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

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...
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 an overview 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. 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 avenues for finding additional information and expanding 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

---

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

---

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.
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.2 We then explored how they managed 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] 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.”4 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” (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 per-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
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.
While it is harder to push and a wee bit heavier than the smallest product, it is easier to use and earned a higher score ... ……
ManeuverabilityWheel typeDurability
ManeuverabilityWheel typeCustomer review
Price
ManeuverabilityWeight and size

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.

💡Analyze

While it is harder to push and a wee bit heavier than the smallest product, it is easier to use and earned a higher score ... ……

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 the cursor 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 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]) to first summarize an arbitrary web page given its title and initial five paragraphs (with the temperature set to 0 to minimize LLM hallucination). 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
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 1c1), 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 “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. 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]. In addition, it surpasses the limited heuristics employed in previous approaches such as Crystalline [72], which rely on page titles and HTML <div> 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].

4.2.2 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 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 performance.

Please refer to section D.3 in the Appendix for the detailed prompt design.

8While 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.

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

10The model can be accessed on-demand through a remote API service that we implemented.
Ensuring the stroller has proper safety features such as a secure harness, sturdy construction, and reliable brakes.

Having easy & smooth maneuverability, with features like swivel wheels, suspension systems, and the ability to navigate tight spaces.

Ensuring the stroller is built to last, with high-quality materials and strong construction.

Allowing for easy folding and compact storage, as well as being lightweight for convenient transportation.

Ensuring the stroller is built to last, with high-quality materials and strong construction.

Offering ample storage space for carrying essentials such as diaper bags, snacks, and personal items.

Having user-friendly features like adjustable handles, intuitive controls, and easy-to-clean fabrics.

Providing a comfortable seat with adequate padding and support for the baby, as well as adjustable recline positions.

Taking into account feedback and recommendations from other parents who have used the stroller.

Considering the affordability and value for money in relation to the features and quality of the stroller.

Having the option to face the baby towards the parent or away from the parent.

Offering the ability to attach a car seat to the stroller for convenient travel.

Checking for a warranty or guarantee that covers any potential defects or issues with the stroller.

Having a reliable brake system that is easy to engage and disengage.

Considering the reputation and reliability of the brand manufacturing the stroller.

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]).

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

<table>
<thead>
<tr>
<th>Criterion Name</th>
<th>Criterion Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety</td>
<td>Ensuring the stroller has proper safety features such as a secure harness, sturdy construction, and reliable brakes.</td>
</tr>
<tr>
<td>Maneuverability</td>
<td>Having easy &amp; smooth maneuverability, with features like swivel wheels, suspension systems, and the ability to navigate tight spaces.</td>
</tr>
<tr>
<td>Durability</td>
<td>Ensuring the stroller is built to last, with high-quality materials and strong construction.</td>
</tr>
<tr>
<td>Storage</td>
<td>Offering easy folding and compact storage, as well as being lightweight for convenient transportation.</td>
</tr>
<tr>
<td>Folding and Portability</td>
<td>Allowing for ample storage space for carrying essentials such as diaper bags, snacks, and personal items.</td>
</tr>
<tr>
<td>Versatility</td>
<td>Providing features that allow the stroller to adapt to different terrains, weather conditions, and age ranges.</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>Having user-friendly features like adjustable handles, intuitive controls, and easy-to-clean fabrics.</td>
</tr>
<tr>
<td>Ease of Cleaning</td>
<td>Ensuring the stroller is easy to clean and maintain, with removable and washable fabric components.</td>
</tr>
<tr>
<td>Adjustable Canopy</td>
<td>The stroller should have adjustable handlebars and footrests to accommodate different caregivers and growing babies.</td>
</tr>
<tr>
<td>Reversible seat</td>
<td>A large and adjustable canopy to protect the baby from the sun and other elements.</td>
</tr>
<tr>
<td>Car seats compatibility</td>
<td>Offering the ability to attach a car seat to the stroller for convenient travel.</td>
</tr>
<tr>
<td>Adjustable height</td>
<td>Allowing for adjustable handlebars to accommodate different heights of caregivers.</td>
</tr>
<tr>
<td>Easy assembly</td>
<td>Providing clear instructions and easy assembly process for the stroller.</td>
</tr>
<tr>
<td>Design and aesthetics</td>
<td>Considering the overall design and aesthetics of the stroller to match personal preferences.</td>
</tr>
<tr>
<td>Weight capacity</td>
<td>Specifying the maximum weight limit the stroller can safely carry.</td>
</tr>
<tr>
<td>Warranty</td>
<td>Checking for a warranty or guarantee that covers any potential defects or issues with the stroller.</td>
</tr>
<tr>
<td>Brand reputation</td>
<td>Considering the reputation and reliability of the brand manufacturing the stroller.</td>
</tr>
<tr>
<td>Accessories</td>
<td>Offering additional accessories such as rain covers, mosquito nets, or parent organizers for added convenience.</td>
</tr>
</tbody>
</table>

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 skipped11 (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

11The 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 as 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 while still being relevant to what users cared about. That is, we need to find a subgraph \( G' \) of size \( n \), which maximizes a weighted \((\beta > 0)\) sum of vertex weights \( \omega_V \) (relevance) and edge weights \( \omega_E \) (diversity):

\[
\arg \max_{G' \subseteq G, |G'| = n} \beta \cdot \omega_V(G') + \sum_{(u,v) \in E'} \omega_E(u,v)
\]

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_aboutCriterion)” 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].

## 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] may cause readers to make easily miss information that is indeed included in the page but not reflected 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
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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Topic-level</th>
<th>Paragraph-level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#GT (total)  #Selenite (total)</td>
<td>Precision  Recall  F1</td>
</tr>
<tr>
<td>Best washing machines</td>
<td>19  24</td>
<td>0.88  1.0  0.93</td>
</tr>
<tr>
<td>Birthday gift ideas</td>
<td>11  21</td>
<td>0.57  0.91  0.70</td>
</tr>
<tr>
<td>Best hybrid app frameworks</td>
<td>15  21</td>
<td>0.86  0.93  0.89</td>
</tr>
<tr>
<td>Best time tracking tools</td>
<td>20  21</td>
<td>0.81  0.95  0.87</td>
</tr>
<tr>
<td>Deep learning frameworks</td>
<td>25  20</td>
<td>0.80  0.84  0.82</td>
</tr>
<tr>
<td>Best sleeping bags</td>
<td>19  21</td>
<td>0.81  0.89  0.85</td>
</tr>
<tr>
<td>Best air purifiers</td>
<td>20  24</td>
<td>0.83  1.0  0.91</td>
</tr>
<tr>
<td>Best robot vacuums</td>
<td>23  28</td>
<td>0.82  1.0  0.90</td>
</tr>
<tr>
<td>Best baby strollers</td>
<td>22  24</td>
<td>0.92  1.0  0.96</td>
</tr>
<tr>
<td>Best tropical vacation spots</td>
<td>15  19</td>
<td>0.74  0.93  0.82</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>19.0  22.3</strong></td>
<td><strong>0.80  0.95  0.87</strong></td>
</tr>
</tbody>
</table>

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 frameworks, 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*. 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.

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
I enjoyed the features provided by the tool. Using this tool would make solving sensemaking problems more efficient and effective.

Temporal demand 1
Performance 3
Frustration 6

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 *.

<table>
<thead>
<tr>
<th>Question category</th>
<th>Statement</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensibility</td>
<td>I would consider my interactions with the tool to be understandable and clear.</td>
<td>6 (6.53 ± 1.10)</td>
</tr>
<tr>
<td>Learnability</td>
<td>I would consider it easy for me to learn how to use this tool.</td>
<td>7 (6.71 ± 1.04)</td>
</tr>
<tr>
<td>Enjoyability</td>
<td>I enjoyed the features provided by the tool.</td>
<td>6 (6.13 ± 1.72)</td>
</tr>
<tr>
<td>Applicability</td>
<td>Using this tool would make solving sensemaking problems more efficient and effective.</td>
<td>6 (6.28 ± 1.39)</td>
</tr>
<tr>
<td>Recommendability</td>
<td>If possible, I would recommend the tool to my friends and colleagues.</td>
<td>6 (6.23 ± 0.94)</td>
</tr>
</tbody>
</table>

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).

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.15

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 that 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).

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 “relax 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 mentions 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” (P9). 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

15To see all the topics that participants explored in this study, please refer to Table 7 in the Appendix.
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 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 grounding the reading process with domain knowledge would also work with other sources of knowledge bases as well, for example, Unakitke + 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, Ziyang Yan, Yi Zhou, Zheng Yao, Tianshi Li, Eric Yiyi Wang, Amber Horvath, Daye Nam, Nimo Ni, Meng Xia, Yongsong Kim, Andrew Kuznetsov, Hyeoums B. Kang, Wesley Hanwen Deng, and Queenie Kravitz for their extraordinary feedback and support.

REFERENCES


### A UNFAMILIAR TOPICS PARTICIPANTS EXPLORED IN THE FORMATIVE STUDY

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

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.”

![Comparison Article](https://www.babygearlab.com/topics/getting-around/best-stroller)

**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](https://www.babygearlab.com/topics/getting-around/best-stroller).
## B.2 Dataset of Groundtruth Criteria vs Selenite-retrieved Criteria

<table>
<thead>
<tr>
<th>Topic</th>
<th>Groundtruth Criteria</th>
<th>Selenite-retrieved Criteria</th>
</tr>
</thead>
</table>

### 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

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

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] was set as the following:

```
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:

```
[User Message] Step 1
Given the following information of a page:
Title: [title]
First few paragraphs: [content of the web page]

What options does this page provide?

[Assistant Message]
This page provides ...

[User Message] Step 2
I want to find pages similar to this one in terms of the general topic. What should I search for? Output one search phrase (in double quotes).
```

\(^{10}\)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 behaviour 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-apis)
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"

Reasoning: model's reasoning process...
Verdict: "one specific option / multiple options"

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]):

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.

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

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:
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\textsuperscript{17}: 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) \textit{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) \textit{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 \texttt{roberta-large-mnli} and \texttt{bart-large-mnli} models that are fine-tuned for multi-genre natural language inference (MNLI) tasks\textsuperscript{18}, 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.

\textsuperscript{17}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.

\textsuperscript{18}These two models are considered to be able to achieve state-of-the-art performance as of June 2023.