotional profile of SE in response to online inquiries—including those of children—has likely evolved as well. Alongside these advancements, information access is also evolving beyond traditional SE. Digital tools and platforms such as agents based on Large Language Models (**LLM**), social media platforms, and voice-controlled assistants are also becoming popular choices of information access systems (**IAS**), especially among young users [2, 11]. Notably, the generative capabilities of LLM have introduced a new way for individuals to seek information online [20, 46]. LLM agents can be especially helpful for those who find it challenging to locate relevant information via traditional SE [45]. In fact, for children, LLM hold particular promise, as they can address some of the known barriers children face when using mainstream SE, including struggling with query formulation, favouring looping behaviour from the query to the link, and focusing on finding a specific answer without trying to understand the contents [4, 25]. By providing direct answers, generative models could help reduce the cognitive effort children would otherwise face when sifting through numerous SE results. Furthermore, LLM agents accommodate more natural language queries, aligning better with the way children typically phrase their queries. Still, the emotional tone of LLM responses remains unexplored. We argue this needs to be further investigated as young searchers would receive a single response from such agents and the strong emotional content—or its absence—may directly affect their emotional development and learning experiences [41].

We conduct a **reproducibility study** on the `InsideOut Framework` [26]. Besides validating the study conducted by Landoni et al. [26], our motivation to revisit the framework is threefold:

(1) **Scrutinizing the impact of emotion detection strategies on the emotional profile of SE:** The original work examined the emotional profile of SE in response to children's queries in a classroom setting, using a lexicon-based approach to grasp the valence of sentiments and emotions. The study was conducted on two languages, using the EmoLex lexicons [31] on English responses and a Google-translated version of the lexicon for the Italian responses. The authors acknowledged this discrepancy as a limitation of the study. Given the development of NLP techniques to analyze emotions in text, we are motivated to revisit the framework, questioning whether the choice of lexicon and the use of a more advanced technique influence the emotional profile of the SE responses generated for children's inquiries.
(2) **Exploring the emotional tone of SE responses over time:** Since the original study was conducted in 2020, significant advancements in the IR community have impacted the resources retrieved and displayed on a SERP in response to user queries [3]. We hypothesize that the results may have evolved as well. This raises the need to examine the evolution of the emotional profile of SE over the past 5 years, as portrayed by the `InsideOut Framework`.
(3) **Extending `InsideOut Framework` beyond SE:** Unlike traditional SE, LLM agents provide direct and more personalized responses, which may affect children's search experience. With LLM being widely prevalent nowadays [8, 44], we posit that expanding our understanding of the emotional content presented by these generative models is essential. By extending the `InsideOut Framework` beyond SE, we investigate how the emotional profile varies across IAS when responding to children's queries.

The key contributions of this work are: (i) we conduct a reproducibility study on the `InsideOut Framework`, providing a deeper analysis of the importance of selecting appropriate components to generate emotional profiles of IAS and (ii) we extend the study scope to capture the emotional profiles of additional IAS that children commonly use in educational settings. By doing so, we examine how these tools differently shape their undertones when addressing children's inquiries. Along the way, we offer a nuanced analysis of how various IAS may expose children to different emotions during their search activities, considering the intensity of these emotions over time and across systems. We also juxtapose our findings with those reported in [26] to identify gaps and limitations while reflecting on the modifications made to the `InsideOut Framework`. Finally, we contemplate the role of the emotional profiles of IAS in shaping children's search experience in classroom settings, particularly pertinent due to the pivotal role of emotions within the search-as-learning (**SAL**) paradigm [9]. We share code and other resources on a public repository: https://github.com/SOLandChildren/SIGIR2025/tree/main/InsideOut2.

2 Experimental Setup

Here, we describe the components of the setup used to reproduce the original `InsideOut Framework` (see Figure 1).

The Framework. The `InsideOut Framework` was designed to model the emotional tone of responses an *IAS* generates. Given a *log* capturing a snapshot of user-system interactions related to information seeking (generally queries and associated SERP) and using a particular *emotion detection strategy*, the framework yields

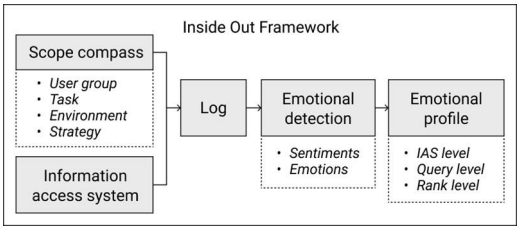


Figure 1: An overview of the `InsideOut Framework`.

a vector representation capturing an intensity distribution across a pre-defined set of sentiments and emotions. This vector is treated as the *emotional profile* of the IAS.

IAS. The IAS of the original study was Bing[1] (`setLang = 'it-IT', mkt = 'it-IT', safeSearch = 'Strict'`), which we refer to as **Bing'20**. We also probe the current Bing, referred to as **Bing'25** (same API and parameters). To extend the analysis of LLM-driven agents, we turn to two types of LLM: one proprietary and one open-source. For the former, we turn to OpenAI's **ChatGPT**, specifically `chatgpt-4o-latest`[2]. For the latter, we use Google's **Gemma** (`gemma-2-2b-it`[3])–a family of lightweight open models built using the same research and technology used to develop models for their flagship LLM-driven chatbot, Gemini [15].

Scope Compass. Landoni et al. [26] anchored their 2-phase study on four pillars [25]–*user group*, *environment*, *task*, and *strategy*–to control scope to specific inquiry activities. Particularly, Italian-speaking children, aged 9 to 11 years old, undertaking information discovery tasks in the classroom on topics common to the primary five curriculum using an SE. Phase 1 of the study focused on **General** tasks related to topics including tornadoes, endangered animals, and sports; Phase 2 on **Emotionally-Charged** tasks about the environment–a topic that the authors assumed would lead to emotionally-charged search experiences. Each Emotionally-Charged task was labelled as either **Positive** or **Negative** based on the sentiment it was likely to evoke.

Logs. In the `InsideOut Framework` [26], the IAS-scope compass pair prompted the generation of two logs: **QL-IT** and **QL-IT-Em**, capturing user-system interactions of children conducting online inquiries on General and Emotionally-Charged tasks, respectively. For brevity, we use **QL-IT-Em**$^{+}$ and **QL-IT-Em**$^{-}$ to indicate the disjoint search sessions in **QL-IT-Em** associated with Positive and Negative tasks. Each log consists of `<query,response>` pairs, where `response` is the top 5 SERP results retrieved by Bing'20 in response to `query`; each result is represented by its title and snippet. QL-IT has 2,610 `<query,response>` pairs capturing interactions of 75 children, whereas QL-IT-Em has 840 pairs originating from interactions of 66 children.

To enable reproducibility, we focus on the information needs expressed by children in the original study, which is why we generate synthetic logs relying on unique queries extracted from QL-IT and QL-IT-Em to elicit responses from Bing'25. This results in **B-G** and **B-Em** for General and Emotionally-Charged tasks, respectively. Similarly, we create two logs for ChatGPT and two for Gemma: **GPT-G**, **GPT-Em**, **GEM-G**, **GEM-Em**. In these logs, `response` refers to the LLM response generated for `query`. Of note, '+/−' appended at the end of the respective logs for Emotionally-Charged tasks indicates the sessions associated with Positive or Negative tasks, respectively. We instruct both LLM to generate text only in Italian by following each model's syntax for prompt formulation as exemplified below.

```
#Gemma prompt
{"role":"user", "content":""},
{"role": "assistant", "content": "Follow these two instructions
 in all your responses:
 1. Use Italian language only; 2. Do not use English
 except in programming language if any."},
{"role": "user", "content": query}

#ChatGPT prompt
{"role": "system",
 "content": "Follow these two instructions in all your responses:
 1. Use Italian language only; 2. Do not use English
 except in programming language if any."  },
{  "role": "user", "content": query }
```

Emotion Detection. The original framework used a lexicon-based strategy, which we denote `EmoLex`, to generate the **emotional vector** of a text sample by concatenating the sentiment and emotion vectors of the sample. `EmoLex` computes the sentiment/emotion vectors as the element-wise average of the sample's constituent word vectors using an Italian version of the EmoLex lexicon adapted from the original English lexicon using Google Translate [31]. The lexicon consists of 14,182 unigrams and 25,000 word senses, each annotated with a vector containing binary values for three sentiments, *Positive*, *Negative*, *Objective*; and eight emotions: *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, and *Trust* (words that are not in the lexicon are assigned a value of 1 for *Objective* sentiment, and 0 for all other sentiments and emotions).

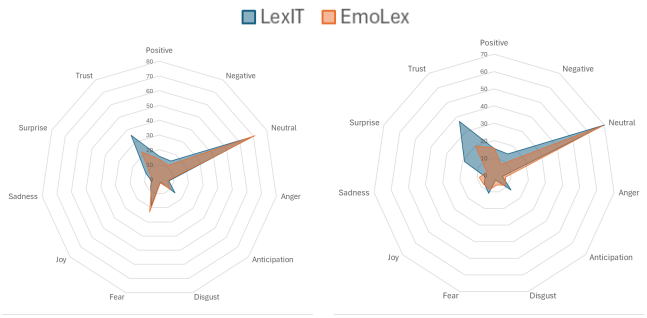
In our work, we turn to `LexIT`, another lexicon-based alternative that generates the emotional vector of a sample the same way as `EmoLex`, but instead of using a translated lexicon, it relies on lexicons created specifically for the Italian language. To our knowledge, no Italian lexicon simultaneously accounts for sentiments and emotions. Hence, the sentiment lexicon for `LexIT` is the Italian lexicon from the Distributional Polarity Lexicon [7] consisting of 75,021 Italian lemmas (concatenated with their parts-of-speech, i.e, PoS tag), each assigned a probability distribution comprised of 3 scores reflecting its *Positive*, *Negative*, and *Neutral* (*Objective* in `EmoLex`) polarity. The emotion lexicon is the Italian EMotive lexicon [36] inferred from the FB-NEWS15 corpus [35], which consists of crawled data from Facebook pages of the most important newspapers in Italy. The lexicon has 29,999 Italian words (and their corresponding PoS tag) each associated with 8 scores (on a -1 to 1 scale) reflecting the strength of a word's association to the same eight emotions as in `EmoLex` and akin to the EmoLex lexicon, the emotion lexicon is transformed into an 8-element binary vector annotation for each word, where a 1 is assigned to the emotion with the highest association score and 0 to others.

Note that, instead of averaging the sentiment and emotion values of words with multiple PoS (as done in `EmoLex`), `LexIT` assigns values to a word as per its PoS tag (identified using the Italian

[1] https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/
[2] https://platform.openai.com/docs/models#current-model-aliases
[3] https://huggingface.co/google/gemma-2b-it

Table 1: Overview of experiments conducted in our in-depth reproducibility study of the `InsideOut Framework`.

	In [26]	Δ lexicon	Δ time	Δ text context	Extension to LLM	Δ IAS
Environment	Classroom					
Task	Open/Fact online inquiries					
Topic	General & emotionally-charged curriculum topics					
User group	9-11 (primary 5 grade)					
Language	Italian					
Logs	QL-IT, QL-IT-Em	=	QL-IT, QL-IT-Em, B-G, B-Em	B-G, B-Em	GPT-G, GPT-Em, GEM-G, GEM-Em	B-G, B-Em, GPT-G, GPT-Em, GEM-G, GEM-Em
Emotion Detection	EmoLex	LexIT	=	FEEL-IT	FEEL-IT, LexIT	=
IAS	Bing '20	=	Bing'25	=	ChatGPT, Gemma	Bing'25, ChatGPT, Gemma



(a) QL-IT. **(b) QL-IT-Em.**

Figure 2: Emotional profiles based on `EmoLex` and `LexIT`.

spaCy model `it_core_news_sm`[4]). Furthermore, `LexIT` removes stopwords from a given text sample (using NLTK[5]) and then generates the emotional vector of the sample.

As a state-of-the-art alternative, we turn to a language-model-based strategy: `FEEL-IT`, based on a benchmark corpus consisting of Italian Twitter posts labelled with emotions [5]. Unlike `EmoLex` and `LexIT`, rather than considering each word in isolation to determine the emotional vector of a text sample, `FEEL-IT` favours contextual representations and considers the sample in its entirety. Specifically, `FEEL-IT` takes advantage of two `UmBERTo`[6]-based models: `feel-it-italian-sentiment`[7] and `feel-it-italian-emotion`[8] to generate the emotional vector of a sample as a distribution of weights across *Positive* and *Negative* sentiments and *Anger*, *Fear*, *Joy*, and *Sadness* emotions.

Emotional Profile. The `InsideOut Framework` produces $EP^X(\texttt{log})$ which represents the emotional profile for an IAS—based on a log—by computing the element-wise average of the emotional vectors of all responses per query and then averaging across all queries in `log`. We use X to denote the detection strategy applied to produce emotional vectors, with X being *OG* for `EmoLex`, L for `LexIT`, or F for `FEEL-IT`. `log`={QL-IT, QL-IT-Em, QL-IT-Em$^+$, QL-IT-Em$^-$, B-G, B-Em, B-Em$^+$, B-Em$^-$, GPT-G, GPT-Em, GPT-Em$^+$, GPT-Em$^-$, GEM-G, GEM-Em, GEM-Em$^+$, GEM-Em$^-$}.

For contextualization, we also apply the `InsideOut Framework` to produce emotional profiles at a query level, i.e., $EP_{query}{}^X(\texttt{log})$, by taking an element-wise average of the emotional vectors of queries in `log`; and when pertinent (i.e., for SE) at the rank level, i.e., $EP_{rank}{}^X(\texttt{log})$, by taking the average of vectors of responses across different ranks.

3 Reproducing Inside Out

In this section, we describe the experiments (summarized in Table 1) and the analysis concerning our reproducibility study.

3.1 The Influence of Lexicon Choice

A core component of the `InsideOut Framework` is the strategy used to produce emotional profiles. As previously stated, `EmoLex`, used in the original study, depends on English lexicons translated into Italian. We posit that a native Italian lexicon could impact emotional intensities. Hence, we apply the `InsideOut Framework` with `LexIT` (native Italian) to produce EP^L(QL-IT) and EP^L(QL-IT-Em), which we juxtapose with the original counterparts EP^{OG}(QL-IT) and EP^{OG}(QL-IT-Em) in [26].

As reported in Table 2 and illustrated in Figure 2, `LexIT` leads to prominently different emotional profiles for QL-IT and QL-IT-Em (independent T-test p<0.05 for significance). Particularly, the intensity of Trust—the emotion with the highest intensity in profiles generated using either strategy—is significantly higher when computed using `LexIT` compared to `EmoLex`. In fact, salient emotions in the profiles computed using `LexIT` have a higher intensity compared to the corresponding ones computed using `EmoLex`. Even the emotional profile of the queries from each log generated based on `LexIT`—$EP_{query}{}^L$(QL-IT) and $EP_{query}{}^L$(QL-IT)—showcase significantly higher intensities across most sentiment and emotions compared to their counterpart profile computed based on `EmoLex`, i.e., $EP_{query}{}^{OG}$(QL-IT) and $EP_{query}{}^{OG}$(QL-IT-Em). The outcomes of this experiment reveal that even the choice of lexicon to use for emotional analysis can lead to considerable changes in the profiles yielded using the `InsideOut Framework`.

3.2 The Impact of Time

The `InsideOut Framework` generated profiles based on Bing'20, using data from late 2019 to early 2020. Due to the evolving and dynamic nature of web collections, along with advancements in retrieval and ranking [1, 3, 10, 38], we examine how the emotional profile of Bing'25 differs from that of Bing'20.

We use the framework to produce EP^L(B-G), EP^L(B-Em), EP^L(B-Em$^+$), and EP^L(B-Em$^-$) which we compare to EP^L(QL-IT), EP^L(QL-IT-Em), EP^L(QL-IT-Em$^+$), and EP^L(QL-IT-Em$^-$) respectively. We see in the profiles reported in Table 2 that profiles of Bing'25 are predominantly Neutral, much like the profiles of Bing'20. The intensity of Trust—the emotion with the highest intensity across all profiles—is significantly (independent T-test, p<0.05) stronger in the emotional profiles of Bing'25 compared to the corresponding profiles of Bing'20. There is an increase in the intensity of Fear in profiles of Bing'25 compared to their corresponding profile of Bing'20, although the increase is significant only when comparing EP^L(B-Em$^+$) to EP^L(QL-IT-Em$^+$). Intensities of negative emotions

[4]https://spacy.io/models/it
[5]https://www.nltk.org/index.html
[6]https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1
[7]https://huggingface.co/MilaNLProc/feel-it-italian-sentiment
[8]https://huggingface.co/MilaNLProc/feel-it-italian-emotion

Table 2: Emotional profile of Bing'20 based on EmoLex (as in [26]) and LexIT, along with the profile for Bing'25 based on LexIT; sentiment/emotion vectors sum up to 100%.

	Strategy			Sentiment			Emotion							
				Positive	Negative	Neutral	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
QL-IT	EmoLex	Queries	Overall	6.02	16.38	61.54	0.45	1.45	0.00	31.57	4.98	5.53	4.76	7.10
		Response	Overall	13.65	11.33	71.15	4.59	8.18	0.96	22.71	6.08	2.08	6.80	21.27
	LexIT	Queries	Overall	15.6	16.38	68.02	13.48	5.86	0.3	37.08	6.55	5.08	0	18.92
		Response	Overall	15.7	14.78	69.52	6.57	13.97	2.08	18.65	7.78	5.41	10.01	35.46
			Rank 1	14.83	13.91	71.26	6.71	15.7	1.85	20.63	6.53	5.75	8.58	34.25
			Rank 2	15.86	14.75	69.38	7.09	14.34	1.9	17.07	8.09	5.69	10.56	34.88
			Rank 3	15.84	14.77	69.39	6.57	14.33	2.44	18.55	8.9	5.42	9.73	34.07
			Rank 4	15.96	15.24	68.8	5.09	12.43	1.98	18.66	7.36	5.71	12.14	36.63
			Rank 5	16.04	15.24	68.72	7.37	12.99	2.23	18.31	8.05	4.51	9.04	37.5
B-G	LexIT	Response	Overall	15.65	14.78	69.57	6.99	11.79	2.78	19.2	7.78	6.06	7.4	38.01
			Rank 1	15.41	14.64	69.95	8.44	14.45	3.08	18.69	7.32	5.62	6.66	35.74
			Rank 2	15.58	14.7	69.72	6.46	11.74	4.3	18.87	7.93	6.19	6.62	37.88
			Rank 3	15.75	14.91	69.34	5.48	11.01	2.24	20.83	8.71	5.9	8.66	37.17
			Rank 4	15.73	14.8	69.47	6.88	11.92	2.3	19.5	6.13	5.75	7.68	39.85
			Rank 5	15.77	14.86	69.38	7.67	9.83	1.97	18.09	8.8	6.83	7.39	39.41
QL-IT-Em	EmoLex	Queries	Overall	3.58	2.98	67.22	1.67	1.19	0.24	1.43	0.00	7.15	0.00	0.00
			Positive Task	7.65	0.00	65.82	0.00	0.00	0.00	0.00	0.00	12.24	0.00	0.00
			Negative Task	0.00	5.59	68.46	3.13	2.24	0.45	2.68	0.00	2.68	0.00	0.00
		Response	Overall	15.79	7.35	68.93	6.85	8.76	5.83	8.39	8.10	8.60	4.61	20.13
			Positive Task	19.74	5.02	60.95	2.99	9.15	2.14	4.91	12.91	10.20	4.39	22.44
			Negative Task	12.11	9.72	75.93	10.24	8.41	9.07	11.44	3.88	7.20	4.81	18.10
	LexIT	Queries	Overall	11.48	11.43	77.08	5.98	4.55	0	5.18	8.13	0	22.73	30.71
			Positive Task	11.29	9.71	79	2.97	0	0	1.82	13.65	0	0	44.06
			Negative Task	11.53	12.33	76.14	9.09	6.06	0	8.59	3.03	0	45.45	18.69
		Response	Overall	15.53	14.57	69.9	4.94	12.93	2.84	10.81	7.39	5.21	18.91	36.97
			Positive Task	15.52	13.38	71.11	4.52	11	3.29	5.09	10.76	6.53	15.28	43.53
			Negative Task	15.53	15.63	68.83	5.32	14.64	2.44	15.87	4.41	4.04	22.12	31.16
			Rank 1	15.6	14.5	69.9	5.82	12.21	3.28	8.91	8.4	5.21	20.19	35.98
			Rank 2	15.51	14.5	69.98	4.29	16.21	2.61	11.32	5.66	3.09	18.36	38.47
			Rank 3	15.59	14.44	69.97	5.87	13.21	2.81	10.76	8.54	4.78	16.82	37.21
			Rank 4	15.46	14.83	69.72	4.34	10.49	3.23	11.52	8.12	5.96	20.6	35.72
			Rank 5	15.48	14.6	69.93	4.38	12.56	2.25	11.57	6.22	7.01	18.54	37.48
B-Em	LexIT	Response	Overall	15.6	14.55	69.85	3.16	14.23	1.46	11.66	5.87	6.34	17.07	40.21
			Positive Task	16.1	13.5	70.41	3.83	14.32	2.16	6.73	8.07	8.35	9.64	46.9
			Negative Task	15.08	15.48	69.44	2.57	13.59	0.74	16.62	3.15	4.54	24.29	34.5
			Rank 1	15.37	14.36	70.27	4.71	15.18	1.69	10.9	5.59	6.12	17.4	38.41
			Rank 2	15.39	14.82	69.79	3.01	15.28	1.29	10.87	6.74	6.9	14.58	41.33
			Rank 3	15.91	14.79	69.31	2.56	13.31	0.85	11.6	7.23	6.58	18.32	39.56
			Rank 4	15.53	14.31	70.16	2.85	13.26	1.24	13.66	5.98	7.01	16.01	40
			Rank 5	15.78	14.49	69.73	2.69	14.12	2.22	11.26	3.79	5.09	19.07	41.75

Anger and Disgust are significantly lower in EP^L(B-Em$^-$) compared to EP^L(QL-IT-Em$^-$). Looking at the rank-level emotional profiles of Bing'25, we do not observe significant variations across the ranks.

Overall, responses produced by Bing still generally portray a neutral and trustworthy tone for children's queries; the intensity of negative emotions for responses to queries related to Emotionally-Charged inquiry tasks has considerably reduced over the past 5 years. However, the opposite is true for the responses to queries related to General tasks, where the negative emotions conveyed on SERP results are significantly stronger than before.

3.3 The Effect of Context on Emotion Inference

The original InsideOut Framework used EmoLex, which generates the emotional vector of a text sample through word-level emotion detection. Although strategies like EmoLex continue to be popular amongst the research community due to their simple implementation [32], newer detection strategies tend to consider contextual factors by analysing the text in its entirety for emotional intensity detection. This prompts our exploration of the impact of using context-based strategies as a core component of InsideOut Framework, as opposed to lexicon-based ones. For this, we build EP^F(B-G) and EP^F(B-Em). As LexIT and FEEL-IT detect different sets of sentiments and emotions, it is not possible to do a direct profile comparison based on these strategies [9]. Instead, we discuss observed direct trends in the emotional tones conveyed by Bing'25, as inferred using FEEL-IT and LexIT.

EP^F(B-G), EP^F(B-Em), and EP^F(B-Em$^-$) (Table 3) show a high intensity for Negative sentiment; EP^F(B-Em$^+$) a higher intensity for Positive. Fear and Sadness are the salient emotions, except in EP^F(B-Em$^+$) where Joy has the highest intensity. From a rank-level perspective, there are no major variations across positions,

[9]The Pearson correlation coefficient across vector dimensions common to LexIT and FEEL-IT strategies yielded a positive correlation for all dimensions. The strongest correlations were for Fear and Joy, but even then, the coefficient value was at most 0.5.

Table 3: Emotional profiles of Bing'25 based on FEEL-IT; sentiment/emotion vectors sum up to 100%.

			Sentiment		Emotion			
			Positive	Negative	Anger	Fear	Joy	Sadness
B-G	Queries	Overall	26.42	73.58	10.95	27.7	26.54	34.81
	Response	Overall	32.84	67.16	1.82	41.8	26.22	30.15
		Rank 1	32.49	67.51	1.78	40.43	26.51	31.28
		Rank 2	29.6	70.4	3.76	43.75	22.47	30.02
		Rank 3	31.75	68.25	1.31	42.44	23.99	32.26
		Rank 4	34.5	65.5	0.75	39.78	28.91	30.56
		Rank 5	35.86	64.14	1.52	42.6	29.23	26.65
B-Em	Queries	Overall	20.12	79.88	33.38	17.6	28.56	20.46
		Positive Task	40.07	59.93	38.38	3.28	55.78	2.56
		Negative Task	0.03	99.97	29.52	31.97	0.11	38.4
	Response	Overall	39.4	60.6	9.84	23.21	35.73	31.22
		Positive Task	75.46	24.54	8.62	6.03	66.01	19.33
		Negative Task	2.6	97.4	11.32	40.57	4.44	43.67
		Rank 1	39.34	60.66	13.13	16.37	28.84	41.66
		Rank 2	33.33	66.67	10.97	22.98	33.14	32.91
		Rank 3	42.93	57.07	6.22	26.04	41.72	26.02
		Rank 4	43.51	56.49	7.03	26.41	39.79	26.76
		Rank 5	37.89	62.11	11.85	24.25	35.17	28.73

although $EP_{rank1}{}^{F}$(B-G) is closest to EP^{F}(B-G), and $EP_{rank1}{}^{F}$(B-Em) is closest to EP^{F}(B-Em) compared to the profiles for other ranking positions. Overall, compared to LexIT, FEEL-IT tends to amplify the intensities of salient sentiments and emotions.

3.4 The Spotlight on LLM

SE are no longer the only IAS children use to access information. Children now favour the direct responses generated by online tools like ChatGPT over the list of results produced by an SE [37]. Mindful of this preference shift, we apply the InsideOut Framework to compute the emotional profiles of ChatGPT and Gemma using both LexIT and FEEL-IT, and conduct element-wise comparisons across different pairs of profiles produced based on a given strategy with the independent T-test (p<0.05) for statistical significance.

We see in the profiles reported in Table 4 that all four profiles, i.e., EP^{L}(GEM-G), EP^{L}(GEM-Em), EP^{L}(GPT-G) and EP^{L}(GPT-Em) are mostly Neutral with EP^{L}(GEM-Em) portraying a significantly stronger Neutral sentiment than EP^{L}(GPT-Em). Trust is a salient emotion across all profiles with a significantly stronger intensity in EP^{L}(GEM-Em) than in EP^{L}(GPT-Em), and also when compared to EP^{L}(GEM-G). Comparing EP^{L}(GPT-G) and EP^{L}(GPT-Em), the former has a higher intensity for Trust but the difference is not significant. Through independent paired comparisons of EP^{L}(GPT-Em^{+}), EP^{L}(GPT-Em^{-}), EP^{L}(GEM-Em^{+}), and EP^{L}(GEM-Em^{-}), we find significant differences across intensities for Fear, Joy, and Trust. EP^{L}(GPT-Em^{+}) and EP^{L}(GEM-Em^{+}) have a significantly higher intensity for Joy and Trust, than EP^{L}(GPT-Em^{-}) and EP^{L}(GEM-Em^{-}) respectively, in contrast, Fear is significantly higher in EP^{L}(GPT-Em^{-}) and EP^{L}(GEM-Em^{-}) compared to the other two profiles.

In the four FEEL-IT-based profiles in Table 5, i.e., EP^{F}(GPT-G), EP^{F}(GPT-Em), EP^{F}(GEM-G), and EP^{F}(GEM-Em), the Negative sentiment has a higher intensity than the Positive. Juxtaposing EP^{F}(GPT-G) with EP^{F}(GPT-Em), and EP^{F}(GEM-G) with EP^{F}(GEM-Em) we see that the profiles for Emotionally-Charged tasks have a significantly higher intensity for Negative sentiment than the those for General tasks. In terms of emotions, both EP^{F}(GEM-G) and EP^{F}(GEM-Em) have the strongest intensity for Sadness and Joy, with the intensity of Sadness significantly higher and Joy significantly lower in EP^{F}(GEM-Em) than in EP^{F}(GEM-G). For ChatGPT, Fear is the most salient emotion in EP^{F}(GPT-G) and EP^{F}(GPT-Em), with higher, but not significant, intensity in EP^{F}(GPT-G). Further, the intensity of the Negative sentiment is significantly higher in EP^{F}(GPT-Em^{-}) and EP^{F}(GEM-Em^{-}) than in EP^{F}(GPT-Em^{+}) and EP^{F}(GPT-Em^{+}), respectively. In EP^{F}(GEM-Em^{+}), the salient emotion is Joy; in EP^{F}(GEM-Em^{-}) Sadness followed by Fear. In the case of EP^{F}(GPT-Em^{-}), Fear has the highest intensity; oddly in EP^{F}(GPT-Em^{+}) both Joy and Sadness—two fundamentally contrasting emotions—have similarly strong intensities.

Overall, the sentiment polarity of LLM is not unique to open- or close-sourced LLM. Contextualizing the profiles at LLM and query level revealed that the salient sentiment observed in the emotional profiles of the two LLM was a result of the LLM *reflecting* the most salient sentiment of the children's queries which is Neutral in LexIT-based profiles, and Negative sentiment in FEEL-IT-based profiles. Based on the profiles for Emotionally-Charged tasks, it emerges that the target sentiment of the task led both LLM to generate responses that *amplified* the target sentiment, i.e., Positive tasks led to even more positive responses and those related to Negative tasks led to more negative responses.

3.5 The Difference across IAS

SE and LLM agents handle users' inquiries in different ways: SE provide searchers with a list of ranked web resources to browse, whereas LLM agents offer a singular and direct response. This motivates us to compare and contrast the emotional profiles of Bing'25, ChatGPT, and Gemma (one-way ANOVA for significance testing with p<0.05) produced based on both LexIT and FEEL-IT.

Scrutiny of the LexIT-profiles in Table 2 and Table 4 reveal no prominent differences across Bing'25, ChatGPT and Gemma. The common trend across all the profiles is that the responses of

Table 4: Emotional profiles for LLM based on LexIT; sentiment/emotion vectors sum up to 100%.

			Sentiment			Emotion						
			Positive	Negative	Neutral	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
	Queries	Overall	15.6	16.38	68.02	13.48	5.86	0.3	37.08	6.55	5.08	0	18.92
GEM-G	Response	Overall	16.25	15.77	67.98	10.17	9.54	0.39	23.19	6	5.28	5.37	35.31
GPT-G	Response	Overall	16.46	14.86	68.68	5.71	9.87	2.82	16.92	7.85	4.39	7.62	44.82
		Overall	11.48	11.43	77.08	5.98	4.55	0	5.18	8.13	0	22.73	30.71
	Queries	Positive Task	11.29	9.71	79	2.97	0	0	1.82	13.65	0	0	44.06
		Negative Task	11.53	12.33	76.14	9.09	6.06	0	8.59	3.03	0	45.45	18.69
		Overall	14.14	13.9	71.96	1.43	7.65	0.45	12.18	5.16	1.3	17.48	53.13
GEM-Em	Response	Positive Task	13.79	13.3	72.91	2.16	3.85	0.51	3.53	8.93	2.22	13.03	65.76
		Negative Task	14.46	14.44	71.1	0.78	11.1	0.39	20.02	1.74	0.47	21.5	41.67
		Overall	15.99	15	69.01	3.74	10.6	1.54	17.2	4.33	4.98	13.73	43.89
GPT-Em	Response	Positive Task	16.3	13.85	69.86	3.52	9.56	1.82	10.72	6.62	5.69	11.87	50.2
		Negative Task	15.72	16.04	68.24	3.94	11.55	1.29	23.08	2.24	4.34	15.41	38.16

Table 5: Emotional profiles of LLM based on FEEL-IT; sentiment/emotion vectors sum up to 100%.

			Sentiment		Emotion			
			Positive	Negative	Anger	Fear	Joy	Sadness
	Queries	Overall	46.67	53.33	4.42	29.99	40.07	25.51
GEM-G	Response	Overall	40.17	59.83	3.92	21.09	33.34	41.65
GPT-G	Response	Overall	38.57	61.43	1.76	50.32	26.22	21.7
		Overall	20.12	79.88	33.38	17.6	28.56	20.46
	Queries	Positive Task	40.07	59.93	38.38	3.28	55.78	2.56
		Negative Task	0.03	99.97	29.52	31.97	0.11	38.4
		Overall	23.72	76.28	15.33	19.19	28.08	37.4
GEM-Em	Response	Positive Task	47.28	52.72	17.62	3.34	56.24	22.8
		Negative Task	2.35	97.65	13.26	33.56	2.54	50.63
		Overall	25.77	74.23	2.46	44.55	17.31	35.69
GPT-Em	Response	Positive Task	54.08	45.92	2.39	23.47	36.26	37.88
		Negative Task	0.09	99.91	2.51	63.68	0.11	33.7

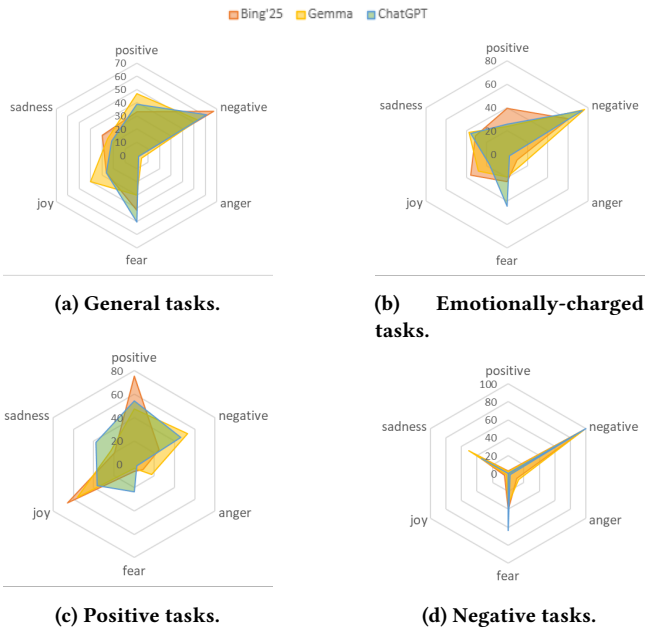
all three IAS are predominantly Neutral and Trust is the salient emotion. On the other hand, skewed distributions in the profiles based on FEEL-IT, as presented in Table 3 and Table 5, highlight prominent differences across the considered IAS. This is perhaps more evident in Figure 3, where we observe that EP^F(GEM-G) has significantly higher intensities for Positive sentiment and Joy compared to EP^F(B-G) and EP^F(GPT-G); whereas EP^F(B-G) has a higher intensity for Sadness compared to the other profiles (although the difference is not significant). For emotionally-charged tasks, EP^F(B-Em) has a significantly higher intensity for Positive sentiment and Joy compared to EP^F(GEM-Em) and EP^F(GPT-Em); Fear in EP^F(GPT-Em) has significantly higher intensity compared to EP^F(B-Em) and EP^F(GEM-Em).

Juxtaposing the IAS profiles for search sessions related to Positive tasks, we find that EP^F(GPT-Em^+) has a significantly higher intensity for Sadness compared to EP^F(B-Em^+) and EP^F(GEM-Em^+); while EP^F(B-Em^+) has significantly higher intensities for Positive sentiment and Joy compared to the other two corresponding profiles. EP^F(B-Em^+) also has a significantly lower intensity for Negative sentiment compared to the corresponding LLM profiles. For the search sessions related to Negative tasks, EP^F(B-Em^-), EP^F(GPT-Em^-) and EP^F(GEM-Em^-) have a high intensity for the Negative sentiment with Fear and Sadness being the most salient emotions. IAS profiles pertaining to sessions related to Negative tasks show no notable differences except for the significantly low intensity of Anger in EP^F(GPT-Em^-) compared to the counterpart profiles.

4 Discussion

Reflecting on the experimental results reported in Section 3, we discuss findings, implications, and suggestions for future work.

Emotion Detection Strategies. Adopting native Italian lexicons in the InsideOut Framework impacted Bing's emotional profile by intensifying the detected sentiments and emotions across responses, for General and Emotionally-Charged tasks alike. The higher intensity values reflected a more expressive emotional profile of the SE which indicates that native lexicons are better suited for capturing the emotional tone of an SE responding in the target



(a) General tasks.

(b) Emotionally-charged tasks.

(c) Positive tasks.

(d) Negative tasks.

Figure 3: Emotional profiles of SE and LLM using FEEL-IT.

language. A "one-fits-for-all" approach, therefore, presents a limitation, especially from the language perspective, as English-centric strategies, like EmoLex, might fail to capture the full range of emotional nuances of non-English languages even after the strategy is adapted to the target language through translations. This calls for adaptable models that consider specific lexical expressions, idioms, and cultural contexts of each language. This would facilitate more accurate analyses of sentiment and emotion detection, leading to expanding the research area to other countries and extending the IR community's knowledge in non-English cultural contexts as well.

While changing the lexicon was useful in capturing the emotional nuances of the Italian language, the emotional profiles produced portrayed similar tones. The profiles generated using FEEL-IT—that analyses text samples as a whole rather than as a collection of isolated words—revealed prominent differences in emotional tones across IAS. These profiles, however, capture intensities based only on two sentiments and four emotions as opposed to three sentiments and eight emotions yielded by LexIT or EmoLex. The two types of emotion detection strategies present a tradeoff– the lexicon-based strategies generate a diverse profile with weak intensity across the emotions, while FEEL-IT amplifies the intensity, allowing for an in-depth scrutiny of the captured emotions, but at the cost of a less diverse profile. Therefore, we suggest that the choice of emotion detection strategy used in the InsideOut Framework is driven by the intended purpose behind producing the emotional profiles. For instance, in our study, FEEL-IT was particularly useful as it amplified negative emotions in IAS responses, which highlights a concerning trend of children's exposure to distressing content during online search.

Time. In reproducing the InsideOut Framework and juxtaposing the emotional tone of the past and current versions of Bing when responding to children's queries, we observed that the current version of Bing responds more negatively as compared to the past, with a notable increase to the intensity of Fear. For example, in 2020, for queries about tornados and their formation (e.g., *"Cosa sono e come nascono i tornadi?"*), which is classified as a General topic in the primary school curriculum, Bing's emotional profile was mostly Neutral, with Surprise as the dominant emotion, followed by Fear. In 2025, Bing's emotional profile in response to the same query still has a Neutral undertone, but the intensity of Fear has increased, followed by Anticipation, Sadness, and Trust. This decrease in positive emotions indicates that children are now being exposed to stronger negative emotions when conducting General search tasks. Fortunately, in the case of children's queries pertaining to Emotionally-Charged tasks, Bing's responses show the opposite trend. Negative emotions are reduced, leading to a more balanced emotional distribution of Bing's responses in 2025. For example, for the query *"Perchè gli orsi polari sono a rischio di estinzione?"*, formulated by a child to search for reasons for the extinction of polar bears—a task related to a distressing topic of extinction—Bing's current response shows a slight increase in positive emotions like Joy and Trust, and a slight decrease in negative emotions such as Disgust and Anger, compared to the SE's response in the past. These findings highlight an important aspect regarding the lack of necessary adjustments to the emotional tone of Bing's responses over time. The dominance of Fear is particularly concerning as children are more susceptible to emotional content and, considering the rise in their online presence [24], strong intensities of such negative emotions in Bing's responses do not align with children's emotional needs and instead would expose the young searchers to unnecessary distress. Although eliminating all negative emotional content is not desirable, as suggested in the original study [26], moderating negative emotions could help children engage with more appropriate content, ensuring a safer and more engaging learning environment, which we fortunately found Bing capable of when at least responding to children's queries pertaining to Emotionally-Charged search tasks.

Emotional Diversity across IAS. Upon extending the InsideOut Framework to capture the emotional tone of ChatGPT and Gemma responses, we note the dominance of Negative sentiment in responses generated by either LLM for General inquiries. In the case of Emotionally-charged ones, however, we see that the emotional charge of the LLM responses is biased by the sentiment evoked by the task, i.e, Negative tasks resulted in LLM responding with strong Sadness and Fear intensities, Positive tasks led to LLM responses with high intensity of Joy. A similar pattern was noted in [26] when studying the role of emotions when children use SE in the classroom. In this case, the SERP results reflected the emotional profile of the child's query. In contrast, the LLM's direct response tends to amplify emotions, exposing children to a non-contextualized, single emotional voice, as opposed to the SE entries that can portray not necessarily uniform emotional tones in the list of results in their SERPs. This can have a stronger impact on children, particularly when exploring distressing topics like natural disasters or tragic historical events. Since these topics typically evoke negative emotions, children may formulate queries with a negative tone [26], which the LLM agent could then amplify.

Since in this study we explore the emotional tones a child would be exposed to when using an LLM "in the wild", the prompt we used to elicit LLM responses did not contain any information about the users themselves. Still, research has shown that prompt engineering does influence outcomes [29, 40]. For instance, ChatGPT can modify the complexity of its responses when explicitly instructed to produce text fitting the reading skills of a child [27, 33]. With that in mind, future work should gauge whether explicitly stating that the user is a child in the prompt impacts the emotional undertone of LLM responses. Further, with LLM driving response generation in emerging technologies like Retrieval-Augmented Generation (RAG) [45], a natural next step is examining the interplay of emotions conveyed by hybrid tools for information access–particularly when LLM could amplify emotional charges already present in the SERP.

Reproducing the Framework. From our experimental results, we infer that the emotional profiles are heavily dependent on the emotion detection strategy the framework adopts. `LexIT` pre-processes IAS responses by removing the stopwords from each response before computing the emotional vector of the response, but we cannot ascertain if the same was done in `EmoLex` due to the lack of details on the text-preprocessing steps in [26]. Thus, LexIT's pre-processing steps could have impacted the intensities of the sentiment and emotion computed by the strategy. By providing details and code for our study, we hope to eliminate the ambiguity for future reproductions and extensions of the `InsideOut Framework`.

Turning to the Scope Compass component, the environment in the original and this reproducibility study was the classroom. However, children's information access is not limited to this setting. Considering the difference in search behaviour when seeking information for learning or leisure, the emotional tone a child may be exposed to during informal search tasks also needs to be examined. Further, like in [26], we applied the `InsideOut Framework` to logs containing only textual content. However, with children turning more to voice-controlled assistants like Siri and Alexa, and social media platforms such as YouTube and TikTok, for information seeking [13, 16], these young searchers consume information in modalities that convey emotions through cues that cannot be captured in text. With that in mind, investigating strategies that can infer sentiments and emotions expressed in multimodal content (e.g., audio or images) becomes a must.

Emotions & SAL. Given the dominance of negative emotions in responses from LLM and Bing'25, it is evident that when designing tools to support children searching in the classroom *the emotional dimension needs full attention.* Since search is not just an aid to learning but learning itself [44], the aesthetic and algorithmic side of IAS must account for the emotional needs of children while also satisfying their information needs. Furthermore, children may not always have the necessary support to help them healthily navigate the effects of being exposed to amplified (negative) emotions of the responses generated by such tools [43]. This evinces the need to account for the emotional aspect when improving search literacy in children. Based on the results of our study, by enabling children to formulate more emotionally tempered queries, they can use these tools to explore controversial topics without being exposed to highly distressing emotional undertones.

The Search-As-Learning (**SAL**) paradigm places a high value on engagement. In educational settings, children are drawn to tools that evoke emotional engagement [26]. As children must read and interact with IAS responses to extract information that meets the needs expressed in their queries, the strong intensity of negative emotions in IAS responses could affect their engagement levels, potentially diminishing their motivation to interact with the content. This issue is connected to children's perceptions of the concept of relevance, which encompasses both the usefulness of the information and its motivational or emotional quality [23]. Given our findings on the emotional undertone of SE and LLM agents, the known concept of relevance in search tools [26] should be re-examined. It is crucial to assess whether the shift to LLM-based tools affects children's perceptions of relevance, ultimately impacting their ability to complete tasks effectively.

5 Conclusions

We successfully reproduced the `InsideOut Framework`, confirming that the choice of emotion detection strategy directly impacts the emotional profile generated for an IAS. From a reproducibility perspective, we find that culturally adaptable models that can capture the emotional nuances across different languages are beneficial. Regarding the IAS probed in this study, our findings reveal that Bing responses in general have become more negative in the past 5 years, with a notable increase in the intensity of Fear. In fact, upon extending the framework to LLM agents, we observe a prominent negative undertone in responses produced by different IAS, more so in the case of LLM agents, underscoring the need to ensure more emotionally appropriate content for children's search-as-learning experience. Particularly, with the growing popularity of LLM in educational settings [17, 42] and these models also driving the response generation in emerging hybrid technologies like RAG [45], it is crucial to explore prompt engineering strategies to temper the intensity of negative emotions in responses.

This work lays the foundations for empirical explorations of emotional profiles of different IAS but is based on synthetic data generated using real queries formulated by children. While the `InsideOut Framework` is inherently meant to examine system responses, a comprehensive understanding of emotions in the search process also requires considering the human aspect of search. With this in mind, involving children to gauge their reactions to emotions conveyed by different IAS could provide meaningful insights that can guide the development of future IR algorithms that consider the emotional impact of IAS, especially when interacting with vulnerable groups such as children.

Children are not the only demographic of digital users who are vulnerable to the emotional charges of IAS responses. For instance, IAS responses with strong negative emotional charges can be extremely distressing for users suffering from mental health disorders [30]. Therefore, the `InsideOut Framework` could also be used to examine the emotional profile of IAS response logs procured from search sessions involving different vulnerable demographics.

Acknowledgments

Supported by SNSF Award #[IC00I0-227887 project n.10000973 SOL]

References

[1] Eytan Adar, Jaime Teevan, Susan T Dumais, and Jonathan L Elsas. 2009. The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. 282–291.

[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. doi:10.1016/j.caeai.2021.100040

[3] Omar Alonso and Ricardo Baeza-Yates (Eds.). 2024. *Information Retrieval: Advanced Topics and Techniques* (1 ed.). Vol. 60. Association for Computing Machinery, New York, NY, USA.

[4] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. 2017. Online searching and learning: YUM and other search tools for children and teachers. *Information Retrieval Journal* 20 (2017), 524–545.

[5] Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

[6] Dania Bilal and Li-Min Huang. 2019. Readability and word complexity of SERPs snippets and web pages on children's search queries: Google vs Bing. *Aslib Journal of Information Management* 71 (03 2019). doi:10.1108/AJIM-05-2018-0124

[7] Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2016. A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 38–45.

[8] Gobinda Chowdhury and Sudatta Chowdhury. 0. AI- and LLM-driven search tools: A paradigm shift in information access for education and research. *Journal of Information Science* 0, 0 (0), 01655515241284046. doi:10.1177/01655515241284046 arXiv:https://doi.org/10.1177/01655515241284046

[9] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as learning (dagstuhl seminar 17092). (2017).

[10] Miguel Costa. 2021. Full-text and URL search over web archives. In *The past web: exploring web archives*. Springer, 71–84.

[11] Judith H. Danovitch, Adam K. Dubé, Cansu Oranç, Jessica Szczuka, and Svetlana Yarosh. 2025. *Children's Understanding and Use of Voice-Assistants: Opportunities and Challenges*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69362-5_84

[12] Luna De Bruyne. 2023. The Paradox of Multilingual Emotion Detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Jeremy Barnes, Orphée De Clercq, and Roman Klinger (Eds.). Association for Computational Linguistics, Toronto, Canada. doi:10.18653/v1/2023.wassa-1.40

[13] Adobe Express. 2024. Using TikTok as a search engine | Adobe Express. https://www.adobe.com/express/learn/blog/using-tiktok-as-a-search-engine

[14] Carlos Flavián-Blanco, Raquel Gurrea-Sarasa, and Carlos Orús-Sanclemente. 2011. Analyzing the emotional outcomes of the online search behavior with search engines. *Computers in Human Behavior* 27, 1 (2011), 540–551.

[15] Google AI for Developers. 2024. Gemma models overview. https://ai.google.dev/gemma/docs

[16] Katherine Haan. 2024. Is social media the new Google? gen Z turn to google 25% less than gen X when searching. https://www.forbes.com/advisor/business/software/social-media-new-google/

[17] Mohanad Halaweh. 2023. ChatGPT in education: Strategies for responsible implementation. *Contemporary educational technology* 15, 2 (2023).

[18] Kailash A. Hambarde and Hugo Proença. 2023. Information Retrieval: Recent Advances and Beyond. *IEEE Access* 11 (2023), 76581–76604. doi:10.1109/access.2023.3295776

[19] Hanna Jochmann-Mannak, Theo WC Huibers, Leo Lentz, and Ted Sanders. 2010. Children searching information on the Internet: Performance on children's interfaces compared to Google. In *Workshop on Accessible Search Systems 2010*.

[20] Thashmee Karunaratne and Adenike Adesina. 2023. Is it the new Google: Impact of ChatGPT on Students' Information Search Habits. (2023).

[21] Gabriella Kazai, Paul Thomas, and Nick Craswell. 2019. The Emotion Profile of Web Search *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1097–1100. doi:10.1145/3331184.3331314

[22] Kalin Kopanov. 2024. Comparative Performance of Advanced NLP Models and LLMs in Multilingual Geo-Entity Detection. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference* (undefinedstanbul, Turkiye) *(AICCONF '24)*. Association for Computing Machinery, New York, NY, USA, 106–110. doi:10.1145/3660853.3660878

[23] 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* (Siena, Italy) *(ECCE '21)*. Association for Computing Machinery, New York, NY, USA, Article 30, 5 pages. doi:10.1145/3452853.3452885

[24] Monica Landoni, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2024. Good for Children, Good for All?. In *European Conference on Information Retrieval*. Springer, 302–313.

[25] 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 *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer, 101–113.

[26] 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* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 136–144. doi:10.1145/3340631.3394847

[27] Max Z Li. 2024. Using Prompt Engineering to Enhance STEM Education. In *2024 IEEE Integrated STEM Education Conference (ISEC)*. IEEE, 1–2.

[28] Irene Lopatovska. 2014. Toward a model of emotions and mood in the online information search process. *Journal of the association for information science and technology* 65, 9 (2014), 1775–1793.

[29] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*. Springer, 387–402.

[30] 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.

[31] Saif M Mohammad. 2011. NRC Emotion Lexicon in Various Languages. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[32] Saif M Mohammad. 2020. Ten Years of the NRC Word-Emotion Association Lexicon. https://medium.com/@nlpscholar/ten-years-of-the-nrc-word-emotion-association-lexicon-eaa47a8dd03e

[33] Emiliana Murgia, Maria Soledad Pera, Monica Landoni, and Theo Huibers. 2023. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 311–316. doi:10.1145/3563359.3596996

[34] Diane Nahl. 2004. Measuring the affective information environment of web searchers. *Proceedings of the American Society for Information Science and Technology* 41, 1 (2004), 191–197. doi:10.1002/meet.1450410122 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.1450410122

[35] Lucia Passaro, Alessandro Bondielli, Alessandro Lenci, et al. 2016. FB-NEWS15: A topic-annotated Facebook corpus for emotion detection and sentiment analysis. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. Accademia University Press, 228–232.

[36] Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating Context Selection Strategies to Build Emotive Vector Space Models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/637.html

[37] Ellie Prosser and Matthew Edwards. 2024. Helpful or Harmful? Exploring the Efficacy of Large Language Models for Online Grooming Prevention. In *European Interdisciplinary Cybersecurity Conference*. 1–10.

[38] Antonio J Roa-Valverde and Miguel-Angel Sicilia. 2014. A survey of approaches for ranking on the web of data. *Information Retrieval* 17, 4 (2014), 295–325.

[39] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. 2008. The Google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, Vol. 60. Emerald Group Publishing Limited, 290–310.

[40] Douglas C Schmidt, Jesse Spencer-Smith, Quchen Fu, and Jules White. 2023. Cataloging prompt patterns to enhance the discipline of prompt engineering. *URL: https://www. dre. vanderbilt. edu/˜ schmidt/PDF/ADA_Europe_Position_Paper. pdf [accessed 2023-09-25]* (2023).

[41] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 903, 20 pages. doi:10.1145/3613904.3642152

[42] Olivia Sidoti and Jeffrey Gottfried. 2023. About 1 in 5 U.S. teens who've heard of ChatGPT have used it for schoolwork. https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-us-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/

[43] Aída Walqui. 2006. Scaffolding Instruction for English Language Learners: A Conceptual Framework. *International Journal of Bilingual Education and Bilingualism* 9, 2 (2006), 159–180. doi:10.1080/13670050608668639

arXiv:https://doi.org/10.1080/13670050608668639

[44] Xinyue Wang and Chang Liu. 2023. Finding the Aha! Moment of Search: A Preliminary Examination of Insight Learning During Search. *Proceedings of the Association for Information Science and Technology* 60, 1 (2023), 421–432. doi:10.1002/pra2.800 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.800

[45] Ryen W. White. 2024. Advancing the Search Frontier with AI Agents. *Commun. ACM* 67, 9 (Aug. 2024), 54–65. doi:10.1145/3655615

[46] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136. doi:10.1109/JAS.2023.123618