



# Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models

Michael Xieyang Liu\*

Carnegie Mellon University, USA  
xieyangl@cs.cmu.edu

Franklin Mingzhe Li

Carnegie Mellon University, USA  
mingzhe2@cs.cmu.edu

Tongshuang Wu

Carnegie Mellon University, USA  
sherryw@cs.cmu.edu

Tianying Chen

Carnegie Mellon University, USA  
tianyinc@andrew.cmu.edu

Brad A. Myers

Carnegie Mellon University, USA  
bam@cs.cmu.edu

## ABSTRACT

Sensemaking in unfamiliar domains can be challenging, demanding considerable user effort to compare different options with respect to various criteria. Prior research and our formative study found that people would benefit from reading an overview of an information space upfront, including the criteria others previously found useful. However, existing sensemaking tools struggle with the “cold-start” problem — it not only requires significant input from previous users to generate and share these overviews, but such overviews may also turn out to be biased and incomplete. In this work, we introduce a novel system, Selenite, which leverages Large Language Models (LLMs) as reasoning machines and knowledge retrievers to automatically produce a comprehensive overview of options and criteria to jumpstart users’ sensemaking processes. Subsequently, Selenite also adapts as people use it, helping users find, read, and navigate unfamiliar information in a systematic yet personalized manner. Through three studies, we found that Selenite produced accurate and high-quality overviews reliably, significantly accelerated users’ information processing, and effectively improved their overall comprehension and sensemaking experience.

## CCS CONCEPTS

- Human-centered computing → Interactive systems and tools.

## KEYWORDS

Human-AI Collaboration, Sensemaking, Large Language Models, Natural Language Processing

### ACM Reference Format:

Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A. Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing*

\*The author is currently affiliated with Google Research and can be contacted at lxieyang@google.com.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.

*Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3613904.3642149>*

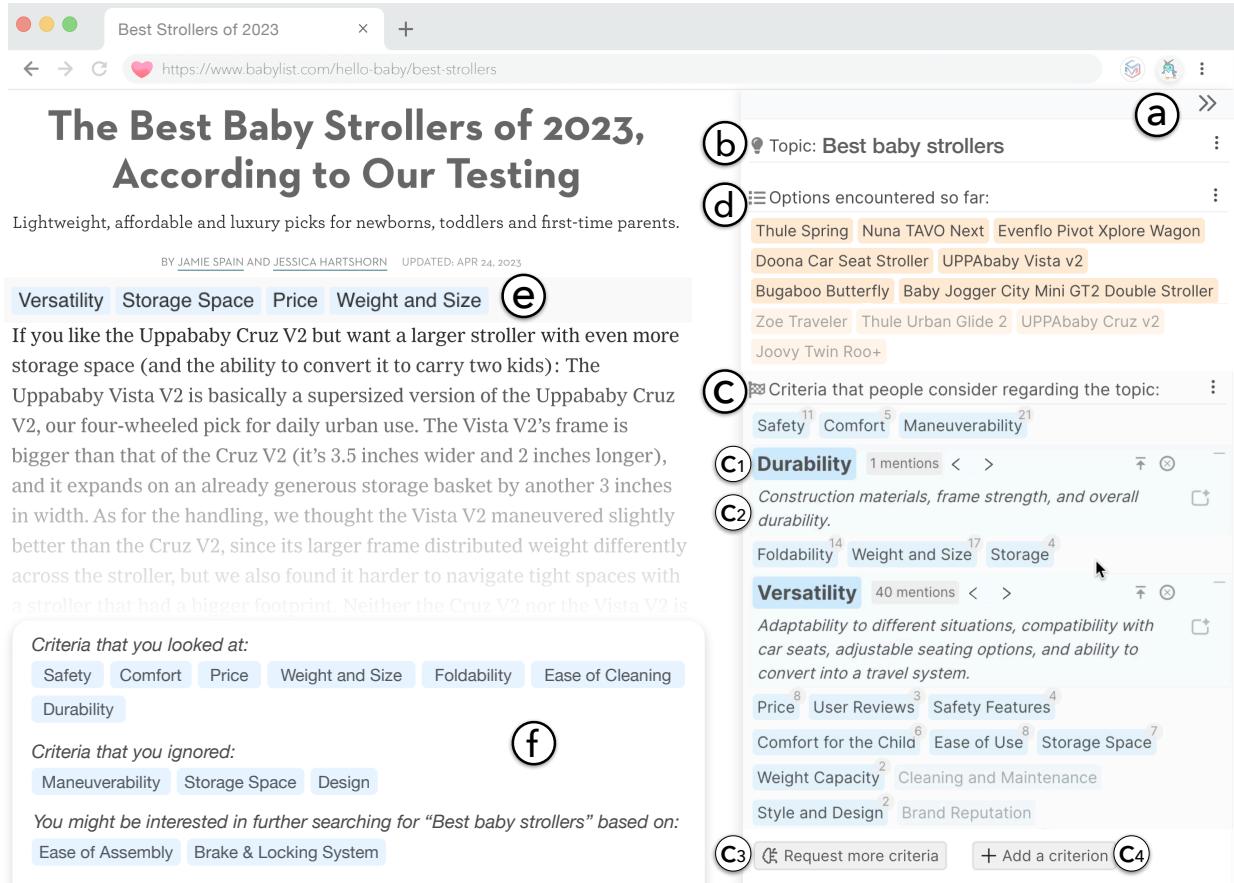
## 1 INTRODUCTION

Whether it is parents delving into the vast sea of baby stroller choices or developers comparing different JavaScript frontend frameworks, people frequently find themselves having to make sense of unfamiliar domains and performing comparisons to make informed decisions. In these situations, people often have to iteratively find, read, collect, and organize large amounts of information about different options with respect to various criteria [56, 57, 69], which can become a messy and overwhelming experience [15, 72].

One challenge lies within the initial reading process — due to unfamiliarity with the topic, people may struggle to fully understand certain content or fail to recognize important aspects that should otherwise warrant their attention [60, 89] as they read, leading to a limited viewpoint or ultimately misguided decisions [44, 114]. For example, novice developers might not be aware of criteria crucial to the utility of a software library, such as its stability, community size and support, or ease of integration with existing codebases, resulting in making a sub-optimal choice.

Another challenge arises when people have to sift through numerous online reviews and comparison articles but with limited time and cognitive bandwidth, making a complete understanding of the information space impractical or impossible. Instead, people often adopt a “selective” (or non-linear) reading strategy [109], only reading the paragraphs that discuss information that they consider relevant or valuable and bypassing the rest [26, 108], e.g., in a focused session where they would like to compare different options with respect to the same criterion. However, it is challenging for people to gauge the potential value of articles or paragraphs, especially long-winded ones, just by skimming and without performing a more thorough read [12, 35, 39]. For instance, information snippets about the maneuverability of different baby strollers may be dispersed throughout a review and appear in diverse variations (e.g., “agile enough to go through tight spots” and “easy to steer and navigate small corners” both discuss “maneuverability”), making it difficult for people to spot these variations effectively and navigate efficiently among such scattered details.

In response to these difficulties, prior work on sensemaking and knowledge reuse has shown promising evidence that people would benefit from seeing an *overview* of the information space *before* they dive into the sensemaking process [16, 37, 80, 85, 99] — for example, Kittur et al. reported that having read an overview of the criteria



**Figure 1: The main user interface of Selenite, which provides users with a comprehensive overview of the information space in the sidebar (a). When users encounter an unfamiliar topic (b), Selenite offers them a global overview based on commonly considered criteria (c) as well as the options encountered so far (d), helping them develop quick intuitions of the topic. As users read articles that they haven't seen before, Selenite provides local grounding through page-level and paragraph-level summaries and annotations (e), enabling effective comprehension and efficient navigation between the content of their interests. Before leaving a page, Selenite dynamically summarizes users' progress and suggests avenues for finding additional new information (f) in subsequent searches.**

that earlier users found useful can help people build intuition and understanding of the decision space upfront, leading to significantly improved digestion of the source material, better-structured mental models, and ultimately more well-informed decisions [56, 57]. Similarly, our formative study found that people expressed a desire for such comprehensive overviews to help them more systematically read, understand, and strategically navigate unfamiliar information. However, existing sensemaking systems have struggled with the “cold start” issue — they often require substantial effort from the current user to personally go through the unfamiliar content and gather and structure information to obtain such an overview [6, 55, 57, 70, 96], which defeats the premise of receiving one upfront. And even if a previous user has generated such an overview, it can often be incomplete [71, 85], biased [38, 51], or in idiosyncratic formats that make it hard for the current user to readily understand and leverage [52, 53, 68, 70].

To overcome these challenges and go beyond prior systems, we explore the idea of providing users with a comprehensive overview

of the information space upfront to jumpstart as well as guiding their subsequent sensemaking processes in a novel system named Selenite.<sup>1</sup> At a high-level, when users encounter an unfamiliar topic, Selenite leverages GPT-4, an LLM developed by OpenAI, as a knowledge retriever to offer them a global grounding based on commonly considered criteria, helping users develop quick intuitions of the topic (Figure 1c). As users read new articles, Selenite contextualizes that overview and uses it as an index to help users effectively comprehend and efficiently navigate among the content of their interest (Figure 1e). Upon leaving a page, Selenite dynamically summarizes users’ progress and suggests unique search queries that would help users find additional information and expand their perspectives rather than duplicating existing knowledge (Figure 1f). Through an intrinsic evaluation of Selenite, we verified its feasibility to provide a sufficiently accurate and high-quality global overview to users. Furthermore, additional usability and case studies revealed that

<sup>1</sup>Selenite is named after a soft and transparent gemstone, and stands for “Smart Environment for Logical Extraction and Navigation of Information using Technological Expertise.”

Selenite accelerated users' information processing, facilitated their comprehension, and improved their overall reading and sensemaking experience. The contributions described in this work include:

- a formative study showcasing people's barriers and needs when reading and understanding information during online sensemaking, despite recent advances in information management tools,
- Selenite, a novel system providing users with an upfront comprehensive overview of the information space as well as interactively guiding their subsequent reading and sensemaking processes,
- as part of Selenite, a novel user interface to interact with LLM-generated content and a novel user experience to contextualize that content into people's existing sensemaking workflows, going beyond the widely-adopted conversational interfaces in generative AI research and applications [3, 9, 13, 76, 83, 110].
- an intrinsic evaluation of Selenite that demonstrates the feasibility of our approach, as well as usability and case studies that offer preliminary insights into its usefulness and effectiveness,
- a discussion of design implications for future LLM and AI-powered sensemaking systems and tools.

## 2 RELATED WORK

### 2.1 Making Sense of Online Information

Building on theories of sensemaking as defined as developing a mental model of an information space in service of a user's goals [28, 59, 95], at a high level, a typical online sensemaking process involves first *reading and understanding* information from various data sources (e.g., online comparison articles, blog posts, video reviews, etc.), and then *collecting and organizing* such information to form a schema or representation (e.g., a comparison table, a decision tree, etc.) to interpret the space [15, 87]. Notably, there have been many research as well as commercial tools and systems developed to support this latter stage of sensemaking, i.e., collecting and organizing information. For example, tools like SearchPad [8], Hunter Gather [96], as well as commercial systems like the Evernote clipper [33], allow a user to capture entire pages or portions of web content, categorize them, and later assemble them into a coherent document or structure for their own sensemaking, decision making, or sharing and collaboration. Furthermore, to reduce the disruption to users' overall workflow, prior work has also explored using lightweight interactions [14, 20, 52, 73] to streamline the process of collection and triaging information, or even automatically keeping track of content of interest on behalf of the user [17, 31, 50, 72].

However, an inherent assumption that these prior systems make is that the user has a well-developed mental model that will allow them to effectively grasp the content that they are about to collect using the system. But, in many circumstances, building sufficient context about the information space can be both time and effort-intensive, requiring individuals to iteratively read through and understand unfamiliar content by themselves without expert guidance or systematic strategy, as has been discussed in the literature [82, 98] and confirmed in our formative study. Despite being feature-rich, aforementioned sensemaking tools generally fail to address the "cold start" issue during the initial reading and comprehension stage where users need help deciding "what is important to read" [39] and

"how to most effectively read it" [5, 46, 92]. Discovering new and unseen content after reading can also be surprisingly difficult, as prior work revealed that a generic search query could return many redundant and near-duplicate documents [30, 86, 90] despite the extensive use of duplicate detection algorithms in modern search engines [88]. Therefore, in this work, we focus on supporting users' reading and understanding of unfamiliar content in the first place.

### 2.2 Tool Support for Reading and Comprehension

One crucial way to address the "cold start" issue is to provide users with a summary of the information landscape. Previous research has investigated mechanisms that prompt domain experts to iteratively and collectively summarize a single document or discussion thread [40, 116, 117]. In addition, prior work has introduced tools and systems that facilitate finding, extracting, and summarizing relevant information from multiple documents and sources into cohesive themes and knowledge [15, 16, 70]. However, these methods are dependent on the significant investment of time and effort from previous users and are contingent upon their availability, and may still result in summaries that are limited in scope [32, 71, 85, 116] or with a biased perspective [38, 51].

More recently, researchers have been experimenting with approaches to extract key information and build underlying knowledge structures automatically with machine learning [2, 10, 27, 39, 54], lexical and HTML patterns [31, 61, 72], or crowdsourcing [18, 22, 42]. However, many studies have demonstrated that these techniques can often produce structures that are incoherent and difficult for users to comprehend and contextualize [18, 24, 47].

To address these limitations, Selenite leverages the wealth of world knowledge embedded within LLMs to generate overviews of information spaces that are aimed at being comprehensive and high-quality, eliminating the need for manual effort. In addition, Selenite provides detailed explanations of the generated overview as well as contextualizes it within the original content of the source documents to further enhance user understanding and facilitate navigation.

### 2.3 Eliciting Knowledge from LLMs

Recent advances in LLMs like GPT-4 [84], PaLM [23], and LLaMa [103] showcase impressive capabilities in answering user questions. These models are trained on large volumes of data, and as a result, their parameters might contain a significant body of factual as well as synthesized knowledge across a wide range of domains [25, 106]. In this work, we leverage GPT-4 as a knowledge retriever to retrieve a list of commonly considered aspects given an arbitrary topic, which is used to directly help users understand that topic at the beginning and systematically read about and explore that topic afterwards. However, LLMs face well-known challenges like hallucination and falsehood [7, 101, 102], which could make their outputs uncertain and less trustworthy, often requiring manual inspection and verification before use [36, 41, 74, 101]. In this work, we address these issues with a two-prong approach: 1) reducing hallucination through techniques such as Self-Refine [77]; 2) grounding LLM generations with the content that users would actually read, enabling natural verification.

### 3 FORMATIVE STUDY & DESIGN GOALS

To better understand the obstacles people encounter in their reading and sensemaking in unfamiliar domains, we first conducted a formative study.

#### 3.1 Formative Study

**3.1.1 Methodology.** Participants were a convenience sample of eight information workers (five male, three female) recruited through social media listings and mailing lists. To capture a variety of processes, we recruited three doctoral students, two professional software developers, two researchers, and one administrative staff member. While we do not claim that this sample is representative of all information workers, the interviews were very informative and helped motivate the design of Selenite.

We began by asking participants to recall experiences of conducting sensemaking tasks on topics that they were not familiar with.<sup>2</sup> We then explored how they manage those situations. We asked participants to provide context by reviewing their browser histories to cue their recollections while retrospectively describing those tasks. We solicited their workflows, strategies, frustrations, and needs. Finally, we had participants use the open-source Unakite system [70]<sup>3</sup> to make sense of a topic that they were not familiar with (e.g., for people who have not yet had children to figure out the best baby strollers to purchase for their future child) and externalize their workflows, strategies, and mental models.

**3.1.2 Findings.** Participants reported a total of 28 distinct topics that they encountered and explored, such as “*choosing a hybrid app framework*,” “*selecting the best time tracking tool*,” and “*picking an engagement ring*.<sup>4</sup> Below, we report major findings from the study:

**People often find themselves feeling lost or unsure of where to begin and desire a big-picture understanding of important criteria (or aspects) of an information space before diving deeper.** When approaching an unfamiliar topic, one common strategy that participants reported employing is to find some sort of “*overview of different aspects*” (P3) that would give them “*an intuition of what to care about and some guidance on what to look out for regarding each option*” (P5) in their subsequent exploration. For example, when investigating which time-tracking app to use, P5 was able to find a few articles that provided such overviews at the beginning, e.g., under the section “*What makes the best time tracking software?*” However, participants complained that such lists of criteria are often “*subjective, incomplete*” (P1), can contain aspects that they “*most likely don’t care about*” (P2), and worse yet, “*do not represent how the rest of an article would be structured*” (P6). In addition, for certain topics such as “*best birthday gift ideas*”, such overviews of criteria are hard to find upfront, in which case participants would have to employ a bottom-up approach by reading through a series of articles back-and-forth, which is often considered “*time-consuming*” (P1) and “*hard to actually follow through*” (P2). Without these “*important criteria to keep in mind*”

<sup>2</sup>We subsequently kept track of these topics and used them in our system evaluations.

<sup>3</sup>Unakite is a Chrome extension that helps users collect and organize information into comparison tables in a lightweight fashion while searching and browsing web articles. We used Unakite since it has been shown to be easy to learn and use, and can support a diversity of web page styles and structures [70].

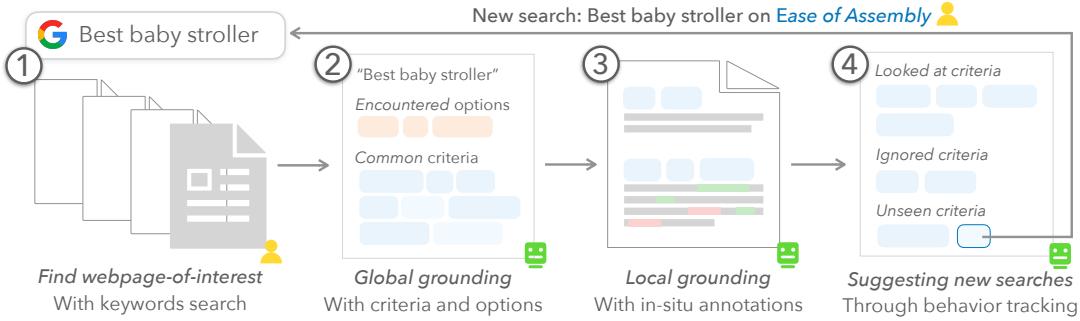
<sup>4</sup>For a complete catalog of these topics, please refer to Table 5 in the Appendix.

(P4) upfront, participants reported feeling “*overwhelmed by large amounts of unfamiliar information*” (P6), lacking “*a sense of clarity and structure*” (P7), and can easily lose focus during sensemaking. These findings prompted us to generate an initial overview of the commonly considered criteria given an information space to provide users with some global grounding and an anchor point for their subsequent reading and sensemaking.

**Identifying and consistently keeping track of criteria is challenging.** Participants cited identifying and aggregating criteria while reading content as a “*significant cognitive load*” (P4). One challenge is that the same criterion can be discussed in various ways across different articles (and even within the same article), making it hard for participants to recognize those variations and time-consuming to flip back and forth to make sure they are eventually aggregated and consistently represented in their Unakite tables. For example, when investigating the topic of “*baby strollers*,” P6 first saw a stroller should be “*agile and nimble* to be able to go through tight spots and sidewalks,” and recorded it as “*nimbleness*” in Unakite; later when she saw another segment that stated that a particular stroller is “*easy to steer and handle* and can smoothly navigate tight corners,” she created another criterion called “*steering*.” It wasn’t until when she saw a segment in a third article that described a stroller having great “*maneuverability and control*” did she realize that all of these were practically describing the same aspect, “*maneuverability*,” and she had to go back and readjust and combine those criteria and their associated evidence in Unakite, a common refactoring challenge for users in unfamiliar domains [56]. Additionally, P7 recounted a similar experience when searching for washers and dryers for his first house and admitted that “*oftentimes, coming up with the right keyword or jargon to summarize what I saw can be surprisingly hard, and I really wish someone would just do that for me.*” Selenite tackles this issue by providing a comprehensive list of frequently considered criteria upfront, reducing the need for individuals to haphazardly find and aggregate criteria themselves.

**People need reading and navigation guidance both at paragraph level as well as article level.** Participants reported often having “*limited attention span*” (P8) when reading online articles and can only focus on a certain amount of information, usually the first few paragraphs or the first few sentences within a paragraph, before getting distracted or lost. For example, P5 in her quest to find a suitable time-tracking app pointed to a typical situation where “*sometimes a paragraph, even a short one, could be quite convoluted and have a lot of intertwined information, for example, and at first I thought this paragraph was just about money, but the rest of the paragraph was actually about lots of other things like platform compatibility*.” But, since participants tend to skim through content quickly, it often leads to potential misunderstandings or missing important details. In such situations, participants desired “*some simple metadata of what’s covered in a paragraph*” (P4) to give them an intuition of what the paragraph is about and whether it is worth reading. These findings prompted us to provide in-situ paragraph summaries and the option for users to clarify convoluted paragraphs by “*zooming in*” on them in Selenite.

The same applies to the page level, where participants wanted to be able to “*preview a page before investing time reading it*” (P3) to understand whether it discusses detailed aspects that they care about. Additionally, such preview can also help them “*maximize the*



**Figure 2: Main stages and features of Selenite:** After the user 1) searches and finds an initial webpage-of-interest to read, Selenite provides: 2) global grounding with a set of common criteria as well as options encountered so far, 3) local grounding with in-situ annotations of criteria per paragraph, and 4) suggestions on what to search for next to gain new information.

information gained from each page” (P5), i.e., help them avoid reading duplicate information and aspects without learning anything new. As P4 put it, “if I’ve already learned about all the aspects from the other pages, I don’t have to read this one.” However, as discussed previously, such previews (even if they are in the form of an abstract or table of contents) are not always available for each article, nor grounded in a person’s past reading and information collection activity. Selenite addresses the page-preview need by offering users a concise overview of what’s covered (and not) in a page to help users gauge its value. Additionally, Selenite tackles the issue of personalization by presenting users with a progress summary based on their previous sensemaking activities at the end of a page.

As participants became more familiar with a topic, their reading patterns started to get increasingly selective and non-linear. For example, we have observed that participants use a combination of keyword searches and flipping back and forth in an effort to find relevant information about a particular criterion that they cared about (with respect to different options), which they thought was “haphazard” (P1) and “inefficient” (P7). This led us to suggest potentially fruitful search keywords to users for discovering more unseen information in the end-of-page progress summary.

### 3.2 Summary of Design Goals

We postulate that an effective user interface/interaction paradigm for helping users find and read about key information during sensemaking should support:

- [D1] As the user starts investigating a topic, **provide a global grounding using common criteria as well as the options encountered** to help users build intuitions of the information space and promote structured thinking;
- [D2] During their reading, **provide a local grounding using page-level as well as paragraph-level summary and annotation** to enable an accurate understanding of and effective navigation within and across articles;
- [D3] Upon finishing, **dynamically suggest next steps in sensemaking based on users’ existing reading and information collection activities** to avoid missing important aspects after reading as well as maximize the information gain in future readings.

## 4 THE SELENITE SYSTEM

Based on the design goals, we designed and implemented the Selenite Chrome extension prototype to help people read about and make sense of unfamiliar topics with the help of global grounding (summarized in Figure 2). We will first illustrate how an end-user, Mirri, would interact with Selenite.

### 4.1 Example Usage Scenario

Mirri, an expectant mother, is seeking guidance in picking a stroller for her upcoming baby. As someone without prior experience in child-rearing, she decided to rely on Selenite to help her while going through review articles and product pages of baby strollers.

Mirri did a quick Google search and clicked on the first result page, which appeared to be a review article titled “The 10 Best Baby Strollers Put To The Test”. Upon opening the page, Selenite automatically recognized the **topic** of the page as “best baby strollers” (Figure 1b), and then automatically presented an overview in a global sidebar available on every page (Figure 1a). The overview contained a list of **criteria** (Figure 1c) that are commonly considered by people when investigating the topic, such as `maneuverability` and `durability`. In addition, Selenite also automatically parsed the web page content and extracted the different baby stroller **options** and presented them under the “Options encountered so far” section (Figure 1d) in the sidebar. After quickly skimming the overview, Mirri now felt that she has already built an intuition about what criteria she should care about when picking baby strollers before even delving into the article itself.

Based on the global options and criteria, Selenite *contextualizes* the ones that are covered on the current page by highlighting them in the sidebar (and conversely low-lighting the ones that are not present, for example, see Figure 1c&d), helping users better understand and find specific information of interest while browsing. In addition, as Mirri read the article, she noticed that Selenite provided **in-situ annotations** of **mentioned criteria** at the beginning of each paragraph (Figure 1e). She quickly learned that she could just skim those mentioned criteria to get a rough idea of what a particular paragraph is about and decide if that paragraph is worth reading.

When she came across information about the `maneuverability` of a specific stroller, Mirri became interested in finding out if there were any details about the maneuverability of other stroller options

best umbrella strollers we tested...' (with 'Price' highlighted in red). Each section has a 'Price' button, a 'Maneuverability' button, a 'Wheel type' button, and a 'Durability' button."/&gt;

**Figure 3: Selenite enables structured and efficient navigation by criterion through clicking the “previous/next” (shown as “<” and “>”) buttons (a), after which Selenite will automatically scroll the page to reveal the previous/next mentioning of the target criterion.**

as well. To facilitate this, she used the “previous/next” buttons (Figure 3a) to quickly navigate among the paragraphs that discussed maneuverability. Here, the aforementioned annotations not only offer paragraph overviews during *linear* skimming but also act as bookmarks for *non-linear* navigation between distinct parts of the page that pertain to similar criteria.

Later, when Mirri encountered a particularly convoluted paragraph with multiple criteria and options that she couldn’t quite absorb after a first pass, she decided to leverage the “zoom in” feature that Selenite offers — querying for more comprehensive descriptions that clarify which sentences or phrases within the paragraph pertain to specific criteria and sentiments (positive, neutral, or negative) (Figure 4) by clicking the “Analyze” button (Figure 4a) that appears when hovering over a paragraph.

After Mirri reached the end of the current article, Selenite presented a **summary** block (Figure 1f), automatically summarizing her research progress (e.g., criteria from the overview that she has actually read about). With the help of this summary, Mirri realized that she hasn’t seen evidence related to *ease of assembly* or *brake & locking system* before. She then specifically searches for them on Google, finding new articles that contain information about these previously unencountered criteria.

## 4.2 Detailed Designs

We now discuss how the various Selenite features are designed and implemented to support the design goals.

**4.2.1 [D1] Providing Global Grounding using Common Criteria and Options Encountered.** In Selenite, we explore the idea of having the system provide users with an initial overview of criteria that are typically significant and frequently considered by people when exploring a particular topic. Selenite also performs information extraction on each page to identify the options that a user has encountered during their sensemaking process. Naturally, users have the flexibility to reorder, pin, edit, add, or delete any

best full-size strollers and jogging competitors, many users feel its higher quality and easy-to-use features more than justify the higher cost.' (with 'Price' highlighted in red), and 'its higher quality features more than justify the higher cost.' (with 'Durability' highlighted in green). Each section has a 'Price' button, a 'Maneuverability' button, a 'Wheel type' button, and a 'Durability' button."/&gt;

**Figure 4: When encountering a particularly convoluted paragraph (e.g., the paragraph on the left) with multiple criteria and options that users can’t quite absorb in the first pass, they can click the “Analyze” button (a) and leverage the “zoom in” feature that Selenite offers to query for more comprehensive descriptions that clarify which sentences or phrases pertain to which specific criteria and sentiments. Selenite wraps phrases and sentences in colored boxes, with green denoting “positive” (b), red denoting “negative”, and grey denoting “neutral” (not shown).**

options and criteria to tailor them precisely to their specific preferences. We discuss the relevant designs and the rationale behind those designs below:

**Automatically recognizing topics.** Selenite goes beyond previous sensemaking systems [17, 56, 58, 69, 71–73] by autonomously identifying and classifying web pages into broad topics based on their titles and content. Unlike previous systems that require users to manually determine the topic [15, 43, 70], automatic topic recognition further lowers the barrier for entry, enabling users to quickly begin reaping benefits from the list of commonly considered criteria based on that topic. To achieve this, we frame the topic recognition as a *summarization* task for an LLM — specifically, we asked GPT-4 (taking advantage of its generalizability to various domains [84])<sup>5</sup> to first *summarize* an arbitrary web page given its title and initial five paragraphs (with the temperature set to 0 to minimize LLM hallucination).<sup>6</sup> Selenite then clusters the semantically similar topics (note that each web page has an associated topic generated by GPT-4) based on the cosine distances on topic semantic embeddings computed using SentenceBERT [94]. Therefore, for example, websites titled “React vs. Svelte: Performance, DX, and more,” “Angular vs React vs Vue: Which Framework to Choose,” and “What are the key differences between Meteor, Ember.js and Backbone.js?” would all be regarded as “Comparison of JavaScript frameworks.” Naturally, users have the flexibility to manually create, edit, and remove topics, as well as reassign pages to different topics based on their personal opinions.

**Automatically retrieving commonly considered criteria.** If we adopt the “bottom-up” approach discussed in prior work [43, 64, 70], an intuitive method for obtaining criteria would involve extracting them from individual paragraphs on a page. However, in our initial

<sup>5</sup>Please refer to section D.1 in the Appendix for the detailed prompt design.

<sup>6</sup>Empirically, we found this step helped GPT-4 to better engage with the context provided. It also aligns with the idea of Chain-of-thought prompting proposed by [107], and then offer a *search phrase* that would help one to find similar web pages using a modern search engine, which we use as the topic.

attempts, we found that this method faced significant challenges that limited its effectiveness. One of the main issues was the lack of uniformity among the criteria extracted from different paragraphs – each paragraph presented its own variations and nuances, making it difficult to establish a cohesive and standardized set of criteria (similar to what was reported by our formative study participants). Additionally, the approach lacked a comprehensive global perspective, failing to consider the broader context and overarching themes of the topic. As a result, manual review, correction, and unification of the extraction results were frequently necessary, making the process impractical and inefficient.

To address these challenges, we instead explored an alternative “top-down” approach, where we directly query an “oracle” for a globally applicable and comprehensive set of criteria. We were particularly inspired by recent research suggesting that *the majority of people’s information-seeking needs are not novel* [75] – “previous people” have experimented with most search needs and synthesized information into summarized knowledge such as review articles. While it is impractical for individuals to process and synthesize vast amounts of information online, LLMs excel at this. Recent studies suggest that LLMs can be highly effective in processing and integrating information, making them potentially valuable for tasks like knowledge graph querying and retrieving common sense information [1, 106, 113], and, in our particular case, a suitable “oracle” for providing a set of commonly considered criteria given a particular topic.

In Selenite, we use GPT-4 as a *knowledge retriever* – for any given topic, we prompt it to produce a list of around 20 commonly considered criteria (Figure 1c), complete with their respective names (Figure 1c1) and descriptions (Figure 1c2) for each topic. To minimize potential anchoring biases, we strive to achieve a balance between relevance and diversity in our prompting strategy. First, we specifically requested an initial set of criteria that are deemed as “most relevant to the topic,” “frequently considered,” and can “cover a broad range of perspectives.” Then, we iteratively prompted GPT-4 while applying the Self-Refine technique [77], where in each iteration, we requested the generation of five additional criteria that were “different, more diverse, and more important” than the previous ones.<sup>7</sup> We also relied on GPT-4 for ranking the criteria based on their importance. Still, users have the freedom to request additional criteria (without repetition) if they believe the existing list is not comprehensive enough (Figure 1c3), or manually add criteria (Figure 1c4). We present in Table 1 the list of criteria that Selenite retrieves for the topic of “best baby strollers,” offering an intuition of their quality and coverage. We further validate our approach through a performance evaluation in section 5 that provides initial evidence that our approach is sufficient for our prototyping purposes. We leave for future work to experiment with advanced approaches, such as retrieval-augmented generation (RAG) [67], that would potentially provide increased perceived external validity.

*Automatically recognizing encountered options.* Instead of relying on GPT-4 to access its internal knowledge and retrieve a set of commonly considered options, we instead leverage its zero-shot *information extraction* capability and expansive context window size [84] to directly extract options from the entire text content of a web page. This approach ensures that the options presented

in the sidebar align with a user’s sensemaking process, i.e., they are indeed what users have encountered as opposed to something that users would potentially never run into. It also circumvents the potential concern where the world knowledge of an LLM is out-of-date, for example, at the time of writing, GPT-4 only “knows” information up to September 2021 [84].<sup>8</sup> In addition, it surpasses the limited heuristics employed in previous approaches such as Crystalline [72], which rely on page titles and HTML <h>-tags as sources for options. This is crucial, because studies have consistently demonstrated that web pages frequently disregard semantic web standards and best practices [49, 78].<sup>9</sup>

**4.2.2 [D2] Providing Local Grounding using Page & Paragraph-level Summary and Annotation.** To address the challenges identified in the formative study, where complex paragraph and page structures frequently led to overlooked information and hindered user comprehension, Selenite introduces the following features:

*In-situ summaries and annotations of paragraphs.* With access to the initial set of common criteria as well as the options extracted from each page, Selenite performs content analysis on each paragraph within a given page to identify the specific criteria being discussed and presents them as in-situ annotations above the respective paragraph (Figure 1e). This feature enables users to swiftly scan through a page, understand the key points of each paragraph, and selectively concentrate on the paragraphs that are valuable and engaging for gathering information.

Such content analysis is enabled by recent advances in large pre-trained transformer models [29, 66, 105] fine-tuned to perform zero-shot text classification tasks following a natural language inference (NLI) paradigm [115]. Specifically, for example, to assess if a given text (e.g., “Angular is very hard to pick up”) covers the criterion of *learning curve*, we can input the text as the *premise* and a *hypothesis* of “This content discusses {learning curve}.” into the NLI model. The entailment and contradiction probabilities are then converted into label probabilities, indicating the likelihood that the content pertains to the specified criterion. We used the bart-large-mnli model<sup>10</sup> for this purpose and considered options and criteria with a score above 0.96 as true positives, displaying them in descending order of scores. We determined this threshold empirically, prioritizing recall over precision, as discussed further in section 5.

In scenarios where users still struggle to comprehend content despite the presence of in-situ annotations, Selenite can perform a deeper analysis on-demand by leveraging the advanced reasoning capabilities of GPT-4. Specifically, through parallel and carefully orchestrated prompts, Selenite produces a more comprehensive description that clarifies which sentences or phrases pertain to specific criteria and sentiments (positive, neutral, or negative) (Figure 4). Although a formal evaluation of this method is beyond the scope of this work, recent research suggests that content analysis conducted by the latest generation of LLMs achieves state-of-the-art

<sup>8</sup>While the direct retrieval of criteria from LLMs may also face this potential issue, in practice, we operate under the assumption that criteria are unlikely to suddenly emerge or become outdated.

<sup>9</sup>For instance, it is common to find pages where nearly every piece of content is enclosed in <div> tags regardless of their semantic roles.

<sup>10</sup>The model can be accessed on-demand through a remote API service that we implemented.

<sup>7</sup>Please refer to section D.3 in the Appendix for the detailed prompt design.

Criterion Name	Criterion Description
Safety	Ensuring the stroller has proper safety features such as a secure harness, sturdy construction, and reliable brakes.
Comfort	Providing a comfortable seat with adequate padding and support for the baby, as well as adjustable recline positions.
Maneuverability	Having easy & smooth maneuverability, with features like swivel wheels, suspension systems, and the ability to navigate tight spaces.
Durability	Ensuring the stroller is built to last, with high-quality materials and strong construction.
Storage	Offering ample storage space for carrying essentials such as diaper bags, snacks, and personal items.
Folding and Portability	Allowing for easy folding and compact storage, as well as being lightweight for convenient transportation.
Versatility	Providing features that allow the stroller to adapt to different terrains, weather conditions, and age ranges.
Ease of Use	Having user-friendly features like adjustable handles, intuitive controls, and easy-to-clean fabrics.
Price	Considering the affordability and value for money in relation to the features and quality of the stroller.
Customer Reviews	Taking into account feedback and recommendations from other parents who have used the stroller.
Weight and size	Considering the weight and size of the stroller to ensure it is manageable and fits well in different environments.
Ease of cleaning	Ensuring the stroller is easy to clean and maintain, with removable and washable fabric components.
Adjustability	The stroller should have adjustable handlebars and footrests to accommodate different caregivers and growing babies.
Canopy	A large and adjustable canopy to protect the baby from the sun and other elements.
Reversible seat	Having the option to face the baby towards the parent or away from the parent.
Brake system	Having a reliable brake system that is easy to engage and disengage.
Car seats compatibility	Offering the ability to attach a car seat to the stroller for convenient travel.
Adjustable height	Allowing for adjustable handlebars to accommodate different heights of caregivers.
Easy assembly	Providing clear instructions and easy assembly process for the stroller.
Design and aesthetics	Considering the overall design and aesthetics of the stroller to match personal preferences.
Weight capacity	Specifying the maximum weight limit the stroller can safely carry.
Warranty	Checking for a warranty or guarantee that covers any potential defects or issues with the stroller.
Brand reputation	Considering the reputation and reliability of the brand manufacturing the stroller.
Accessories	Offering additional accessories such as rain covers, mosquito nets, or parent organizers for added convenience.

**Table 1: Commonly considered criteria (name & description) that Selenite retrieves for the topic of “best baby strollers.”**

performance in terms of quality, accuracy, and granularity [11, 84], making it suitable for our purposes.

*Page-level overview of options and criteria.* As evidenced by our formative study, providing a page-level overview of the information space to users can greatly assist them in reading and sensemaking tasks. To facilitate this, Selenite consolidates paragraph-level metadata into the sidebar’s options and criteria entries, with the entries that are present on the page highlighted (Figure 1c&d). This offers two key benefits. First, it provides a comprehensive summary of all the available options and criteria specific to the current page, which allows users to quickly understand the focus of the page as well as judge its value against their personal interests and needs. Second, it enables structured and efficient navigation. By utilizing the “previous/next” button for a given criterion (Figure 3a), users can swiftly navigate between distinct parts of the page related to identified criteria (Figure 3). This feature saves users time and effort, as it eliminates the need for manual searching and filtering, which our formative study found to be the common practice.

It is worth noting that the combination of page and paragraph-level annotations effectively addresses a significant prior limitation highlighted by Crystalline [72] and further revealed in our formative study: the inability to manually recognize “latent/implicit criteria,” where the same criterion can be expressed in various forms without being explicitly mentioned (for instance, it can identify the criterion of “price” from a statement like “I bought this mp3 player for almost nothing” [91]).

**4.2.3 [D3] Dynamically Suggesting Next Steps in Sensemaking.** As users navigate through a page and consume the content, Selenite implicitly keeps track of the criteria and aspects that they paid attention to on the page based on dwell time, that is, the amount of time they roughly spent reading a block of content, as specified in [21, 72]. Selenite then uses this information to summarize the user’s research progress in the aforementioned summary block (Figure 1f), which consists of three sections: (1) criteria that users **cared about** and have seen evidence for based on implicit tracking (emphasizing their focus and priorities), (2) the remaining ones that are discussed on the page but users **ignored or skipped**<sup>11</sup> (reminding users of potentially overlooked criteria), and (3) a set of **recommended** criteria from the global overview that haven’t been discussed in any of the past articles but could still be worth exploring (encouraging users to find additional information about these unseen criteria, for example, by conducting additional searches on them, thereby broadening their perspectives and maximizing information gain).

Specifically, to achieve this third objective, we leverage users’ the subset of criteria that users *cared about* and the subset they have intentionally *skipped* to *recommend* additional **relevant** and **diverse** criteria to search for and read about from the remaining global list. This requirement for the suggested criteria to be both

<sup>11</sup>The threshold of determining if the user indeed paid attention to a paragraph is set to 2 seconds based on our empirical testing. Future work can investigate more adaptive methods, such as taking into account the length of a paragraph, the amount of new information contained in a paragraph compared to users’ existing knowledge, or if users appear to be idling and performing irrelevant activities.

relevant and diverse is similar to the exploration-exploitation trade-off in information retrieval and recommender system literature [4], helping maintain user engagement and interest while avoiding over-fit or filter bubbles [63, 100].

To operationalize this idea, we consider the problem a graph problem (following the approach by [112]): by constructing a fully connected graph using the global list of criteria as *vertices*, we assign *edge* weights as distances between respective criteria in a semantic embedding space and vertex weights as the criterion's relevance to the subset of criteria that the user cared about. Our objective then is to recommend a *diverse* subset of criteria (vertices) that have large distances between each other while still being *relevant* to what users cared about. That is, we need to find a sub-graph  $G'$  of size  $n$ , which maximizes a weighted ( $\beta > 0$ ) sum of vertex weights  $w_V$  (relevance) and edge weights  $w_E$  (diversity):

$$\arg \max_{G' \subset G, |G'|=n} \beta \cdot w_V(G') + w_E(G')$$

To build the graph, we measure relevance with the perplexity score of the sentence “{global\_criterion} tend to be considered together (or is a trade-off) with {cared\_about\_criterion}” using GPT-2 [93] and characterize diversity with the cosine distance between the SentenceBERT [94] embeddings of the two vertices (criteria). Here, we follow the classic greedy peeling algorithm [111] by dropping vertices with the lowest weights (the sum of vertex and every edge weight) one at a time in a greedy fashion until the graph size reaches  $n$  (empirically determined as 2). We then present these additional criteria as the ones that users “might be interested in further searching for” (Figure 1f) in the summary block.

### 4.3 Implementation Notes

The Selenite browser extension is implemented in HTML, TypeScript, and CSS and uses the React JavaScript library [34] for building UI components. It uses Google Firebase for backend functions, database, and user authentication.

As explained previously, we implemented Selenite using state-of-the-art NLP models: off-the-shelf GPT-4 [84] and NLI models finetuned on BART [66]. These models were chosen for their strong performance and efficiency that would satisfy our prototyping needs as well as their generalizability across different application domains (c.f. Section 5). However, it is important to note that **our contributions lie more in the concept of grounded reading, interface design, and underlying NLP task abstractions, which are independent of specific model usage**. We anticipate that these designs will remain valid as AI techniques continue to advance [74].<sup>12</sup>

## 5 STUDY 1: INTRINSIC EVALUATION

While Selenite can help ground users in what to read, its impact may backfire if the list of options and criteria is not accurate or comprehensive — anchoring bias [104]<sup>13</sup> may cause readers to more easily miss information that is *indeed included in the page but not reflected*

<sup>12</sup>Additional implementation details can be found in section E of the Appendix.

<sup>13</sup>Anchoring bias refers to people’s inclination towards relying too excessively on the initial set of information they were exposed to on a topic. Regardless of the accuracy or quality of that information, people use it as a reference point, or an “anchor,” to make subsequent judgments or decisions [118].

in the Selenite-generated options and criteria list. Here we evaluate on a diverse set of topics whether Selenite can: 1) **accurately** report options that are present on a web page; 2) **comprehensively** report critical criteria people commonly consider.

### 5.1 Methodology

**5.1.1 Topic Sampling.** We collected ten topics that exhibit a mixture of practicality and diversity (Table 2): (1) we randomly sampled five topics (out of the 28 topics reported) reported by participants in the formative study, and (2) we collected five more from Wirecutter, a popular review site — the three most popular product guides listed in their 2021 year-in-review (at the time of writing, the 2022 year-in-review has not been published) as well as their two most recently updated guides for June 2023.

**5.1.2 Groundtruth Dataset Creation for Options and Criteria.** To collect groundtruth criteria that the general audience would care about for each topic, we mimic a typical information collection workflow, where people rely on top sources from popular search engines for their authenticity and credibility. Specifically, we first gathered the top five Google search results using the query template “best [product or category]” (excluding promotions or ads). We show a partial snapshot of a representative web page in Figure 5 in section B.1 of the Appendix. Then, for each web page, two authors independently first read through and annotated the options and criteria mentioned in every paragraph, and then merged all the annotations, excluding duplicate ones. Note that since many criteria are mentioned in a descriptive manner (e.g., the phrase “It is available in a black finish” implicitly refers to aesthetics), the two authors had some variance in how they named essentially the same criteria. Therefore, the two authors iteratively discussed and resolved their conflicts, merging criteria that they believed were semantically equivalent, producing the groundtruth criteria list (see Table 6 in section B.2 of the Appendix for the groundtruth criteria list). Finally, for the topics that we sampled from the formative studies where participants explicitly collected options and criteria using Unakite, we double-checked and were able to verify that all the criteria that they identified were indeed included in our groundtruth dataset, providing preliminary evidence to the soundness of our groundtruth dataset.

#### 5.1.3 Evaluation Metrics.

**Option Extraction.** Since Selenite directly extracts options from web pages, we evaluated this capability using the *accuracy*, that is, the percentage of options extracted by Selenite out of all the options available on a page.

**Criteria Retrieval.** We also evaluated Selenite’s ability to retrieve the right set of criteria on two levels. First, to answer whether Selenite helps find useful criteria *for each topic*, we compute *topic-level precision* (“the fraction of criteria retrieved by Selenite that coincided with the groundtruth”) and *recall* (“the fraction of groundtruth criteria that are were also retrieved by Selenite”).

Second, to measure whether Selenite provides high-quality groundings *per paragraph*, we additionally randomly sampled 20 paragraphs per topic, and computed *paragraph-level precision* (“the fraction of criteria recognized by Selenite that were indeed mentioned

Topic	Topic-level						Paragraph-level				
	#GT (total)	#Selenite (total)	Precision	Recall	F1	#GT (avg)	#Selenite (avg)	Precision	Recall	F1	
Best washing machines	19	24	0.88	1.0	0.93	3.05	3.30	0.91	1.0	0.95	
Birthday gift ideas	11	21	0.57	0.91	0.70	1.40	3.45	0.57	0.96	0.72	
Best hybrid app frameworks	15	21	0.86	0.93	0.89	2.55	3.00	0.83	1.0	0.91	
Best time tracking tools	20	21	0.81	0.95	0.87	2.95	3.20	0.88	0.98	0.93	
Deep learning frameworks	25	20	0.80	0.84	0.82	3.15	3.05	0.87	0.95	0.91	
Best sleeping bags	19	21	0.81	0.89	0.85	2.85	3.15	0.95	1.0	0.97	
Best air purifiers	20	24	0.83	1.0	0.91	3.05	3.75	0.83	0.98	0.90	
Best robot vacuums	23	28	0.82	1.0	0.90	3.10	4.05	0.95	1.0	0.97	
Best baby strollers	22	24	0.92	1.0	0.96	3.45	3.65	0.81	1.0	0.90	
Best tropical vacation spots	15	19	0.74	0.93	0.82	2.55	3.10	0.92	1.0	0.96	
Mean	19.0	22.3	0.80	0.95	0.87	2.81	3.37	0.85	0.98	0.91	

**Table 2: Statistics of the accuracy and coverage evaluation on Selenite’s capability to retrieve a high-quality set of commonly considered criteria by topic. For topic-level, we report the number of groundtruth criteria, i.e., “#GT (total)”, the number of criteria retrieved by Selenite, i.e., “#Selenite (total)”, as well as the precision, recall and F1-score. For paragraph-level (recall that we randomly sampled 20 paragraphs per topic), we report the average number of groundtruth criteria mentioned per paragraph, i.e., “#GT (avg)”, the average number of criteria reported by Selenite, i.e., “#Selenite (avg)”, as well as the precision, recall and F1-score.**

in the paragraph”) and *recall* (“the fraction of criteria mentioned in the paragraph that were indeed reported by Selenite”).

## 5.2 Results

**5.2.1 Option Extraction.** Selenite achieved 100% accuracy on extracting options from web pages, i.e., as long as there was an option explicitly mentioned on a web page, Selenite was able to correctly extract it. This directly speaks to the strong reasoning and information extraction capabilities of GPT-4 as described in OpenAI’s technical report [84].

**5.2.2 Criteria Retrieval.** We present the result of criteria retrieval evaluation metrics in Table 2, which provides initial evidence to Selenite’s strong capability in presenting to the user a comprehensive set of criteria that people commonly consider. Notice that for most of the topics, Selenite retrieved *more* criteria compared to the groundtruth set. This is not surprising, partly due to the fact that GPT-4 has likely synthesized information from significantly more sources than what was considered during the construction of the groundtruth dataset (five web pages for each topic). Theoretically, there is also a possibility that GPT-4 hallucinated some criteria that are largely irrelevant to a given topic, however, upon further manual inspection, we did not see evidence of hallucination, at least for the 10 topics considered in this evaluation (for example, Table 1 shows a list of commonly considered criteria that Selenite retrieves for the topic of “best baby strollers”).

**Topic-level recommendations.** Selenite achieved both high recall and high precision on multiple topics (e.g., *best washing machines*, *best air purifiers*, *best robot vacuums*, and *best baby strollers*), and usually achieves higher recall than precision, suggesting that Selenite has the tendency of finding *supersets* of what users would generally be able to identify from reading, i.e., criteria in the groundtruth set.

We qualitatively analyzed the topics with a lower-than-average topic-level criteria recall, and found two contributing reasons: (1) Some web pages cover factual information that is not necessarily relevant. Multiple pages describing *Best Hybrid App Frameworks*

mentioned “First Release Date,” which arguably is not a *criterion* necessary for selection. (2) Some criteria are inter-correlated. For example, in the case of *deep learning framework*, whereas it did not explicitly mention “growth speed,” Selenite did suggest *innovation*, whose description is “the ability of the framework to stay up-to-date with the latest research and developments in deep learning, and to incorporate new techniques and architectures as they emerge.” While we did not count these two as equivalent in the evaluation, in practice, these two have a high correlation, and we believe having one included might be sufficient. Still, this potential mismatch reflects the necessity of allowing users to edit the criteria and descriptions. Meanwhile, upon initial observation, Selenite’s lower precision on certain topics may suggest its inclination towards retrieving unnecessary criteria. However, a closer examination revealed an interesting insight: for instance, when it comes to topics like *birthday gift ideas*, popular web pages often present a list of 10+ diverse options that lack strict comparability and are all described using generic terms such as “fun” or “sweet.” This lack of specificity makes it challenging to determine a comprehensive set of groundtruth criteria. In contrast, Selenite offers comprehensive overviews that encompass factors like *personalization*, *uniqueness*, *practicality*, *sentimentality*, and *presentation (wrapping)*, among others.

**Paragraph-level Grounding.** Selenite also achieved high per-paragraph performances, again with a bias towards higher recalls. This is intentional – we tuned the parameters of the NLI-based method such that it is more likely for Selenite to claim non-existing criteria than overlooking actual existing ones. This approach prioritizes avoiding information loss, which, suggested by prior work [74], is a more expensive mistake compared to user verification. We order the criteria based on their probability score from the NLI model and will, in future iterations, fade the ones with a lower score.

We did notice that in some rare scenarios, the NLI performance can be influenced by a criterion’s description, e.g., changing “appropriate for age” to “appropriate for kids, adults, or elderly” can reduce

Selenite's error on recognizing arbitrary numbers as ages. Therefore, in future iterations of Selenite, we will provide a hint to users, prompting them to try tweaking the description when they attempt to delete a criterion due to its seemingly low grounding efficacy.

## 6 STUDY 2: USABILITY EVALUATION

We also conducted an initial usability study to evaluate if the features provided by Selenite are usable and if the approach of providing global as well as contextual grounding can allow users to read, navigate, and comprehend information more efficiently. Specifically, we were interested in the following quantitative research questions:

- [RQ1] Does using Selenite speed up people's process of reading and understanding information?
- [RQ2] Does using Selenite help people achieve a more comprehensive understanding of an information space?
- [RQ3] Can Selenite help people obtain new information in addition to their existing knowledge?

### 6.1 Methodology

We recruited 12 participants (five female, seven male) aged 21-40 ( $\mu = 28.9$ ,  $\sigma = 5.2$ ) through social media. Participants were required to be 18 or older and fluent in English. All participants reported that they regularly engage in the process of seeking and sifting through large volumes of online information, whether for professional or personal purposes, on a weekly basis.

The study was a within-subjects design, where participants were presented with two tasks and were asked to complete each one under a different condition, counterbalanced for order. For each task, participants were given a topic that they needed to investigate and two web pages relevant to the topic that they were required to read and process. The two topics were "*best baby strollers*" and "*best robot vacuums*".<sup>14</sup> The provided two web pages for each topic were all product comparison pages used in the previous study (see section 5.1). For each task, participants were asked to read through the two required pages, either by themselves without any aid (a *control* condition simulating how people normally read) or with Selenite (*experimental* condition). While reading, they were instructed to write down *as many* criteria as they learned and thought were important for the topic as well as the reason why they were important as if they needed to thoroughly explain the topic to a friend later. Then, participants were instructed to optionally search (using Google) and gather additional information that they still wanted to learn about but weren't able to from reading the two required pages. We imposed a 25-minute limit per task to keep participants from getting caught up in one of the tasks. However, they were instructed to inform the researcher that they felt like they could make no further progress, i.e., having learned as much as they could about the given topic.

<sup>14</sup>To ensure realism and participant engagement, the tasks were selected based on actual topics that the formative study participants reported investigating. Rather than letting participants search for their own pages to read from the get-go, we provided them with a predefined set of pages to enable a fair comparison of the results (e.g., speed, etc.). Requiring participants to use predefined pages (each contains, on average, 15 screenfuls of content) for the first portion of the study also helps ensure that the two tasks are of roughly equal difficulty in terms of reading and cognitive processing effort. As described in the results, there was no significant difference by task.

Each study session started by obtaining consent and having participants fill out a demographic survey. Participants were then given a 5-minute tutorial showcasing the various features of Selenite and a 5-minute practice session before starting. At the end of the study, the researcher conducted a NASA TLX survey and a questionnaire, eliciting feedback on their experience in both conditions. Each study session took around one hour, using a designated Macbook Pro computer with the latest version of Chrome and Selenite installed, and was conducted remotely via Zoom. Each participant was compensated with \$15 USD. The study was approved by our institution's IRB.

### 6.2 Results

All participants were able to complete all tasks in both conditions, and nobody went over the pre-imposed time limit. Below, we present primarily quantitative evidence to evaluate the usability of Selenite with respect to our research questions.

First, we were interested in understanding if Selenite can help participants read and process information faster compared to the baseline condition (RQ1). To examine this, we measured the time it took for them to finish reading all the materials in each task. A two-way repeated measures ANOVA was conducted to examine the within-subject effects of the condition (baseline vs. Selenite) and task on completion time. There was a statistically significant effect of condition ( $F(1, 20) = 102.5$ ,  $p < 0.01$ ) such that participants completed tasks significantly faster (36.3%) with Selenite (Mean = 840.3 seconds,  $SD = 102.7$  seconds) than in the baseline condition (Mean = 1319.3 seconds,  $SD = 120.0$  seconds). There was no significant effect of task ( $F(1, 20) = 0.40$ ,  $p = 0.53$ ), indicating the two tasks were indeed of roughly equal difficulty. These results suggest that Selenite helped participants read and comprehend information more efficiently. We discuss additional qualitative insights into why Selenite was more efficient in the following open-ended case study (section 7.2).

In addition, we were interested in understanding if Selenite can help participants achieve a more comprehensive understanding of a topic (RQ2). To measure this, we first compared the *quantity* of criteria that participants externalized under each condition. As a pre-filtering step, two researchers rated all the criteria that participants externalized as either *valid* or *invalid* blind to the conditions. Valid criteria are considered as ones that are *relevant* to the topic and *backed by specific evidence* that can be traced back to the content, consistent with those standards used by prior work in judging the quality of subjective evidence [15]. After resolving conflicts (which were minimal) between the two researchers and filtering out the criteria that were invalid, we found that the average total number of valid criteria increased by 90.4% when using Selenite (Mean = 12.93,  $SD = 3.90$ ) compared to the baseline condition (Mean = 6.79,  $SD = 4.07$ ), which is statistically significant ( $p < 0.01$ ) under a t-test. Thus, using Selenite appeared to enable participants to identify and learn significantly more criteria about a topic compared to people's current way of reading information.

In addition to quantity, we also examined the *quality* of the criteria by comparing the ones that participants externalized with the groundtruth criteria curated in the previous accuracy and coverage evaluation — we can calculate the precision (calculated as

	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration
Selenite	3.0 (3.03 ± 1.76)*	1.0 (0.51 ± 1.74)	2.5 (2.26 ± 1.68)*	8.5 (8.47 ± 1.32)*	3.5 (4.08 ± 1.88)*	0.5 (0.33 ± 1.51)
Baseline	6.5 (6.43 ± 2.07)*	1.0 (0.79 ± 1.98)	4.0 (4.29 ± 2.08)*	6.5 (6.54 ± 1.73)*	6.0 (5.98 ± 2.23)*	1.0 (0.89 ± 1.91)

**Table 3: Study 2 participants' responses to NASA TLX questions (on a scale from 0 to 10) in study 2. Format: median (mean ± standard deviation). Statistically significant differences ( $p < 0.05$ ) through t-tests are marked with an \*.**

Question category	Statement	Response
Comprehensibility	I would consider my interactions with the tool to be understandable and clear.	6 (6.33 ± 1.10)
Learnability	I would consider it easy for me to learn how to use this tool.	7 (6.71 ± 1.04)
Enjoyability	I enjoyed the features provided by the tool.	6 (6.13 ± 1.72)
Applicability	Using this tool would make solving sensemaking problems more efficient and effective.	6 (6.28 ± 1.39)
Recommendability	If possible, I would recommend the tool to my friends and colleagues.	6 (6.23 ± 0.94)

**Table 4: Study 2 participants' responses to System Usability Scale questions (on a scale of 1 to 7, where 1 represents “strongly disagree” and 7 represents “strongly agree”) in study 2 regarding their Selenite experience. Format: median (mean ± standard deviation)**

$n_{\text{Hit}}/n_{\text{Total}}$ ) and recall (calculated as  $n_{\text{Hit}}/n_{\text{Groundtruth}}$ ) of participants' criteria that *hit* the groundtruth (where  $n_{\text{Total}}$  is the total number of valid criteria participants externalized, and  $n_{\text{Groundtruth}}$  is the number of groundtruth criteria for each task). On average, participants in the Selenite condition achieved significantly higher precision (98.8% vs. 78.4%,  $p < 0.05$ ) and recall (73.0% vs. 30.4%,  $p < 0.05$ ) in both tasks. Thus, using Selenite appeared to have enabled participants to improve the quality of their understanding of an information space in terms of its criteria.

Furthermore, to understand if Selenite can help participants obtain new information in addition to what they have already learned from reading the two required pages (RQ3), we examined: 1) the number of additional *searches* that they performed in the Selenite condition (Mean = 2.01, SD = 1.39), which turned out to be significantly more ( $p < 0.05$ ) than the baseline condition (Mean = 0.33, SD = 0.62); 2) the number of additional *pages visited* in the Selenite condition (Mean = 2.76, SD = 2.32), which turned out to be significantly more ( $p < 0.05$ ) than the baseline condition (Mean = 0.42, SD = 0.74); and 3) the number of additional *criteria* that participants externalized in the Selenite condition (Mean = 1.58, SD = 0.91), which turned out to be significantly more ( $p < 0.05$ ) than the baseline condition (Mean = 0.33, SD = 0.22). These results suggest that Selenite did encourage and help participants to seek additional information beyond their existing perspective.

Last but not least, participants filled out a NASA TLX [45] cognitive load scale and a System Usability Scale (SUS) [65] questionnaire for each condition. SUS Likert items were integer-coded on a scale from 1 (strongly disagree) to 7 (strongly agree). The median response values are presented in Tables 3 and 4. Notably, participants perceived Selenite to have significantly lowered workload across mental, temporal, and effort demands as well as significantly increased perceived performance based on paired t-tests. This suggests that using Selenite can reduce the cognitive load and interaction costs when reading and understanding information, even when users had to learn and get used to a new user interface.

## 7 STUDY 3: OPEN-ENDED CASE STUDY

Encouraged by the promising performance outcome of the previous two studies, we conducted a third open-ended case study to understand the usefulness and effectiveness of the Selenite prototype from a qualitative perspective.

### 7.1 Methodology

We recruited eight participants (three male, five female; three students, two software engineers, one dermatologist, one accountant, and one researcher) aged 24–55 years old (Mean = 33.6, SD = 8.1) through emails and social media. The same recruitment requirements were applied, but individuals who participated in the previous usability study were excluded from this study.

Each participant first completed two pre-defined tasks, where they used Selenite to help them read information about an unfamiliar topic. From the topics that participants reported having explored in the formative study, we randomly selected two that the participant was unfamiliar with (indicated in their screening survey). For each task, participants were presented with a set of three web pages that covered the topic that the formative study participants had gone through. The provided web pages were primarily review articles comparing several options together or product detail pages. We imposed a 20-minute limit per task to keep participants from getting caught up in one of the tasks. To further explore Selenite's potential, all participants then used Selenite to make sense of a third topic that they intend to explore in real-life. Here, we purposefully did not limit the topic to be “unfamiliar,” allowing participants to revisit previous topics of interest and potentially uncover fresh perspectives. Each study session began by obtaining consent and demographic information. Participants were then given a 5-minute tutorial showcasing the various features of Selenite and a 5-minute practice session before starting. At the end of the study, the researcher elicited feedback on using Selenite through a semi-structured interview, which was recorded and later transcribed for coding and thematic analysis [19]. Each study was conducted via Zoom for up to one hour. Each participant was rewarded with \$15 USD. The study was approved by our institution's IRB.

## 7.2 Results

Below, we present the major qualitative findings from the observation of participants' behaviors using Selenite as well as their feedback from the post-study interviews.<sup>15</sup>

*Time and effort savings.* All of the participants mentioned that using Selenite would save them a lot of time and effort compared to using their typical reading and information collection workflow, echoing the quantitative results reported in the usability evaluation (see section 6). First of all, having access to the global overview felt like “*a game-changer*” (P8) that offers a “*bird’s-eye view*” (P4) or access to “*on-demand expert opinion*” (P1) that “*took away the anxiety and guesswork of wondering what other folks would actually care about*” (P5). P7 suggested that “*this is something that I always wished for when reading about stuff that I’m not an expert in. It seriously saves me a ton of time that I’d otherwise spend trying to wrap my head around it little by little*,” while P6, who couldn’t “*stand the huge deal of work of figuring out stuff that I’m not used to*” said “*now I really feel like I’m chilling in the passenger seat and not having to do all the heavy-lifting personally.*”

Second, participants seemed to appreciate the in-context annotations and summaries of each paragraph provided by Selenite. They thought that this feature “*made things incredibly easy*” (P3) by “*helping me grasp the key points without wasting time reading a paragraph through*” (P1), and “*felt like back in the day when my classmate would mark all the important stuff in the textbook after a class when I couldn’t make it.*” However, some did report that the in-context annotations can occasionally be “*a little bit distracting*”, especially for paragraphs that are “*apparently unrelated to the main content*” (P2), such as those that talk about related articles or terms of services, suggesting that future versions of Selenite should consider more robust content filtering techniques.

Last but not least, participants also appreciated that Selenite can help them brainstorm search queries that would enable them to find new information more efficiently than was “*almost always one step ahead*” (P4), especially in the third task. For example, after reading two review articles about e-readers, Selenite suggested that P5 could do some additional investigations about “*supported file formats and syncing across devices*”. P5 admitted that “*I’d totally miss those if I’m by myself, and even if I’m trying to be super careful, it would take me forever to figure out that I need to check out those aspects.*” Additionally, we observed that when integrating the Selenite suggested criteria into subsequent search queries, the search engine did return result pages that turned out to be noticeably different yet sufficiently high-quality for users to explore.

*Impact on reading patterns and habits.* Participants all mentioned that they immediately checked out the commonly considered criteria from the sidebar before diving into reading the first web page. They claimed that compared to what they normally do, which is “*just have to hunker down and read*”, reading the overview first helped them “*cut to the chase and get a feel of what’s out there*” (P7) and remind them of criteria that would otherwise “*slip my [their] mind*” (P3).

<sup>15</sup>To see all the topics that participants explored in this study, please refer to Table 7 in the Appendix.

On a per-paragraph level, we noticed an initial hesitation among some participants (3 out of 8) towards relying solely on the provided criteria labels. As a safety precaution, they personally read through a handful of paragraphs to confirm the labels’ accuracy and reliability. We further corroborated this observation with their reflections, such as “*I’ve never seen anything like this before, so honestly, I was a bit skeptical at first. But hey, everything looked legit!*” (P5) After this initial hurdle, participants tend to “*rely on the labels to tell me the gist of a paragraph*” (P4) and only read paragraphs that discuss criteria that they truly cared about. For the content that participants did end up reading, they think the corresponding criteria “*definitely helped me [them] process and digest it better*” (P6), and even “*saved me [them] from otherwise misunderstanding things*” (P7). For example, while exploring healthy diet plans, P7 reflected that he would have initially thought a paragraph detailing the caloric allocation for each meal was seemingly discussing “*calorie intake*,” however, Selenite preemptively clarified that the focus was on portion control, i.e., “*providing guidelines on portion sizes.*”

Participants also enjoyed the easy navigation feature that Selenite offers, and used it to frequently jump between different criteria mentionings for easy comparison and digestion (7/8). They claimed that finding specific criteria about different options in a long article used to be “*link finding a needle in a haystack*” (P8) that they were hesitant to do, but with Selenite, “*it’s more like following a well-lit path*” (P5). For example, P2 reflected on her experience exploring VPN solutions, and claimed that “*now I get it, McAfee Safe Connect seems to be keeping track of all sorts of my information while SurfShark doesn’t do any of that. If I can’t quickly switch between these two points on the page, by the time I reach SurfShark’s no-logging policy, I would have totally forgotten about what McAfee does, or that I should even be concerned about logging at all.*” In addition, participants liked the fact that they can more effectively break out from the original structure and narrative of an article; for instance, P1 recounted that “*you don’t gotta stick to what the authors say anymore, ya know? Because, let’s face it, their storylines can get all tangled and complicated sometimes.*”

Last but not least, we did not observe much usage of the “zoom in” feature, where Selenite can leverage GPT-4 to provide a thorough analysis of a piece of content – only 3 participants tried it for a total of 8 times. We hypothesize that 1) the web pages utilized in the study were all professionally crafted, resulting in content that was relatively easy to comprehend; 2) the criteria labels generated by our NLI pipeline proved to be adequate in addressing the participants’ information needs; 3) the time required for the “zoom in” feature to provide a useful analysis, typically ranging from 5 to 10 seconds, still exceeded the participants’ patience and attention span. Future work could explore solutions to address this limited adoption from these perspectives, for example, with models that boast significantly increased inference speeds.

*Additional findings.* One interesting theme that emerged was that some participants (4/8) opted to use Selenite in the third task to revisit topics that they had previously explored and wanted to be able to “*double-check*” (P3) whether their prior understanding of the topic was truly comprehensive. Consistently, each participant uncovered something new that they hadn’t considered before. For example, P3, who had recently been making plans to move in with his partner, revisited the topic of “*choosing the right mattress*,” and

realized that he had never taken into consideration criteria such as `motion transfer` (i.e., “the extent to which movement on one side of the mattress affects the other side”) or `noise reduction` (i.e., “the ability of the mattress to minimize noise from springs, coils, or other components”), which prompted him to reassess his original mattress purchase. As another example, P4, a professional software engineer, revisited the topic of “*choosing a hybrid app framework*” and discovered that he had neglected to consider the `licensing and legal considerations` (i.e., “compliance with licensing requirements and legal considerations”) as suggested by Selenite. Consequently, P4 was able to find additional evidence to confirm the validity of their original framework choice made back in 2017.

In the post-study interview, many participants (6/8) felt that now they “*can’t imagine reading without a tool like this (Selenite)*” (P3). Half even inquired about the possibility of installing Selenite on their personal computers for post-study usage, and we gladly fulfilled their requests. Despite encountering a few bugs in our research prototype during the study and having no obligation or incentives for continued usage after the study, the fact that they were willing to do so suggests that our grounded reading approach indeed holds value for our participants.

## 8 DISCUSSION

Some of the participants (3/8) from the case study expressed concern about the coverage of Selenite’s overview criteria and the criteria labels for each paragraph at the beginning. They wondered if Selenite might overlook important criteria that they should also consider. This concern was valid, given that we presented the tool as an AI-powered oracle that could potentially be fallible or overlook certain factors and encouraged users to conduct their own explorations in addition to relying on Selenite’s insights. However, our accuracy and coverage evaluation described in section 5 provides an initial validation that the criteria and options provided by Selenite are indeed comprehensive, relevant, and accurate. In addition, after the study, participants also acknowledged that the current set of criteria offered by Selenite already “*far exceeds what I [they] could identify and keep track of on my [their] own*” (P4); therefore, they “*wouldn’t mind at all if the algorithm misses any minor ones*” (P1). Indeed, despite the participants’ awareness of the opportunity to request additional criteria from Selenite in the case of insufficient coverage (as confirmed in the post-study interviews), we did not observe any instances of such usage. Nevertheless, further research is necessary to investigate: 1) ways that would further improve the coverage and accuracy of Selenite, such as leveraging retrieval-augmented models [67]; 2) mechanisms and interventions designed to reduce over-dependence on Selenite as well as encourage critical thinking and user-led explorations.

Though primarily designed as a tool for grounded reading, Selenite might also have the potential to address some of the issues identified by prior work regarding structuring information during sensemaking – prior research suggested that asking users to structure information too early might lead to a more poorly structured information space [57]. In addition, the knowledge structures that people created often become obsolete, and new structures often emerge as their mental representations evolve over the course of

their investigation [37, 48, 57], resulting in having to spend significant effort in refactoring the structures every once in a while. Here, Selenite provides users with a well-structured framework from the outset, including a set of commonly acknowledged criteria. This readily usable scaffold serves as a starting point, aiming to encompass the majority’s perspective and thereby minimizing the necessity of refactoring or restructuring. Hopefully, it simplifies the iterative and cognitive-demanding process of building a mental model, transforming it into a possibly more manageable task of refining and pruning [72].

There was also a concern that GPT-4 could potentially hallucinate or generate irrelevant or even false criteria and thus mislead users in their subsequent exploration. However, it is important to note that in our case study, as well as in study 2, we did not observe any such episodes or evidence of this occurring. This could be attributed to the fact that the topics explored in the study were all common subjects with abundant source materials available online, which were likely encountered by GPT-4 during its training process. We would like to further conjecture that even if hallucination occurs, users can readily identify irrelevant or false criteria by carefully reading their descriptions and comparing them with common sense or their intuitive knowledge about the topic, mitigating the actual impact of hallucination.

## 9 LIMITATIONS & FUTURE WORK

*Connections among criteria.* In Selenite, we made the implicit assumption that criteria are completely independent. However, in reality, there could be connections between the criteria – for instance, when evaluating the “Best Baby Stroller,” the specific criterion of `suspension system` falls under the broader category of `safety (hierarchy)`, while on the other hand, aspects like `price` and `versatility` are typically trade-offs that are impractical to optimize for simultaneously (*correlation*). Currently, Selenite takes into account one form of criteria connections, i.e., *relevance* between criteria, when suggesting the next steps. This proved promising in the study, which gave us reasons to believe that further exploiting these connections between criteria can better support users’ reading. For example, instead of presenting the criteria in a simple list, one can imagine creating a behind-the-scenes knowledge graph where criteria are connected using edges of relations (TypeOf, CompetesWith, etc.). By initially displaying a portion of this graph and allowing users to “zoom in” on the specific criteria they are interested in (e.g., a subset of “safety”-related features), we can help users intuitively reason through an initially overwhelming list. In addition, one can again imagine “overview first, details later”-style UIs [97] that accommodate criteria hierarchies, e.g., multi-level tables or lists, granting users the flexibility to combine or decompose criteria at decision time.

*Availability of domain knowledge in LLMs.* LLMs, such as GPT-4, possess an extensive range of encoded knowledge, yet they might lack domain-specific information for specialized or emerging topics, as well as for topics involving confidential or sensitive information. Our technical implementation in Selenite is primarily based on extracting knowledge (e.g., commonly considered criteria) from commercially available LLMs, and its effectiveness is highly dependent on the LLM’s capability to capture and synthesize relevant

domain knowledge from its training data. Without such knowledge, the guidance provided may be subpar. In addition, LLMs themselves can sometimes be biased, and the response they generate might be incorrect or harmful [62, 81]. However, our approach to ground the reading process with domain knowledge would also work with other sources of knowledge bases as well, for example, Unakite + Strata tables [70, 71], or crowdsourced [42, 79], or a combination of them. Furthermore, we should also urge users to thoroughly examine the Selenite overview when dealing with critical situations.

*Generalizing beyond comparison tasks.* In this work, we focus on helping people with sensemaking tasks that often need users to systematically compare different options with respect to various criteria. As evidenced by our formative as well as case studies, it is beneficial for people to be aware of the criteria that other people commonly consider upfront to help with their subsequent sensemaking journey. However, there are other types of sensemaking tasks, such as those that are purely exploratory or investigatory (e.g., debugging, learning a new skill), that do not entail well-established options and criteria, and as such, they are not ideally compatible for Selenite to assist with. Nevertheless, we postulate that the notion of procuring comprehensive expert perspectives upfront may still apply in these non-comparison tasks – for example, one can imagine asking an LLM for advice on a range of typical strategies to try when debugging or a list of common steps to take to master a new skill. Future research can work on enabling users to obtain these categories of overviews by adapting and customizing the LLM prompts used in Selenite (that were originally used to obtain criteria) and continue to receive similar in-context annotations and reading guidance grounded on those overviews.

*Impact on learning.* The current design of Selenite functions as an “index” to direct users to relevant parts of a web page for reading and processing. However, we need to be cautious about a potential risk associated with this approach – some users might believe they have gained sufficient knowledge about a topic by merely reading the overview and may, therefore, skip engaging with the actual web content. This behavior could lead to incomplete, biased, or even inaccurate understandings of the subject. It is akin to only reading the table of contents or indices of a book without delving into the actual passages. Nevertheless, our studies conducted under controlled settings have shown that participants did, in fact, engage with the actual web content after going through the overviews. To build on this promising evidence, future research should additionally investigate interface and interaction designs that motivate users to explore and read the actual web content with the assistance of Selenite-style guidance. One potential approach could be progressively revealing criteria information to users based on their reading behavior, encouraging deeper exploration and understanding.

*Field Study.* In the future, once all the bugs and usability issues have been thoroughly addressed, we aim to conduct an extensive, long-term field study on a larger scale, where people will have both sufficient motivations to investigate topics relevant to their own personal context and familiarity with Selenite through repeated usage. This could potentially shed light on situations where Selenite performs reasonably well, as well as situations where it may fall short. Additionally, we are also interested in understanding

Selenite’s long-term impact on individuals’ analytical skills and problem-solving abilities.

## 10 CONCLUSION

Sensemaking in unfamiliar domains can often be challenging, with users having to sift through large volumes of information and compare different options with respect to various criteria. Previous sensemaking research as well as our new formative study has shown that people would benefit from seeing an overview of the information space, such as the criteria that others have previously found useful. However, existing systems have been limited by the “cold start” issue — they require substantial effort from previous users to gather and structure information to produce such an overview, and, even if it has been produced, there are no straightforward methods of sharing that overview with future users and making it so that future users would find it to be comprehensive, unbiased, and useful.

In this work, by leveraging recent advances in LLMs and natural language processing, we introduce a novel system named Selenite that automates finding an initial set of options, criteria, and evidence, and provides a comprehensive overview to users at the start of their sensemaking process. In addition, it also adapts as people use it, helping users find, read, and navigate unfamiliar information in a systematic yet personalized manner. As such, it provides a valuable proof of concept of how a future LLM-powered sensemaking tool that provides users with comprehensive overviews and in-context reading guidance can scaffold their sensemaking and learning of an unfamiliar space.

## ACKNOWLEDGMENTS

This research was supported in part by NSF grants CCF-1814826 and FW-HTF-RL-1928631, Google, Adobe, Oracle, Bosch, the Office of Naval Research, and the CMU Center for Knowledge Acceleration. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. We would like to thank our study participants for their kind participation and our anonymous reviewers for their insightful feedback. We sincerely thank Daniel M. Russell, Kenneth Holstein, Toby Jia-Jun Li, Julia Qian, Haojian Jin, Xu Wang, Anhong Guo, Ziyan Wang, Yi Zhou, Zheng Yao, Tianshi Li, Eric Yiyi Wang, Amber Horvath, Daye Nam, Nimo Ni, Meng Xia, Yongsung Kim, Andrew Kuznetsov, Hyeonsu B. Kang, Wesley Hanwen Deng, and Queenie Kravitz for their extraordinary feedback and support.

## REFERENCES

- [1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A Review on Language Models as Knowledge Bases. <https://doi.org/10.48550/arXiv.2204.06031> [cs].
- [2] Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3675–3686. <https://doi.org/10.18653/v1/D18-1403>
- [3] Anthropic. 2023. Claude. <https://claude.ai/chats>
- [4] Kumaripab Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. 2016. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. Association for Computing Machinery, New York, NY, USA, 359–369. <https://doi.org/10.1145/2856767.2856786>

- [5] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* (April 2023). <https://doi.org/10.1145/3589955> Just Accepted.
- [6] Michelle Q. Wang Baldonado and Terry Winograd. 1997. SenseMaker: An Information-exploration Interface Supporting the Contextual Evolution of a User's Interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, New York, NY, USA, 11–18. <https://doi.org/10.1145/258549.258563>
- [7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holly Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. <https://doi.org/10.48550/arXiv.2302.04023> [cs].
- [8] Krishna Bharat. 2000. SearchPad: explicit capture of search context to support Web search. *Computer Networks* 33, 1 (June 2000), 493–501. [https://doi.org/10.1016/S1389-1286\(00\)00047-5](https://doi.org/10.1016/S1389-1286(00)00047-5)
- [9] BlockTechnology. 2023. AskYourPDF: The Best PDF AI Chat App. <https://askyourpdf.com>
- [10] Samuel Brody and Noemie Elhadad. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn (Eds.). Association for Computational Linguistics, Los Angeles, California, 804–812. <https://aclanthology.org/N10-1122>
- [11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. <https://doi.org/10.48550/arXiv.2303.12712> arXiv:2303.12712 [cs].
- [12] Jürgen Buder, Christina Schwind, Anja Rudat, and Daniel Bodemer. 2015. Selective reading of large online forum discussions: The impact of rating visualizations on navigation and learning. *Computers in Human Behavior* 44 (March 2015), 191–201. <https://doi.org/10.1016/j.chb.2014.11.043>
- [13] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? <https://doi.org/10.48550/arXiv.2307.03826> arXiv:2307.03826 [cs].
- [14] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016. Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 61–68. <https://doi.org/10.1145/2984511.2984538>
- [15] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding Comparison Tables for Online Decision Making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 391–405. <https://doi.org/10.1145/3379337.3415865>
- [16] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. Search-Lens: composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, Marina del Ray, California, 498–509. <https://doi.org/10.1145/3301275.3302321>
- [17] Joseph Chee Chang, Yongsung Kim, Victor Miller, Michael Xieyang Liu, Brad A Myers, and Aniket Kittur. 2021. Tabs.do: Task-Centric Browser Tab Management. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 663–676. <https://doi.org/10.1145/3472749.3474777>
- [18] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with Crowds and Computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3180–3191. <https://doi.org/10.1145/2858036.2858411>
- [19] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE. Google-Books-ID: 2ThdBAAAQBAJ.
- [20] Tianying Chen, Michael Xieyang Liu, Emily Ding, Emma O'Neil, Mansi Agarwal, Robert E Kraut, and Laura Dabbish. 2023. Facilitating Counselor Reflective Learning with a Real-time Annotation tool. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3544548.3581551>
- [21] Xiang'Anthony' Chen, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. Marvista: A Human-AI Collaborative Reading Tool. *arXiv preprint arXiv:2207.08401* (2022).
- [22] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1999–2008. <https://doi.org/10.1145/2470654.2466265>
- [23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyoontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezh Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. <https://doi.org/10.48550/arXiv.2204.02311> arXiv:2204.02311 [cs].
- [24] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 443–452. <https://doi.org/10.1145/2207676.2207738>
- [25] Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the Internal Knowledge-Base of Language Models. <https://doi.org/10.48550/arXiv.2301.12810> arXiv:2301.12810 [cs].
- [26] Dick Cunningham and Scott L. Shablak. 1975. Selective Reading Guide-O-Rama: The Content Teacher's Best Friend. *The Journal of Reading* (1975).
- [27] Subha Jyoti Das, Riki Murakami, and Basabi Chakraborty. 2021. Development of a two-step LDA based aspect extraction technique for review summarization. *International Journal of Applied Science and Engineering* 18, 1 (2021), 1–18. [https://doi.org/10.6703/IJASE.202103\\_18\(1\).003](https://doi.org/10.6703/IJASE.202103_18(1).003) Publisher: Chaoyang University of Technology.
- [28] Brenda Dervin. 1983. An overview of sense-making research concepts, methods, and results to date. (1983). <http://www.worldcat.org/title/overview-of-sense-making-research-concepts-methods-and-results-to-date/oclc/733067203>
- [29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* [cs] (May 2019). <http://arxiv.org/abs/1810.04805> arXiv: 1810.04805.
- [30] G.A. Di Lucca, M. Di Penta, and A.R. Fasolino. 2002. An approach to identify duplicated web pages. In *Proceedings 26th Annual International Computer Software and Applications*. 481–486. <https://doi.org/10.1109/CMSAC.2002.1045051> ISSN: 0730-3157.
- [31] Mira Dontcheva, Steven M. Drucker, Geraldine Wade, David Salesin, and Michael F. Cohen. 2006. Summarizing Personal Web Browsing Sessions. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST '06)*. ACM, New York, NY, USA, 115–124. <https://doi.org/10.1145/1166253.1166273>
- [32] Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work (CSCW '92)*. Association for Computing Machinery, Toronto, Ontario, Canada, 107–114. <https://doi.org/10.1145/143457.143468>
- [33] Evernote. [n. d.]. Best Note Taking App - Organize Your Notes with Evernote. <https://evernote.com>
- [34] Facebook. 2018. React - A JavaScript library for building user interfaces. <https://reactjs.org/>
- [35] David K. Farkas and Christopher Raleigh. 2013. Designing Documents for Selective Reading. *Information Design Journal* 20, 1 (Jan. 2013), 2–15. <https://doi.org/10.1075/ijd.20.1.01far> Publisher: John Benjamins.
- [36] Kasra Ferdowsi, Jack Williams, Ian Drosos, Andrew D. Gordon, Carina Negreanu, Nadia Polikarpova, Advait Sarkar, and Benjamin Zorn. 2023. COLDECO: An End User Spreadsheet Inspection Tool for AI-Generated Code. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 82–91. <https://doi.org/10.1109/VL-HCC57772.2023.00017> ISSN: 1943-6106.
- [37] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed Sensemaking: Improving Sensemaking by Leveraging the Efforts of Previous Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 247–256. <https://doi.org/10.1145/2207676.2207711>
- [38] Andrew J. Flanagan and Miriam J. Metzger. 2000. Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly* 77, 3 (Sept. 2000), 515–540. <https://doi.org/10.1177/10776990000770034> Publisher: SAGE Publications Inc.
- [39] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S. Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 476–490. <https://doi.org/10.1145/3581641.3584034>

- [40] Saskia Gilmer, Avinash Bhat, Shuvam Shah, Kevin Cherry, Jinghui Cheng, and Jin L. C. Guo. 2023. SUMMIT: Scaffolding OSS Issue Discussion Through Summarization. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–27. <https://doi.org/10.1145/3610088> arXiv:2308.02780 [cs].
- [41] Andrew D. Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, Jack Williams, and Ben Zorn. 2023. Co-audit: tools to help humans double-check AI-generated content. <https://doi.org/10.48550/arXiv.2310.01297> [cs].
- [42] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2258–2270. <https://doi.org/10.1145/2858036.2858364>
- [43] Nathan Hahn, Joseph Chee Chang, and Aniket Kittur. 2018. Bento Browser: Complex Mobile Search Without Tabs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Montréal QC, Canada, 251:1–251:12. <https://doi.org/10.1145/3173574.3173825>
- [44] Dianne J. Hall and Robert A. Davis. 2007. Engaging multiple perspectives: A value-based decision-making model. *Decision Support Systems* 43, 4 (Aug. 2007), 1588–1604. <https://doi.org/10.1016/j.dss.2006.03.004>
- [45] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.), Human Mental Workload, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [46] Andrew Head, Amber Xie, and Marti A. Hearst. 2022. Math Augmentation: How Authors Enhance the Readability of Formulas using Novel Visual Design Practices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3491102.3501932>
- [47] Marti A. Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM* 49, 4 (April 2006), 59–61. <https://doi.org/10.1145/1121949.1121983>
- [48] Marti A. Hearst. 2014. What's Missing from Collaborative Search? *Computer* 47, 3 (March 2014), 58–61. <https://doi.org/10.1109/MC.2014.77>
- [49] Lawrence J. Henschien and Julia C. Lee. 2009. Using Semantic-Level Tags in HTML/XML Documents. In *Universal Access in Human-Computer Interaction. Applications and Services (Lecture Notes in Computer Science)*, Constantine Stephanidis (Ed.). Springer, Berlin, Heidelberg, 683–692. [https://doi.org/10.1007/978-3-642-02713\\_0\\_72](https://doi.org/10.1007/978-3-642-02713_0_72)
- [50] Andrew Hogue and David Karger. 2005. Thresher: automating the unwrapping of semantic content from the World Wide Web. In *Proceedings of the 14th international conference on World Wide Web (WWW '05)*. Association for Computing Machinery, New York, NY, USA, 86–95. <https://doi.org/10.1145/1060745.1060762>
- [51] Johan F. Hoorn and Teunis D. van Wijngaarden. 2010. Web Intelligence for the Assessment of Information Quality: Credibility, Correctness, and Readability. *Web Intelligence and Intelligent Agents* (March 2010). <https://doi.org/10.5772/8372> Publisher: IntechOpen.
- [52] Amber Horvath, Michael Xieyang Liu, River Hendriksen, Connor Shannon, Emma Paterson, Kazi Jawad, Andrew Macvean, and Brad A Myers. 2022. Understanding How Programmers Can Use Annotations on Documentation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3491102.3502095>
- [53] Jane Hsieh, Michael Xieyang Liu, Brad A. Myers, and Aniket Kittur. 2018. An Exploratory Study of Web Foraging to Understand and Support Programming Decisions. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 305–306. <https://doi.org/10.1109/VLHCC.2018.8506517> ISSN: 1943-6092.
- [54] Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. Association for Computing Machinery, New York, NY, USA, 1195–1204. <https://doi.org/10.1145/1557019.1557148>
- [55] Hyeyoung B. Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. <https://doi.org/10.1145/3526113.3545660> arXiv:2208.03455 [cs].
- [56] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the Schemas of Giants: Socially Augmented Information Foraging. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 999–1010. <https://doi.org/10.1145/2531602.2531644>
- [57] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, Trupti Telang, and Michael R. Bove. 2013. Costs and Benefits of Structured Information Foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2989–2998. <https://doi.org/10.1145/2470654.2481415>
- [58] Aniket Kittur, Bongwon Suh, and Ed H. Chi. 2008. Can you ever trust a wiki? impacting perceived trustworthiness in wikipedia. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08)*. Association for Computing Machinery, San Diego, CA, USA, 477–480. <https://doi.org/10.1145/1460563.1460639>
- [59] G. Klein, B. Moon, and R. R. Hoffman. 2006. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 21, 4 (July 2006), 70–73. <https://doi.org/10.1109/MIS.2006.75>
- [60] Torkel Klingberg. 2009. *The Overflowing Brain: Information Overload and the Limits of Working Memory*. Oxford University Press, USA. Google-Books-ID: IxMSDAAAQBAJ.
- [61] Weize Kong and James Allan. 2014. Extending Faceted Search to the General Web. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. Association for Computing Machinery, New York, NY, USA, 839–848. <https://doi.org/10.1145/2661829.2661964>
- [62] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 3299–3321. <https://aclanthology.org/2023.eacl-main.241>
- [63] Matevž Kunaver and Tomáš Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (May 2017), 154–162. <https://doi.org/10.1016/j.knosys.2017.02.009>
- [64] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. <https://doi.org/10.1145/3526113.3545693> arXiv:2208.14861 [cs].
- [65] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction* 34, 7 (July 2018), 577–590. <https://doi.org/10.1080/1047318.2018.1455307> Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/1047318.2018.1455307>.
- [66] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR* abs/1910.13461 (2019). <http://arxiv.org/abs/1910.13461> arXiv: 1910.13461.
- [67] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205780e1bc26945df7481e5-Abstract.html>
- [68] Michael Xieyang Liu. 2023. *Tool Support for Knowledge Foraging, Structuring, and Transfer during Online Sensemaking*. Ph.D. Dissertation. Carnegie Mellon University. <http://reports-archive.adm.cs.cmu.edu/anon/usr0/ftp/usr/ftp/heii/abstracts/23-105.html>
- [69] Michael Xieyang Liu, Nathan Hahn, Angelina Zhou, Shaun Burley, Emily Deng, Aniket Kittur, and Brad A. Myers. 2018. UNAKITE: Support Developers for Capturing and Persisting Design Rationales When Solving Problems Using Web Resources. *Workshop on Designing Technologies to Support Human Problem Solving at the IEEE Symposium on Visual Languages and Human-Centric Computing* (Oct. 2018). <https://par.nsf.gov/biblio/10152060-unakite-support-developers-capturing-persisting-design-rationales-when-solving-problems-using-web-resources>
- [70] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A. Myers. 2019. Ünakte: Scaffolding Developers' Decision-Making Using the Web. In *Proceedings of the 32Nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. ACM, New Orleans, LA, USA, 67–80. <https://doi.org/10.1145/332165.3347908> event-place: New Orleans, LA, USA.
- [71] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2021. To Reuse or Not To Reuse? A Framework and System for Evaluating Summarized Knowledge. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 166:1–166:35. <https://doi.org/10.1145/3449240>
- [72] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2022. Crystalline: Lowering the Cost for Developers to Collect and Organize Information for Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3491102.3501968> event-place: New Orleans, LA, USA.
- [73] Michael Xieyang Liu, Andrew Kuznetsov, Yongsung Kim, Joseph Chee Chang, Aniket Kittur, and Brad A. Myers. 2022. Wigglite: Low-cost Information Collection and Triage. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3526113.3545661>

- [74] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. “What It Wants Me To Say”: Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–31. <https://doi.org/10.1145/3544548.3580817>
- [75] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. [https://doi.org/10.48550/arXiv.2304.09848 arXiv:2304.09848 \[cs\]](https://doi.org/10.48550/arXiv.2304.09848).
- [76] Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2023. HypoCompass: Large-Language-Model-based Tutor for Hypothesis Construction in Debugging for Novices. [https://doi.org/10.48550/arXiv.2310.05292 arXiv:2310.05292 \[cs\]](https://doi.org/10.48550/arXiv.2310.05292).
- [77] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. [https://doi.org/10.48550/arXiv.2303.17651 arXiv:2303.17651 \[cs\]](https://doi.org/10.48550/arXiv.2303.17651).
- [78] Joaquim Mendes, Nuno Laranjeiro, and Marco Vieira. 2018. Toward characterizing HTML defects on the Web. *Software: Practice and Experience* 48, 3 (2018), 750–757. [https://doi.org/10.1002/spe.2545 \\_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.2545](https://doi.org/10.1002/spe.2545 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.2545).
- [79] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction* 14, 4 (Nov. 2021), 272–344. <https://doi.org/10.1561/1100000083> Publisher: Now Publishers, Inc..
- [80] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: An Interface for Collaborative Web Search. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*. ACM, New York, NY, USA, 3–12. <https://doi.org/10.1145/1294211.1294215>
- [81] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5356–5371. <https://doi.org/10.18653/v1/2021.acl-long.416>
- [82] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 189–199. <https://doi.org/10.1145/3242587.3242666>
- [83] OpenAI. 2023. ChatGPT. <https://chat.openai.com>
- [84] OpenAI. 2023. GPT-4 Technical Report. [https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 \[cs\]](https://doi.org/10.48550/arXiv.2303.08774).
- [85] Sharoda A. Paul and Meredith Ringel Morris. 2009. CoSense: Enhancing Sensemaking for Collaborative Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1771–1780. <https://doi.org/10.1145/1518701.1518974>
- [86] Alvaro Pereira, Ricardo Baeza-Yates, and Nívio Ziviani. 2006. Where and How Duplicates Occur in the Web. In *2006 Fourth Latin American Web Congress*. 127–134. <https://doi.org/10.1109/LA-WEB.2006.39>
- [87] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*. <http://www.phibetaiota.net/wp-content/uploads/2014/12/Sensemaking-Process-Pirolli-and-Card.pdf>
- [88] Yannis Plegas and Sofia Stamou. 2013. Reducing information redundancy in search results. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13)*. Association for Computing Machinery, New York, NY, USA, 886–893. <https://doi.org/10.1145/2480362.2480533>
- [89] Marlene A. Plumlee. 2003. The Effect of Information Complexity on Analysts’ Use of That Information. *The Accounting Review* 78, 1 (Jan. 2003), 275–296. <https://doi.org/10.2308/accr2003.78.1.275>
- [90] G. Poonkuzhal, R. Kishore Kumar, R. Kripa Keshav, P. Sudhakar, and K. Sarukesi. 2011. Correlation Based Method to Detect and Remove Redundant Web Document. *Advanced Materials Research* 171–172 (2011), 543–546. <https://doi.org/10.4028/www.scientific.net/AMR.171-172.543> Publisher: Trans Tech Publications Ltd.
- [91] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*. 28–37.
- [92] Napol Rachatasumit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719. <https://doi.org/10.1145/3490099.3511162>
- [93] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [94] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [https://doi.org/10.48550/arXiv.1908.10084 arXiv:1908.10084 \[cs\]](https://doi.org/10.48550/arXiv.1908.10084).
- [95] Daniel M. Russell, Mark J. Stefk, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 269–276. <https://doi.org/10.1145/169059.169209>
- [96] M. schraefel, Yuxiang Zhu, David Modjeska, Daniel Wigdor, and Shengdong Zhao. 2002. Hunter Gatherer: Interaction Support for the Creation and Management of Within-web-page Collections. In *Proceedings of the 11th International Conference on World Wide Web (WWW '02)*. ACM, New York, NY, USA, 172–181. <https://doi.org/10.1145/511446.511469>
- [97] Ben Shneiderman. 2000. Designing trust into online experiences. *Commun. ACM* 43, 12 (Dec. 2000), 57–59. <https://doi.org/10.1145/355112.355124>
- [98] Alexandra N Spichtig, Elfrieda H Hiebert, Christian Vorstius, Jeffrey P Pascoe, P David Pearson, and Ralph Radach. 2016. The decline of comprehension-based silent reading efficiency in the United States: A comparison of current data with performance in 1960. *Reading Research Quarterly* 51, 2 (2016), 239–259. Publisher: Wiley Online Library.
- [99] Bongwon Suh, Ed H. Chi, Aniket Kittur, and Bryan A. Pendleton. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, Florence, Italy, 1037–1040. <https://doi.org/10.1145/1357054.1357214>
- [100] João Sá, Vanessa Queiroz Marinho, Ana Rita Magalhães, Tiago Lacerda, and Diogo Gonçalves. 2022. Diversity Vs Relevance: A Practical Multi-objective Study in Luxury Fashion Recommendations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2405–2409. <https://doi.org/10.1145/3477495.3531866>
- [101] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. [https://doi.org/10.48550/arXiv.2311.00710 arXiv:2311.00710 \[cs\]](https://doi.org/10.48550/arXiv.2311.00710).
- [102] H. Holden Thorp. 2023. ChatGPT is fun, but not an author. *Science* 379, 6630 (Jan. 2023), 313–313. <https://doi.org/10.1126/science.adg7879> Publisher: American Association for the Advancement of Science.
- [103] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaee, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [104] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131. Publisher: American association for the advancement of science.
- [105] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. [https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 \[cs\]](https://doi.org/10.48550/arXiv.1706.03762).
- [106] Chengguang Wang, Xiao Liu, and Dawn Song. 2020. Language Models are Open Knowledge Graphs. [https://doi.org/10.48550/arXiv.2010.11967 arXiv:2010.11967 \[cs\]](https://doi.org/10.48550/arXiv.2010.11967).
- [107] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. [https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 \[cs\]](https://doi.org/10.48550/arXiv.2201.11903).
- [108] Dale M. Willows. 1974. Reading between the Lines: Selective Attention in Good and Poor Readers. *Child Development* 45, 2 (1974), 408–415. <https://doi.org/10.2307/1127962> Publisher: [Wiley, Society for Research in Child Development].
- [109] Dale M. Willows and G. E. MacKinnon. 1973. Selective reading: Attention to the “unattended” lines. *Canadian Journal of Psychology / Revue canadienne de psychologie* 27, 3 (1973), 292–304. <https://doi.org/10.1037/h0082480> Place: Canada Publisher: University of Toronto Press.
- [110] Wondershare. 2023. Chat with PDF \*Powered by ChatGPT. <https://www.hipdf.com/chat-with-pdf>
- [111] Nicholas C. Wormald. 1995. Differential Equations for Random Processes and Random Graphs. *The Annals of Applied Probability* 5, 4 (1995), 1217 – 1235. <https://doi.org/10.1214/aoap/1177004612> Publisher: Institute of Mathematical Statistics.
- [112] Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis, Christian Kaestner, and Tongshuang Wu. 2023. Beyond Testers’ Biases: Guiding Model Testing with Knowledge Bases using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore,

- 13504–13519. <https://doi.org/10.18653/v1/2023.findings-emnlp.901>
- [113] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 3 (June 2022), 3081–3089. <https://doi.org/10.1609/aaai.v36i3.20215> Number: 3.
- [114] Ilan Yaniv and Shoham Choshen-Hillel. 2012. When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology* 48, 5 (Sept. 2012), 1022–1028. <https://doi.org/10.1016/j.jesp.2012.03.016>
- [115] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. <https://doi.org/10.48550/arXiv.1909.00161> arXiv:1909.00161 [cs].
- [116] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat Through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 196:1–196:27. <https://doi.org/10.1145/3274465>
- [117] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 2082–2096. <https://doi.org/10.1145/2998181.2998235>
- [118] Yi Zong and Xiaojie Guo. 2022. An Experimental Study on Anchoring Effect of Consumers' Price Judgment Based on Consumers' Experiencing Scenes. *Frontiers in Psychology* 13 (2022). <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.794135>

## A UNFAMILIAR TOPICS PARTICIPANTS EXPLORED IN THE FORMATIVE STUDY

Index	Theme	Topic	Participants
1	Software & online services	Choosing a hybrid app framework	P2, P5
2		Selecting a secure password manager	P3, P7
3		Choosing a suitable ERP (Enterprise Resource Planning) solution	P1
4		Choosing a reliable VPN (Virtual Private Network) provider	P1, P5
5		Picking a deep learning framework	P8
6		Deciding on the best data visualization tool	P6
7		Choosing the best time tracking tool	P4
8	Consumer Electronics & Technology	Choosing a high-quality digital camera	P2, P5
9		Choosing the best action camera	P8
10		Selecting a VR headset	P7
11		Picking a drone	P3, P5
12		Picking a smart home ecosystem	P1, P6, P8
13	Home Appliances & Furniture	Picking the best robot vacuum	P2, P5
14		Choosing the best air purifier	P4
15		Selecting the best washing machine	P3
16		Picking the right refrigerator	P3
17		Selecting the best mattress	P3, P6
18	Outdoor & Adventure	Choosing the best city bike	P7
19		Choosing the best barbecue grill	P3
20		Choosing the best tropical vacation location	P4
21	Health & fitness	Choosing an effective diet plan	P1, P7
22		Picking a reliable treadmill	P3
23		Picking the best running shoes	P8
24	Gifts & special events	Choosing a birthday gift	P5, P6
25		Picking the right wedding venue	P2
26		Picking an engagement ring	P2
27	Parenting	Choosing the best baby stroller	P8
28	Pets	Choosing a breed of dog to adopt	P4

**Table 5: Unfamiliar topics (organized by themes) that participants in the formative study reported encountering and exploring. Some topics were explored by multiple participants, such as “Picking a smart home ecosystem” and “Choosing a reliable VPN provider.”**

## B STUDY 1 DETAILS

### B.1 Example web pages

Below in Figure 5, we show a partial snapshot of a representative comparison article on the topic of “best baby strollers.”

**GearLab Baby** IN DEPTH BABY PRODUCT REVIEWS LED BY A PEDIATRICIAN

**10 Best Strollers of 2023**

We've tested strollers from Britax, BOB, Thule, Chicco, and more to find the best options in every category



**Best Light & Compact Baby Stroller**  
Thule Spring



**Editor's Choice**  
\$400 at Amazon

**Weight:** 22.7 lbs | **Folded Size:** 7,650 cubic inches

<b>REASONS TO BUY</b>	<b>REASONS TO AVOID</b>
+ Lightweight for a full-size	- Plastic wheels
+ Compact & quick fold	- Less impressive off-road
+ Reasonable price	

The **Thule Spring** is a full-size stroller that is surprisingly light and compact. It has an ultra-quick fold, comfortable seating, and uses quality materials. It is an easy-to-use, full-size stroller from Thule. It is one of the most compact folded options in its category, making it suitable for families with limited space. It includes a large storage bin, a deep recline, and a large canopy. It is compatible with multiple infant car seats from Chicco, Maxi-Cosi, Nuna, and Cybex.

This full-size stroller is relatively easy to push, but the plastic wheels make uneven terrain, gravel, and grass more challenging, unlike choosing a stroller like the **Thule Urban Glide 2** with rubber, pneumatic tires. However, the plastic wheels are still of higher quality than most plastic competitors like those found on the **Evenflo Pivot Xpand**. We think this impressive stroller works well, is reasonably priced, and includes features parents, and little ones will enjoy.

**Best Overall Baby Stroller**  
UPPAbaby Cruz v2



**Top Pick**  
\$690 at Amazon

**Weight:** 25 lbs | **Folded Size:** 12,469 cubic inches

<b>REASONS TO BUY</b>	<b>REASONS TO AVOID</b>
+ High-quality	- Expensive
+ Great features	
+ Works great with an infant car seat	

If we picked a single option for friends concerned with product quality, it would be the **UPPAbaby Cruz v2**. No question. The Cruz v2 is lightweight (though not as lightweight as the original **UPPAbaby Cruz**), easy to use, and uses high-quality materials that feel durable. Plus, the Cruz v2 works well with infant car seats, including a few of our favorites, like the **UPPAbaby Mesa** and **Chicco KeyFit 35**. We love the vast storage basket with a 30 lb weight limit, the super large canopy, and the addition of suspension on the v2, making the new Cruz better than any full-size competitor.

The wheels on the Cruz are relatively small, and their disappointing size makes it more challenging to traverse uneven surfaces than the larger tires on joggers. However, the Cruz offers excellent maneuverability on smooth surfaces making it highly suitable for running errands and public venues. While the Cruz v2 is more expensive than most of the **best full-size strollers** and jogging competitors, many users feel its higher quality and easy-to-use features more than justify the higher cost. If you need a similar option but plan to grow your family, the **UPPAbaby Vista v2** can hold up to three little ones of different ages and abilities.

**Best High-end Stroller**  
UPPAbaby Vista v2



**Top Pick**  
\$1,000 at Amazon

**Weight:** 27.1 lbs | **Folded Size:** 14,157 cubic inches

<b>REASONS TO BUY</b>	<b>REASONS TO AVOID</b>
+ High-quality	- Stiff price tag
+ Expands for three children	- Heavier
+ Excellent storage	

The **UPPAbaby Vista v2** is a fantastic high-end choice we recommend for parents that want a little more of everything and aren't worried about the price tag that comes with it. The Vista v2 is one of the **best double strollers** as it can grow with your family, carrying up to three little ones with ease (age-dependent). The Vista v2 is stylish, easy to use, and includes quality parts and attention to detail. This product has various seating configurations and consists of a toddler seat and bassinet in the original purchase. The Vista v2 works with two toddler seats, two infant carriers (seats and adapters purchased separately), two bassinets, or any combination of seating facing forward or backward. It features a giant storage bin and canopy with a nap-worthy seat recline. The Vista v2 delivers high functionality, a lower price tag, and unusually lower weight than similar competitors. The Vista v2 moves well for its size, and we think parents will enjoy its adjustable handlebar and giant under-seat storage bin with an unheard-of 30 lb max allowance.

The Vista v2 has one of the highest prices in our reviews, making it a potential no-go for families on a tight budget. The **Evenflo Pivot Xpand Double** is a similarly styled stroller with a lower price, but the lower price also comes with lower quality than the Vista, in our opinion. Also, it is not suitable for running. However, we think it's tough to beat the Vista v2, especially for families with multiple children or plans to grow if your budget allows. If the price tag is not in your wheelhouse or a double option isn't necessary, then the less spendy, smaller, and lighter **UPPAbaby Cruz v2** should make your list instead.

**Big City Taxi Suitable Stroller**  
Doona Combo



**Top Pick**  
\$550 at Amazon

**Weight:** 16.5 lbs | **Folded Size:** 8,150 cubic inches

<b>REASONS TO BUY</b>	<b>REASONS TO AVOID</b>
+ Unique combo with car seat	- Heavy car seat
+ Excels at city life	- Relatively higher price
+ Easy to use	

The **Doona Combo** is a unique car seat included in our **best car seat** review. This strolling combo makes a simple stroller with an infant seat for a useful combination for city dwellers. This cool mash-up makes the Doona a ridiculously easy-to-use product that is good for those who frequent public transportation and sidewalks. This combo lets parents push their baby from place to place, from home to a day on the town, including public transportation. The car seat is easy to use, installs quickly without the base, and fills a niche for parents who live in the city and may otherwise choose to forgo a car seat for convenience.

The Doona Combo may not be ideal for every family, thanks to its relatively higher price, heavy carrying weight, and limited lifespan. If you want to stroll with your car seat carrier but need a budget-friendly solution, then pairing a Chicco brand seat with a frame stroller like the **Chicco KeyFit Caddy** could be a smarter solution than the Doona Combo. Still, you can't beat it for city families that may otherwise legally skip using a car seat because they are a pain to carry or require a strolling companion. We think this is the easiest way to get your baby from an apartment to an Uber to strolling with ease and enviable style, making it a must-have for parents on-the-go, and those who frequent public transportation. However, if you want the best car seat, the Doona can't compare to the **Nuna Pipa**.

**Inexpensive, Casual Jogging Stroller**  
Baby Trend Expedition



**Best Value**

**Figure 5: Example comparison article on the topic of “best baby strollers.” Note that it only contains approximately 1/4 of the article, which originally includes content about 10 baby strollers, along with other long-form commentaries. The article can be found online at <https://www.babygearlab.com/topics/getting-around/best-stroller>.**

## B.2 Dataset of Groundtruth Criteria vs Selenite-retrieved Criteria

Topic	Groundtruth Criteria	Selenite-retrieved Criteria
Best washing machines	Cleaning performance*, Fabric care*, Water usage*, Smart features*, Durability*, Water Temperature Options*, Noise level*, Load type*, Odor control*, Ease-of-use*, Price*, Aesthetics*, Power source*, Speed & Cycle time*, Features*, Capacity*, Warranty & Customer support*, Size*, Brand reputation* (n = 19)	Capacity*, Energy efficiency, Water efficiency*, Noise level*, Durability*, Cleaning performance*, Cycle options*, Spin speed*, Price*, Size*, Brand reputation*, User-friendliness / ease-of-use*, Warranty*, Features*, Aesthetics*, Cycle time*, Load type*, Maintenance, Smart connectivity*, Fabric care*, Water Temperature Options*, Power source*, Child Safety Lock, Odor Control* (n = 24)
Birthday gift ideas	Practicality*, Aesthetics*, Experience*, Price*, Age appropriateness*, Prerequisite, Uniqueness*, Sentimental value*, Personalization*, Gender appropriateness*, Shipping* (n = 11)	Personalization*, Uniqueness*, Practicality*, Sentimental value*, Presentation (wrapping), Price*, Creativity*, Durability, Functionality*, Brand, Age appropriateness*, Gender appropriateness*, Timeliness, Accessibility (can get it in time)*, Social norms, Environment impact, Size, Experience*, Aesthetic*, Relevance, Surprise factor (n = 21)
Best hybrid app frameworks	Programming Language, Code reusability*, Cross-platform compatibility*, Development time*, Ease of learning*, Third-Party Integration*, Popularity*, Features*, Testing and debugging*, Plugin availability*, Native features access*, Performance*, User interface*, Framework size*, Community support* (n = 15)	Cross-platform compatibility*, Performance*, User interface*, Community support*, Plugin availability*, Development time*, Maintenance, Cost*, Third-Party Integration*, Security*, Scalability*, Platform Support*, Ease of learning*, Code reusability*, Customization*, Testing and debugging*, App store compliance, Innovation, Popularity and Adoption*, Framework size*, Native features access* (n = 21)
Best time tracking tools	Real-time tracking*, Editing ability*, Export and Invoice*, Accessibility*, Price*, OS platform*, Features*, Easy-of-use*, Tracking accuracy*, Simplicity*, Extensibility*, Automation*, Privacy*, Learning curve*, Intrusiveness, Customization*, Interface*, Integration*, Focus work feature*, Project management* (n = 20)	User interface (ease-of-use, navigation)*, Features*, Compatibility*, Reporting*, Mobile access*, Pricing*, Customer support, Customization*, Security & privacy*, Time tracking accuracy*, Time tracking methods (manual automatic timer)*, Project management*, Team management*, Invoicing*, Time off management*, Analytics*, User permissions, Offline tracking, Ease of setup*, Work types*, Team sizes (n = 21)
Deep learning frameworks	Programming language*, Data flow / graph structure*, Ease-of-use*, Support for product*, Support for research*, Visualization*, Release date, OS platform*, Licensing*, Community support*, Growing speed (community), Used-by, Documentation*, Extensibility*, Developed-by, Flexibility*, Scalability*, Speed / efficiency*, Performance*, Popularity*, Hardware support*, Supported network architecture*, Distributed compute*, Dependency*, CUDA usage* (n = 25)	Ease of use*, Flexibility*, Performance*, Scalability*, Community support*, Compatibility (programming language, hardware, OS, libraries)*, Model accuracy (available pre-trained models)*, Hardware support*, Model interpretability (support for model understanding), Model size, Development speed, Security, Licensing*, Data preprocessing*, Model architecture*, Documentation*, Innovation (research)*, Deployment*, Debugging and profiling*, Interoperability (integrate the model with other systems and applications)* (n = 20)
Best sleeping bags	Temperature rating*, Comfort*, Weight*, Packed size*, Filling material*, Outside material*, Durability*, Price*, Easy to clean, Space inside*, Shape*, Color*, Hooded or not*, Waterproofness*, Quality of zipper*, For extreme weather*, Easy setup, Length*, Warranty* (n = 19)	Temperature rating*, Weight*, Insulation type*, Comfort*, Ventilation, Size and fit*, Packability*, Durability*, Water resistance*, Zipper location, Zipper quality*, Price*, Hood and collar design and availability*, Versatility (different environments and conditions)*, Brand reputation, Shape (mummy, rectangular, or semi-rectangular)*, Length*, Shell material*, Lining material, Warranty*, Color options* (n = 21)
Best air purifiers	Ease of use*, Coverage*, Noise*, Price*, Air quality indicator*, Speed settings*, Portability*, Design*, Smart features*, Warranty*, Customer reviews*, Customer support*, Filter type*, Sleep mode*, Remove smell*, Power consumption*, Performance*, Ozone safe*, Maintenance requirements*, Filter replacement indicator* (n = 20)	Filter type*, Coverage area*, CADR (Clean Air Delivery Rate), Fan speed settings*, Noise level*, Energy efficiency*, Additional features (such as ionizer, air quality sensors, auto mode, or remote control)*, Design*, Size, Price*, Brand reputation and customer reviews*, effectiveness (dust, pollen, smoke, pet dander, and mold spores)*, Maintenance and filter replacement*, Odor elimination*, Smart features*, Filter indicator*, Warranty*, User-friendly controls*, Portability*, Ozone emissions (high levels can be harmful to health)*, Customer support*, Sleep mode*, Child lock, Certifications (n = 24)
Best robot vacuums	Performance*, Pet hair handling*, Battery life*, Reliability*, Ease of use*, Price*, Design*, Functionality*, Dust bin size*, Customer support*, Suction power*, Ability to handle different floor types*, Noise level*, Compactness*, Navigation capabilities*, Maintenance requirements*, Auto-recharge*, Smart home compatibility*, Scheduling capabilities*, Boundary control features*, Voice control*, Effectiveness in corners and edges*, Automatic dirt disposal* (n = 23)	Cleaning performance*, Navigation and mapping*, Battery life and charging*, Noise level*, Dustbin capacity*, Smart features and connectivity*, Edge cleaning and corner reach*, Maintenance and filter replacement*, Price and value for money*, Customer reviews and ratings, Runtime and automatic docking*, Scheduling and automation*, Filtration system (the efficiency of the vacuum's filtration system in capturing small particles and allergens), Suction power*, Size and design*, Multi-floor cleaning*, Virtual walls and boundary setting*, Pet hair performance*, Durability and reliability*, Brand reputation and customer support*, Accessories and additional features*, User-friendly interface & control*, Warranty and after-sales service, Mopping capabilities, Remote control options (e.g., voice), Energy efficiency, Auto-emptying capabilities*, Integration with smart home systems* (n = 28)
Best baby strollers	Safety features*, Comfort for baby*, Durability*, Foldability*, Steer and maneuver*, Storage space*, Color options*, For twins*, Compatibility with car seats*, Customer support*, Reviews*, Ease of cleaning*, Expandability for growing child*, Suspension system*, Terrain adaptability*, Included accessories (like cup holders, trays, etc.)*, Canopy and weather protection*, Height and angle of handlebars*, Reversible seat*, Price*, Weight and size*, Lockable swivel wheels* (n = 22)	Safety (secure harness, sturdy construction, and reliable brakes)*, Comfort*, Maneuverability*, Durability*, Storage*, Folding mechanism and ability*, Weight and size*, Versatility (terrain)*, Price*, Ease of cleaning & maintenance*, Design and Style*, Suspension*, Canopy & UV and weather protection*, Reversible seat*, Adjustable handlebar (to accommodate parents of different heights and provide comfortable pushing)*, Brake & locking system (ensure the stroller stays in place when needed)*, Travel system compatibility*, Accessories*, Customer reviews and ratings*, Adjustability (to accommodate baby growth)*, Adjustable seat height (to accommodate different table heights), Warranty & customer support*, Easy assembly, Accommodation for twins* (n = 24)
Best tropical vacation spots	Weather*, Cost*, Beach quality*, Cultural experiences*, Outdoor activities*, Ease of travel*, Hotel*, Scenery*, Child-friendly*, Shopping opportunities*, Understanding of language*, Food*, LGBTQ+ friendly, Bars and clubs*, Safety* (n = 15)	Weather*, Beaches*, Activities & Adventure opportunities*, Natural beauty*, Accommodation*, Safety*, Transportation & Accessibility (Ease of travel and transportation options)*, Culture & history*, Dining options*, Cost & Value for money*, Nightlife (bars, clubs, or live music venues)*, Reviews and recommendations, Family-friendly*, Wildlife, Local hospitality, Shopping*, Sustainability (Commitment to eco-tourism practices, conservation efforts, and protection of the environment), Communication & Language barrier*, Accessibility for people with disabilities (n = 19)

**Table 6: Dataset of Groundtruth Criteria vs Selenite-retrieved Criteria. Groundtruth criteria that are retrieved by Selenite are marked with a “\*”, while Selenite-retrieved criteria that appear in the Groundtruth set are marked with a “^”.**

## C TOPICS EXPLORED IN STUDY 3

Participant	Topics		Participant	Topics
P1	Choosing a high-quality digital camera Choosing the best air purifier Picking a suitable hand truck		P5	Picking the best robot vacuum Picking a drone Choosing the best e-reader
P2	Selecting the best washing machine Choosing a reliable VPN provider Deciding on unique thank-you gifts		P6	Picking the right refrigerator Choosing the best barbecue grill Choosing the best tropical vacation location
P3	Selecting the best mattress Choosing the best city bike Choosing a birthday gift		P7	Choosing an effective diet plan Choosing a breed of dog to adopt Selecting a suitable SUV
P4	Selecting a secure password manager Choosing the best tropical vacation location Choosing a hybrid app framework		P8	Picking a reliable treadmill Picking the right wedding venue Choosing the best skiing venue

Table 7: Topics that participants in study 3 explored.

## D GPT-4 PROMPTS USED IN SELENITE

Here, we outline the techniques we employed to guide the GPT-4 model, developed by OpenAI [84], within the context of Selenite. If not explicitly stated, the temperature of the model is set to 0.3 for a balance between consistency and creativity.

If not explicitly stated, the initial [System Message]<sup>16</sup> was set as the following:

[System Message]

You are a helpful assistant that performs content analysis according to user requests. Follow the user's requirements carefully and to the letter.

### D.1 Obtaining topic from a web page

The prompt that we used to obtain a concise topic given a web page (with [title] and [content of the first few paragraphs]) is a two-step prompt:

---

[User Message] Step 1

Given the following information of an article:

Title: [title]

First few paragraphs: [content of the first few paragraphs]

What is this article about?

[Assistant Message]

This article is about ...

---

[User Message] Step 2

I want to find articles similar to this one in terms of the general topic. What should I search for? Output one search phrase (in double quotes).

---

Additionally, we set the n parameter to 10, thereby instructing GPT-4 to produce 10 simultaneous responses. Subsequently, we determined the most commonly occurring one among these 10 as the **topic** for the article. While it could be assumed that a higher value for n would result in a lengthier response time from the model, our observations indicate that such delay is practically insignificant.

### D.2 Obtaining options from a web page

The prompt that we used to obtain extract options from a given web page (with [title] and [content of the web page]) is a two-step prompt:

---

<sup>16</sup>The system message helps set the behavior of the model response, i.e., the [Assistant Message]. However, as stated by OpenAI: "... note that the system message is optional and the model's behavior without a system message is likely to be similar to using a generic message such as 'You are a helpful assistant' ..." (<https://platform.openai.com/docs/guides/gpt/chat-completions-api>)

[User Message] Step 1

Given the following information of an article:

Title: [title]

Is the article likely to be discussing one or more aspects of "one specific option" (e.g., a single javascript framework, for example, React, or a single baby stroller option, or a specific Airbnb listing) or "multiple options/topics"? Output in the following format:

Reasoning: your reasoning process.

Verdict: "one specific option / multiple options"

[Assistant Message]

Reasoning: model's reasoning process...

Verdict: "one specific option / multiple options"

[User Message] Step 2

Now, given the content of the article below, what is/are the options?

Content:

[content of the web page]

Output should be in the following format: ["option\_1", "option\_2", ...]

### D.3 Obtaining commonly-considered criteria from a web page

The prompt that we used to obtain a set of commonly considered criteria resembles something like the following (given a [topic]):

[User Message] Step 1: Ask for an initial set of criteria

What are some common aspects, criteria, or dimensions that people consider on the topic of [topic]? Note that the criteria should be \*\*most relevant to the topic\*\*, \*\*frequently considered\*\*, and can \*\*cover a broad range of perspectives\*\*. Output should be a single bulleted list in the format of:

- Criterion: short description.

Do not output anything else.

[Assistant Message]

- [Criterion 1]: [Short description]
- [Criterion 2]: [Short description]
- [Criterion 3]: [Short description]
- ...

[User Message] Step 2+: Ask for additional criteria until we get around 20.

Give me five more that are different from, more diverse than, and possibly as important as the ones listed above. Output in the same format.

### D.4 Obtaining detailed analysis of text content

The prompts that we used to obtain a detailed analysis of text content given the [text content], the list of NLI criteria, and the list of [options] on the corresponding web page is two-fold:

First, we ask GPT-4 to extract phrases from the content that describes a given criterion as well as determine each extracted phrase's sentiment with respect to the criterion:

[User Message]

Given the following \*\*content\*\* and list of \*\*criteria\*\*:

\*\*Content\*\*:

[content]

\*\*Criteria (with definitions)\*\*:

- [NLI Criterion 1]: [description]

- [NLI Criterion 2]: [description]

...

For each criterion: 1) extract \*\*every possible\*\* utterance that \*\*mentions\*\* or \*\*explicitly describes\*\* that criterion from the content 2) perform sentiment analysis to determine if the utterance is "positive", "neutral", or "negative" with respect to that criterion. Remember to use the \*\*exact same words\*\* from the content. Do not paraphrase!

Output must follow the format below:

```
## criterion_1_name
- "extracted_sentence_or_phrase_1" -> positive,
- "extracted_sentence_or_phrase_2" -> neutral,
## criterion_2_name
NONE FOUND
## criterion_3_name
- "extracted_sentence_or_phrase_1" -> neutral,
- "extracted_sentence_or_phrase_2" -> negative,
- "extracted_sentence_or_phrase_3" -> positive,
```

---

Second, we ask GPT-4 to label each extracted phrase with a possible [option] on the web page (we framed options as “subjects” of a phrase to achieve a better empirical performance):

---

[User Message]

Given the following \*\*content\*\* and the \*\*phrases\*\* extracted from the content below:

\*\*Content\*\*:

[content]

\*\*Extracted phrases\*\*:

- "extracted\_phrase\_1"

- "extracted\_phrase\_2"

- "extracted\_phrase\_3"

...

For each phrase, determine the \*\*subject\*\* of the phrase based on the \*\*content\*\*. Possible subjects are: [option\_1, option\_2, option\_3, ...] Say "N/A" if you cannot determine the subject. Output should be in the following format:

"extracted phrase 1" -> "subject" or "N/A"

"extracted phrase 2" -> "subject" or "N/A"

...

---

## E OTHER IMPLEMENTATION DETAILS

We leverage GPT-4 for several use cases in Selenite, and encountered several challenges<sup>17</sup>: First, due to the limited context window size of GPT-4 (8192 tokens or approximately 6100 English words), we occasionally need to divide the entire text content of a web page into smaller chunks and run parallel queries to extract options. As of July 2023, we don't have access to the version of GPT-4 with a 32k context window, which would significantly reduce the need for chunking and parallel queries. Second, unfortunately, there are occasions when the GPT-4 model becomes overloaded with requests or takes an exceptionally long time to respond. To mitigate these problems and provide uninterrupted user experience to Selenite users, we have employed the following two approaches: 1) *Dual API requests*: We send two identical requests using separate API keys simultaneously. We prioritize the response that returns first with valid information, indicating that it is not an error and contains the requested information from the prompt; 2) *Graceful error handling & retry*: In the event of an error, we introduce a random delay (ranging from 1 to 5 seconds) before retrying the request. We repeat this retry process for up to 5 attempts, allowing sufficient opportunity for a successful response. Note that these issues are attributable, in part, to the current limited beta status of GPT-4. Consequently, it is uncertain whether these issues will persist in the future. Nevertheless, we delve into them here to provide a comprehensive and accurate accounting of our experience interacting with the API.

To efficiently perform natural language inference (NLI) during the analysis of article content to produce per-paragraph summaries and annotations of options and criteria, we experimented with both the `roberta-large-mnli` and `bart-large-mnli` models that are fine-tuned for multi-genre natural language inference (MNLI) tasks<sup>18</sup>, and ended up using the latter for its better performance in our informal testing. In addition, we implemented a REST API service that the Chrome extension can query on demand. To decrease model inference time and ensure a smooth user experience, we ran the service on multiple Google Cloud virtual machines with NVIDIA L4 GPUs.

---

<sup>17</sup>We have documented the specific prompt designs for those tasks in section D, and only discuss a series of challenges we experienced while interacting with the GPT-4 API here in this section.

<sup>18</sup>These two models are considered to be able to achieve state-of-the-art performance as of June 2023.