AbstractExplorer: Leveraging Structure-Mapping Theory to Enhance Comparative Close Reading at Scale

Ziwei Gu
Harvard University
Cambridge, Massachusetts, USA
ziweigu@g.harvard.edu

Joyce Zhou
Cornell University
Ithaca, New York, USA
jz549@cornell.edu

Ning-Er (Nina) Lei
Harvard University
Cambridge, Massachusetts, USA
nlei@college.harvard.edu

Jonathan K. Kummerfeld
The University of Sydney
Sydney, NSW, Australia
jonathan.kummerfeld@sydney.edu.au

Mahmood Jasim
Louisiana State University
Baton Rouge, Louisiana, USA
mjasim@lsu.edu

Narges Mahyar
University of Massachusetts Amherst
Amherst, Massachusetts, USA
nmahyar@cs.umass.edu

Elena L. Glassman
Harvard University
Cambridge, Massachusetts, USA
glassman@seas.harvard.edu

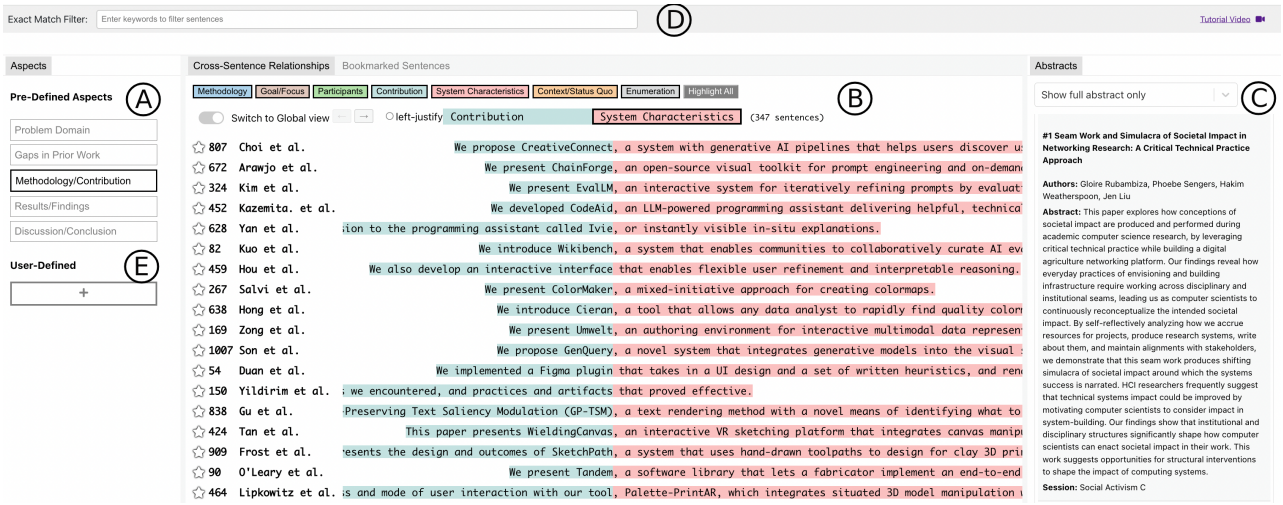


Figure 1: The AbstractExplorer interface, rendering over a thousand CHI2024 paper abstracts. (A) Users can select one of five pre-defined aspects of abstracts to view at a time; *Methodology/Contribution* is currently selected. (B) Sentences in each abstract that reflect the selected aspect are shown; users can skim or read laterally [67] across many abstracts, and engage in comparative close reading at scale. Sentences are segmented into grammatical chunks, categorized into one of the pre-defined roles listed at the top of the (B) pane, and highlighted by that role's assigned color. The sentences are ordered by 'structure': within each selected aspect, the most common structure is initially shown by default, e.g., `Contribution` then `System Characteristics` above, but others can be selected. (C) Clicking on a sentence scrolls the full abstract it was extracted from into view in the right sidebar. (D) Users can use the exact match filter to hide all but a more narrowly scoped set of abstracts, e.g., to only those that mention 'VR'. (E) Users can author their own aspects as well, using natural language.

Abstract

Individual flagship conferences today can have over a thousand papers; even reading just the abstract of every paper at the latest relevant conference to keep up with the research is time and memory prohibitive. Previous visualizations in this domain have ubiquitously followed Shneiderman's Visual Information-Seeking Mantra, with details available on demand. However, recently in other domains, system designers have leveraged Structure-Mapping

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747773

Theory (SMT) to facilitate seeing both the overview and the details at the same time, facilitating abstraction without losing context. We compose and evaluate a system, called AbstractExplorer, with analogous SMT-derived characteristics for the domain of scientific abstract corpus familiarization. AbstractExplorer has a unique combination of LLM-powered (1) faceted comparative close reading with (2) role highlighting enhanced by (3) structure-based ordering and (4) alignment. An ablation study (N=24) validated that these features work best together. A summative study (N=16) describes how these features support users in familiarizing themselves with a corpus of paper abstracts from a single large conference with over 1000 papers.

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**; **Interactive systems and tools**; *Visualization theory, concepts and paradigms.*

Keywords

Structure-Mapping Theory, text, scientific abstracts, reading, sensemaking at scale

ACM Reference Format:

Ziwei Gu, Joyce Zhou, Ning-Er (Nina) Lei, Jonathan K. Kummerfeld, Mahmood Jasim, Narges Mahyar, and Elena L. Glassman. 2025. AbstractExplorer: Leveraging Structure-Mapping Theory to Enhance Comparative Close Reading at Scale. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3746059.3747773

1 Introduction

The only existing ground-truth mechanism for reasoning about a corpus of documents is to read every document in the corpus. Unaugmented reading is a serial process, where connections within and across documents are made because the reader remembers text they have previously read in the corpus and, based on that memory, recognizes a relationship to the current portion of text before them.

Many existing approaches to augmenting humans' reasoning about document corpora are lossy, and therefore risk omitting important context. Non-linguistic lossy representations of documents are typically designed to preserve what the system creators believe is most useful for the intended user task(s), e.g., a network of document-representing dots that preserve citation relationships across documents while omitting most other content. Zooming into the otherwise hidden details often means narrowing one's view to a small subset of the corpus at a time, replacing a focus on cross-document relationships with a focus on individual documents.

Lossy linguistic representations can introduce semantic ambiguities and misrepresentations [24], such as reducing a definitive factual statement to a topic. Lossy approaches limit how many unanticipated user questions the system can help answer, and make it harder for users to recognize objectively wrong or contextually inappropriate choices made about the document or its representation on their behalf [24].

Recent prior work has shown that it is possible to help people read and reason about a corpus of short documents without employing lossy document representations. For example, for collections of code examples written with similar purposes but using different libraries, ParaLib [69] used color-coordinated role highlights to reveal cross-example commonalities and distinctions. The Positional Diction Clustering (PDC) algorithm identified analogous sentences across many LLM responses, which were reified both as color-coordinated cross-document analogous text highlighting (like ParaLib) and in a novel 'interleaved' view where analogous sentences across documents were rendered in adjacent rows to enable more easy comparison [18]. These examples of text-centric lossless techniques do not abstract away or summarize; they strategically re-organize and re-render the existing text to help enhance readers' own perceptual cognition, informed by Structural Mapping Theory (SMT) [17].[1] The human perceptual, comparative mental machinery that SMT describes is part of what enables humans to form more abstract structured mental models from concrete examples, among other critical knowledge tasks.

This SMT-informed approach, which AbstractExplorer shares, tries to give this mental machinery "a leg up," letting users perhaps skip some steps by accepting reified cross-document relationships identified by the computer. The revealed variation within these analogous cross-document relationships can *invite* the user's engagement. This is the essence of comparative close reading, a dialectical activity [73] that requires repeated deep engagement with the texts to reveal new insights.

Lossless SMT-informed techniques have yet to be brought to bear in the context of researchers familiarizing themselves with a corpus of existing literature, e.g., all (> 1000) paper abstracts at recent CHI. Most tools assist researchers in this pursuit by helping them narrow their attention to a manageable set of papers they can sit down with and serially engage with. These tools often rely on lossy representations of the entire corpus and give researchers search affordances to navigate serially through papers of potential interest in a more informed way.

AbstractExplorer instantiates new minimally lossy[2] SMT-informed techniques for skimming, reading, and reasoning about a corpus of similarly structured short documents: phrase-level role classification that drives sentence ordering, highlighting, and spatial alignment. We demonstrate these features' utility in the context of helping researchers' skim and read closely and laterally [67] across a corpus of scientific abstracts.

Three studies inform and validate AbstractExplorer's design: First, a formative study (Section 3) suggested unmet needs and interest in our approach to supporting cross-document reasoning. Second, an ablation study with eye-tracking (Section 5) revealed that the three key features of AbstractExplorer's central cross-sentence relationships pane—sentence order, role-coordinated highlighting, and alignment—work best in concert, not alone. Finally, a summative study (Section 6) describes how researchers used AbstractExplorer to familiarize themselves with a corpus of ~1000

[1] Structural Mapping Theory (SMT) is a long-standing well-vetted theory from Cognitive Science that describes how humans attend to and try to compare objects by finding mental representations of them that can be structurally mapped to each other (analogies).

[2] Just the order of sentences within their respective abstracts is abstracted away.

CHI paper abstracts—reading across a larger and more diverse collection of abstracts and more easily discerning relationships and distributions across prior work. In summary, we contribute:

- Novel SMT theory-informed text analysis and rendering techniques for enabling cross-document skimming and comparative close reading at scale
- AbstractExplorer, which instantiates these techniques for familiarizing oneself with a corpus of ~1000 CHI paper abstracts.
- Three studies informing and evalutaing the benefits, challenges, and interactions between these techniques.

2 Related Work

AbstractExplorer extends prior work on tools that support reading and sensemaking at scale as well as text alignment.

2.1 Tools for Close Reading

Close reading refers to the "conscientious analyzation and interpretation" of text [32], while *distant reading* [49] avoids serially reading documents by looking at alternative summaries such as counting, graphing, and mapping of text [61]. Distant reading tools often employ very lossy approaches, e.g., visualizations of documents that abstract text into topics, such as Hierarchical Topic Maps [63], and do not preserve entire sentences, such as displaying word pairs in a word cloud [11, 70]. Distant reading tools do not support comparing documents at the textual level one needs during close reading. They are also not necessarily superior in terms of cognitive load or more 'unbiased' in their presentation of data, either: large network graphs and scatterplots are known to impose significant cognitive demands [39, 71] and can introduce perceptual biases [66].

Close reading is an important yet cognitively demanding task in scholarly activities [60] and beyond [55]. Close reading requires reasoning about context; for example, ClioQuery [26] assists historians with investigating queries in context through linked views and text highlighting. While not specifically designed for close reading, many tools have been designed to support reading activities within or anchored by a single document, e.g., supporting comprehension, information foraging, and reading efficiency. For example, GP-TSM [24] helps readers read more efficiently by modulating text saliency while preserving grammar. VarifocalReader [36] supports skimming by presenting abstract summaries alongside the source document, with machine-learned annotations highlighting key sentence segments in different colors. Scim [15] helps readers skim academic papers by using colored highlights to guide attention to predefined content types, such as "Objective" and "Method." Qlarify [14] allows scholars to specify additional information needs while reading an abstract, dynamically expanding it with relevant content from the full paper. While still anchored on a single document, the Semantic Reader project [43] supports features that bring information from related papers into the focal paper's reading environment. For example, Relatedly [54], part of the Semantic Reader project, highlights unexplored dissimilar information in related work sections of unread papers while lowlighting previously seen information. As such, while the Semantic Reader can reason over document collections, we still consider it a *document-centric* [23] reading tool.

When working with document corpora, these prior reading systems leave users to manually organize documents using their limited working memory; AbstractExplorer instead lays out the contents of the corpus in a way that, given its cognitive-theory-informed design, may work with humans' limited working memory while reading.

2.2 Text Alignment

Text alignment refers to the process of finding correspondences—similar and diverging patterns—among two or more pieces of text [72]. Methods often segment text into smaller, comparable units and align segments to highlight shared patterns and individual divergences. The alignment step is often cast as a sequence alignment problem, using methods such as edit distances [10] or common grammatical pattern identification [59]. Cross-sentence relationships are often rendered using matrices [48, 76], trees [65], and graphs [25]. However, these approaches can only handle sentences with at least some closely matched structure and overlapping diction.

Some work has explored aligning larger pieces of text, including websites [3], books [56], and passages [13]. However, given that the typical purpose of this work is for machine-powered functions, e.g., obtaining parallel corpora for machine translation or plagiarism detection, renderings for human consumption are not a focal point.

Some alignment strategies aim to scale up the alignment process to more (but not necessarily long) texts. For instance, Tempura [68] employs "structural templates" based on linguistic features to organize short diverse search queries into representative groups. Gero et al. [18] addresses the challenge of aligning a large set of similar yet varying multi-sentence LLM responses with Positional Diction Clustering (PDC), an algorithm that identifies analogous sentences across documents based on shared diction and position within their respective documents. However, the PDC approach ignores *purpose* at both the sentence and sub-sentence level; by leveraging purpose, AbstractExplorer, can handle more diverse short documents, i.e., abstracts written by different authors about very different topics.

In summary, while scientific abstracts, especially within the same community, lack the linguistic consistency required by previous methods, they make up for it in the common purposes that sentences and phrases within sentences fulfill. AbstractExplorer reframes text alignment as a semantic role labeling task, using spatial alignment, color, and proximity to help humans perceive simultaneous text alignment across hundreds of diverse abstracts.

2.3 Text-Centric Tools for Sensemaking over Document Corpora via Shared Structure

Prior systems have employed structure to facilitate *inductive* sensemaking across various domains. For example, Hope et al. [29] segments product descriptions into fine-grained functional aspects to support analogical reasoning. In the space of text-centric [57] sensemaking tools for text and code document corpora, these systems include WordSeer [50, 51, 59], OverCode [20], Foobaz [19], Examplore [22], SOLVENT [6], Paralib [69], Positional Diction Clustering (PDC) [18], and CorpusStudio [9]. Some of these systems exploit shared structure by creating templates, skeletons, or schemas to reveal variation across corresponding parts.

Like many of these prior systems, AbstractExplorer is a text-centric system that exploits shared structure to support lateral reading across short document corpora at scale. Unlike prior systems, it leverages common structures in *purpose*, not lexical or grammatical features, at both the sentence and phrasal level in natural language documents. AbstractExplorer specifically leverages the shared structure in scientific abstracts to facilitate abstract corpus skimming and comparative reading at scale.

This is feasible for scientific abstracts because, as prior work has identified, there are common structures in academic writing [40, 41, 52], particularly in abstracts [12, 30]. Datasets of paper abstracts have been annotated with research aspects by experts [35] and crowds of non-experts [30]. For instance, a large-scale study on medical journal abstracts [12] found that the predominant structure is *Background*, *Methods*, *Results*, and *Conclusions*; similarly, for a different abstract corpus, Chan et al. [6] used *Background*, *Purpose*, *Mechanism*, and *Findings* when building Solvent to support users finding analogies between research papers.

2.4 Theories Operationalized for Supporting Inductive Sensemaking

Most sensemaking systems for corpora are built for serial exploration and reading of document corpora, e.g., by improving information scent as in Paper Forager [46]'s scaled down paper images or the mixed-initiaive Serendyze [33]'s recommendations of what to read next, not close reading and comparison at scale.

But many of the systems in Section 2.3 explicitly reference, as design inspiration or justification, two theories of human cognition, i.e., Variation Theory [45] and/or Structural Mapping Theory (SMT) [17].[3] SMT provides a framework for understanding how humans compare two or more objects by finding common structural alignments between objects. SMT posits that visual alignment helps people perceive relational similarities and differences more clearly, thereby improving their ability to make meaningful comparisons and understand underlying patterns [28, 38, 47]. The prior SMT-informed tools in Section 2.3 for both code and natural language corpora suggest that the cognitive process of comparing texts may be no exception to the cognitive processes SMT predicts.

3 Formative Interview Study

In order to determine (1) the context in which we might offer novel views of scientific abstracts and (2) the intelligibility of various novel prototype designs for reifying cross-abstract relationships, we conducted a formative interview study with 12 active researchers (see Appendix A for participant information). The interview sessions were divided into two parts: an open-ended semi-structured interview about their backgrounds and practices, followed by feedback on a range of mock-ups, including novel reified relationships between analogous sentences in different abstracts (Figure 2). Sessions, which were held on Zoom, lasted 55 minutes on average. Participants were compensated with $15 USD.

[3]SMT is sometimes referred to by alternative names, such as Analogical Learning Theory.

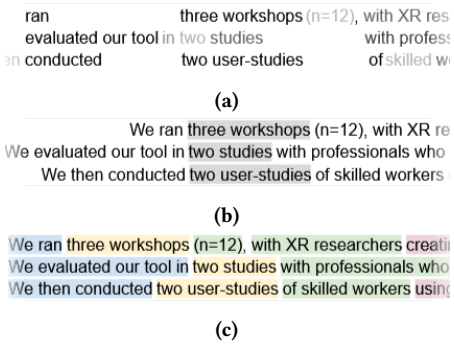


Figure 2: Examples of mock-ups of cross-document relationship visualizations created for formative study participants, which remix ideas inspired by GP-TSM [24], Examplore [22] and ParaLib [69].

3.1 Procedure

In the first part of the session, we asked participants about their strategies for selecting publication venues for their manuscript submissions, how they identify and synthesize information from venues, their approaches to writing manuscripts, and finally, the technology they have used to help with these processes, current technology shortcomings, and ideas for addressing these challenges.

In the second part of the session, we provided participants with mock-ups of possible reifications of cross-document relationships that might help them synthesize information across abstracts. These mock-ups were inspired by prior work. For example, Figure 2(a) shows analogous adjacent sentences rendered with the grayscale skimming support of GP-TSM [24] and the alignment of analogous alternatives inspired by Examplore [22], a code corpora sensemaking tool. Figure 2(b) shows the alignment of a single set of analogous alternatives across the sentences, and Figure 2(c) shows role-reifying color-coordinated highlights inspired by ParaLib [69], another code corpora sensemaking tool. We also presented participants with mock-ups of potential approaches to identifying relevant information types, including custom search or LLM-powered text annotation. See Appendix C for the complete set of mock-ups.

We used these mock-ups as design probes [31] to inspire ideation and elicit creative responses. Specifically, we asked participants to compare and contrast alternative mock-ups and reflect on how they could be used or improved to support their known or emerging synthesis and information-foraging goals. After the interviews, we analyzed the data using the process described in Appendix B

3.2 Key Findings

3.2.1 Existing Challenges to Sensemaking. Participants' approaches to reading and gathering information from papers varied significantly, and despite leveraging some tools, these tools were often less helpful than hoped for. Participants frequently mentioned difficulties with search, e.g., needing to try searching *"at least 10 times with different keywords to find the initial set of relevant papers"* (P8). Some participants also mentioned using generative AI tools to *"help [them] figure out ... the key points of a specific piece"* (P12) but may find it unhelpful if *"they often miss some of the key information"* (P8).

In contrast, many participants described their paper comparison process to be manual, with minimal tool usage.

3.2.2 Feedback on Mock-ups. Participants commented that alignment and/or highlighting were useful for identifying and comparing information: *"this... helps to recognize the differences or similarities"* (P7). However, some thought that this would require some amount of trust in the algorithm performing the highlighting, especially if judgments of importance were being delegated to the system designer or an AI: *"I don't know if I would agree with whoever's classification of what's important"* (P2). Visually, participants commented that color highlights made it easier for them to read "important" parts of the sentences. However, they were divided on alignment formatting: while some participants disliked seeing extra spacing necessary to create vertical alignments because it *"[causes] a little bit of annoyance... why is the gap there?"* (P3), others said that it was helpful if the *"user wants to look at [a particular common component]... instead of looking at the whole sentence"* (P4).

Participants particularly liked a presented mock-up that let them build and review custom search queries by highlighting desired portions of text in one or more papers as example-based specifications for retrieving analogous text from other papers, describing it as *"a smarter Ctrl-F"* (P2) or a way to augment keyword-based search with richer semantic context. One participant described that this *"would be useful just because it gets at what the researcher thinks they want to know ... and it highlights that information for them"* (P12). However, some raised concerns about implementing this "query from a text highlight" affordance using an LLM: *"if I cannot verify [it] myself, there's no way for me to know what I have missed"* (P5). Similar to the discussion about cross-document relationships, participants generally appreciated color highlights.

4 AbstractExplorer

Given the challenges and opportunities of supporting skimming and comparing papers at scale revealed in the formative study, AbstractExplorer is designed to help researchers (1) skim, read, and better familiarize themselves with the contents and composition style of a large corpus of abstracts and (2) reason about cross-paper relationships at scale without abstracting away the author-written sentences about their own work. To do this, AbstractExplorer is designed to support close-reading purpose-defined *slices* through a large collection—1,057 in our studies—of paper abstracts.

4.1 System Components

4.1.1 Slicing abstracts for sentence-level aspect reading and comparison (same-role sentences). To keep details at the forefront of the interface, we designed a mechanism to slice abstracts for viewing them from specific angles, allowing for comparative close reading at scale at the *sentence* level. We chose the *sentence* as our unit for cross-document alignment because: (1) it preserves complete propositional content (unlike phrases or words), (2) maintains grammatical coherence when isolated (unlike arbitrary text spans), and (3) serves as the minimal self-contained unit where aspects can be meaningfully compared. "Aspects" are either pre-defined or user-defined on the fly; they are collections of sentences across abstracts that serve the same or similar purposes within their abstract.

Pre-defined aspects. AbstractExplorer classifies sentences into five pre-defined aspects common in CHI abstracts: *Problem Domain*, *Gaps in Prior Work*, *Methodology/Contribution*, *Results/Findings*, and *Discussion/Conclusion*. These were developed to suit the context of CHI based on the authors' manual annotation of ~100 CHI abstracts.

Viewing one aspect at a time enables users to closely read and compare just the analogous sentences of abstracts, which may be cognitively easier than the comparative close reading of many abstracts in their entirety, especially if cross-sentence relationships are pre-computed and reified in the interface. Clicking an aspect in the list (Figure 1A) displays corresponding sentences from all abstracts in the "Cross-Sentence Relationship" panel (Figure 1B). Each sentence is displayed with a matching origin paper ID for provenance and paper author information for context.

User-defined (custom) aspects. Motivated by formative study feedback on alternative search and document grouping, AbstractExplorer allows users to define custom aspects in addition to the five pre-defined ones. To create a custom aspect, users provide a name, description, and an optional exact-match filter (like Figure 1D) to limit results to relevant abstracts. The interface then displays matching sentences in the "Cross-Sentence Relationship" panel (Figure 1B), highlighting only the chunk most relevant to the aspect definition, as determined by the system backend (e.g., Figure 4). See Section 4.3 for this and other implementation details.

4.1.2 Grammar-preserving sentence segmentation and role highlighting. Inspired by GP-TSM [24], AbstractExplorer first segments sentences into grammar-preserving chunks—segments that respect grammatical boundaries, i.e., an LLM judges that the sentence can be truncated at that chunk boundary without breaking the grammatical integrity of the preceding text. Each chunk is then classified by an LLM as having one of nine pre-defined roles, each of which has its own assigned color.

To define these roles, we used a human-LLM collaboration approach. An LLM produced initial annotations, which we iteratively refined via comparison with the authors' manual annotations on our sample of ~100 CHI 2024 abstracts. Formative study feedback supported the use of role-related color highlights, so we planned to visually indicate the role of each chunk with a unique color. We balanced role specificity with the possible visual indistinguishability of too many roles if each is assigned a unique color, ultimately defining nine roles: `Context/Status Quo`, `Challenge/Problem`, `Contribution`, `Goal/Focus`, `Methodology`, `Participants`, `System Characteristics`, `Finding`, and `Enumeration` for the CHI 2024 corpus. An LLM was used to classify sentence chunks into each of these roles. We adopt the color palette of Tableau 10, which was carefully designed to be clearly distinguishable.[4]

Each role label has a corresponding unique color shown at the top of the "Cross-Sentence Relationship" panel (Figure 1B and Figure 3B). By default, all highlights are turned on. Users can toggle individual role highlights by clicking on the corresponding role label or toggle all highlights via the "Highlight All" button. These role-based color highlights enable quick identification of analogous chunks and visual pattern matching over *sequences* of chunks across sentences.

[4] https://www.tableau.com/blog/colors-upgrade-tableau-10-56782

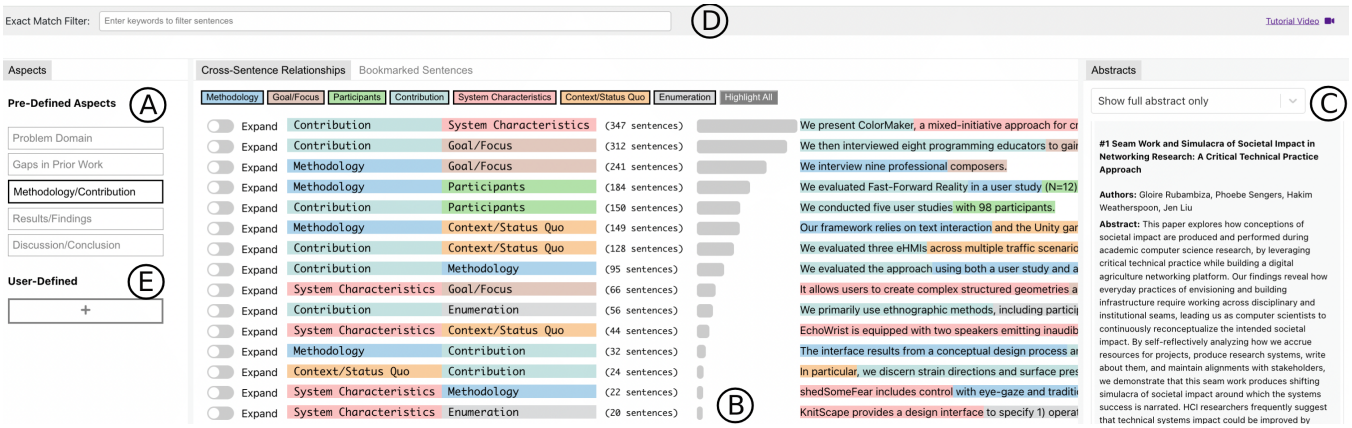


Figure 3: The global view of the AbstractExplorer interface. (A), (C), (D), and (E) are the same as Figure 1. (B) currently shows the distribution over common sentence structures in the selected aspect of the corpus, along with an example for each structure. Users can click the "Expand" toggle button to see the individual sentences represented by each histogram bar.

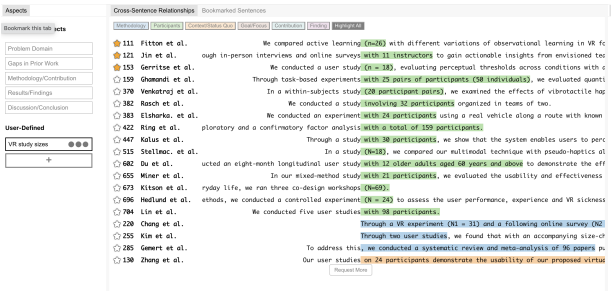


Figure 4: A user-defined aspect showing results of "user study sample sizes" (user-provided description) in "VR" (user-provided filter) papers. The most relevant chunks that contain information related to "user study sample sizes" in each sentence are automatically highlighted and used to vertically align their containing sentences.

4.1.3 Organizing sentences into structure groups. We consider common sequences of chunk roles to be *alignable* structures that could be used to support users in identifying structural similarities and differences across sentences in different abstracts, in line with Structure-Mapping Theory [17]. In SMT terminology, rendering and arranging according to corresponding chunks reify "commonalities in structure," while variation within corresponding chunks are "alignable differences" that users are predicted to notice.

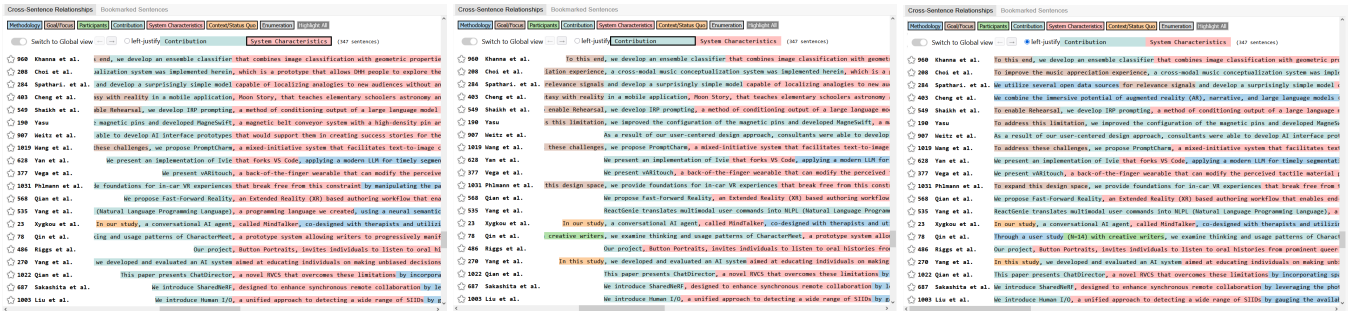
In AbstractExplorer, sentences are grouped by this definition of sentence structure. For example, Figure 5 displays sentences that all start with a description of the paper's `Contribution`, followed by `System Characteristics`. Likewise, Figure 28 in Appendix F shows all the sentences that contain a description of a study `Methodology`, followed by information about the `Participants` who are involved in the study. To navigate to a group of sentences within the same aspect that share a different structure, users can either use the left and right buttons at the top of the Cross-Sentence Relationships pane to step to the next most or next least common structure, or through the "global view" described next.

Structure Global View. In this view (Figure 3B), the sentences with the most common sentence structure are listed together first, followed by the sentences with the next most common structure, and so on. The total number of sentences with each structure in the (possibly filtered) corpus is shown in parentheses and visualized as a histogram. An example sentence is also shown alongside each structure. This allows users to first understand the different structure patterns and their commonality, before diving into close reading at scale of the sentences that share a particular structure by clicking any of the "Expand" toggles.

4.1.4 Ordering and Alignment of Sentences within Structure Groups. Structural mappings between objects are part of the cognitive process of comparison according to the Structure-Mapping Theory [17], and juxtaposition can facilitate humans in recognizing particular possible structural mappings between objects [75]. We design and implement two types of juxtapositions:

Within-structure ordering. The structure groups are defined by tuples of chunk roles, e.g., (`Contribution`, `Participants`), but within the group they may have longer common sequences of chunk roles. AbstractExplorer *orders* sentences within each structure group based on the sequential pattern of chunk roles (*vertical juxtapositions*). Sentences are recursively grouped by sequences of three chunk roles, with groups ordered by decreasing size. Within each of those groups, sentences are arranged by increasing length. This ordering prioritizes dominant structural patterns (largest groups first) while exposing fine-grained variations (via length-sorted triplets), mirroring how humans compare sentences, if SMT is an accurate description in this domain of comparative close reading.

Within-structure alignment. AbstractExplorer also *aligns* the sentences in three different ways, as illustrated in Figure 5: vertical alignment by the middle of the structure tuple (second element), vertical alignment by the left of the structure tuple (first element), and left-justified alignment (*horizontal juxtapositions*). By default,



(a) (b) (c)

Figure 5: Sentences that share the `Contribution` + `System Characteristics` role structure as viewed in the Cross-Sentence Relationship pane: (a) Vertically aligned by the boundary between the chunks with the two respective roles. (b) Aligned by the chunk with the first role. (c) Aligned by the beginning of the sentence.

sentences are vertically aligned by the middle of their shared structure tuple, but users can freely switch between the three alignment options using the button group atop the Cross-Sentence Relationship pane (Figure 1B).[5] For instance, when examining sentences with the `Contribution` + `System Characteristics` structure, users may choose to align by middle of the structure, starting with `System Characteristics`, to easily focus on the systems being built as they form a contiguous visual block (Figure 5a). Within the same aspect, when switching to examine sentences with another structure, e.g., the `Methodology` + `Participants` structure, they may choose to align by the left of the shared structure, starting with `Methodology`, to focus on the methodology being used (Figure 28 in Appendix F). These alignment options are intended to enable users to more easily read analogous chunks across sentences from different abstracts, ignoring details serving other roles within the sentence. This may effectively decrease context switching and lead to more robust mental models without requiring more cognitive load. If users prefer, sentences can be toggled to a left-justified view to facilitate conventional reading and skimming. Together, the vertical and horizontal juxtapositions are designed to help users identify both high-level commonalities and nuanced variations across structurally similar sentences.

4.1.5 *Additional Features for Context and Familiarity.*

Abstract and TLDR panel. Sentence-level reading may be new to readers. To allow users to contextualize individual sentences within their respective abstracts, we link the Cross-Sentence Relationship and Abstract panels (Figure 1B and C): when users click on any sentence in the Cross-Sentence Relationships pane (Figure 1B), the corresponding full abstract is automatically highlighted and scrolled into view in the Abstracts panel (Figure 1C), offering additional context when needed. It also includes paper metadata such as the full author list and the session name. The Abstracts panel can be customized by users to display the full abstract text, an abstract "TLDR" (a shorter abstractive summary generated by an LLM), or both at the same time.

[5]Only the left-justified alignment option appears as a radio button, the other two options are embedded into the corresponding structure labels.

Keyword filtering (search). Filtering (Figure 1D) enables users to narrow their focus to a subset of the corpus while still benefiting from features that help them recognize cross-sentence relationships within the remaining abstracts. Users can enter a search term into the search bar and only papers that include that exact term will appear in the Cross-Sentence Relationships pane (Figure 1B). Users can remove the filter by deleting text from the search bar.

Sentence bookmarking. Sentence bookmarking helps users keep track of papers to revisit later. When users click on a bookmark icon to the left of any specific sentence in the Cross-Sentences Relationships Pane (Figure 1B), that sentence is added to a bookmark list that can be viewed in the Bookmarked Sentences alternate pane. From this pane, users can toggle and view highlighted sentence chunks, click to scroll to the relevant abstract for each sentence, or remove bookmarks from the list.

4.2 User Scenario

Alice wants to learn more about the papers published in CHI 2024 and decides to use AbstractExplorer to explore them. Upon opening AbstractExplorer, Alice sees the Methodology/Contribution aspect selected by default and a list of `Contribution` + `System Characteristics` sentences—the most common sentence pattern of that aspect. After scanning a few sentences, Alice realizes that `Contribution` mostly consists of system names, while the variation lies in `System Characteristics`. She shifts her focus to `System Characteristics`, quickly skimming the vertically aligned list to discover a wide variety of systems. Intrigued by a system involving 3D-printable ceramic materials, Alice clicks on the sentence to view more details about the paper in the Abstract panel, gaining insight into a previously unfamiliar area.

Alice then navigates to a different group of sentences about evaluation methodologies. She becomes curious about the number of participants typically involved in CHI studies. She selects the `Contribution` + `Participants` group to explore the distribution of participants across studies. The green spans representing `Participants` make it easy for her to locate mentions of participant information in different abstracts. Alice notices a wide range in participant numbers, with some studies involving as many as

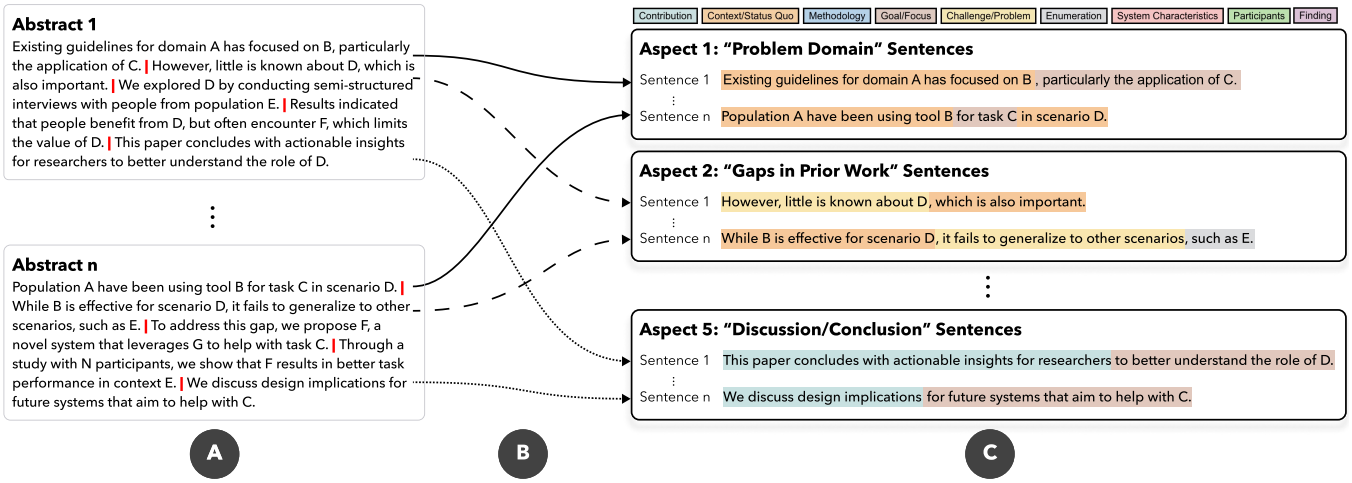


Figure 6: Workflow for automatic sentence- and chunk-level labeling in AbstractExplorer. Two hypothetical abstracts are used to illustrate the workflow. (A) We collect paper abstracts and segment each into individual sentences. (B) An LLM classifies every sentence into one of five pre-defined aspects. Together, (A) and (B) constitute Stage 1: Sentence Segmentation & Categorization. (C) Each sentence is segmented into grammar-preserving chunks (Stage 2: Sub-sentence Segmentation). Every chunk is then assigned to one of nine pre-defined functional roles and color-highlighted accordingly (Stage 3: Chunk Role Annotation).

1,500 participants. However, by looking at `Methodology` spans that identify study types, she observes that these large numbers are primarily from survey-based research. She then refines her working set of abstracts by typing "qualitative" into the search bar, filtering the underlying abstracts accordingly, and notes that most qualitative studies involve around 12 participants.

In addition to participant numbers, Alice also explores study populations by creating a custom aspect, which she describes as "description of qualitative study population," with a filter for "health" papers. The interface returns a dozen highly relevant papers, with key descriptors of study populations—such as queer women, visually impaired developers, and older adults involved in care provision—automatically highlighted for quick identification.

Finally, Alice explores research involving large language models (LLMs). She types "LLM" into the search bar and selects "Gaps in Prior Work" from the predefined aspects panel. The largest structure group is `Challenge/Problem` + `Contribution`, and Alice skims these sentences with a focus on `Challenge/Problem`, quickly identifying recurring themes such as addressing hallucination risks, improving prompting, and examining the impact of LLMs on marginalized groups. This exploration gives her a clearer view of the key challenges and research directions related to LLMs discussed at CHI 2024.

4.3 Implementation Details

The AbstractExplorer interface is a React app loaded with CHI 2024 abstract data from the CHI 2024 Papers Explorer's open-source repository.[6] The dataset consists of paper abstracts and metadata including title, author names, and session. We select only full papers from the dataset, which results in 1057 paper abstracts.

We process this data in a three-stage pipeline (Figure 6). In the first stage, Sentence Segmentation and Categorization, abstracts are split into individual sentences using the NLTK package, and each sentence is classified into one of the five pre-defined aspects as listed in Section 4.1.1. Classification is performed by prompting an LLM (see prompt used in Appendix D.1) with the sentence and its full abstract. Note that an abstract may contain multiple sentences along the same aspect.

Then, we segment sentences within each aspect into grammar-preserving chunks (see prompt used in Appendix D.2). This results in grammatically coherent chunks that are the basis of structure patterns. After identifying chunk boundaries, we again prompt an LLM to generate labels for chunks in a human-in-the-loop approach: starting from an initial set of labels for chunk roles, when a new label is generated, a researcher from the research team examines the new label and merges it with existing labels if appropriate, controlling for the total number of labels.

After obtaining an expanded set of high-level chunk labels, we assign them to each of the sentence chunks by using LLMs in a multi-class classification few-shot learning task, with the initial labels and assignment as examples (see prompt used in Appendix D.3).

All the chunks and corresponding labels are pre-computed and stored as JSON files, ensuring responsiveness and low latency of the web application. All keyword searches are computed dynamically in the React app and have low latency.

Custom aspects are generated dynamically via API calls to a FastAPI back-end, which prompts an LLM to check whether each sentence in the filtered subset matches the aspect description—either in terms of overall content or a matching token—and extracts the most relevant chunk of that sentence to highlight (see prompts used in Appendix D.4). The front-end React app allows users to view partially loaded custom aspects while they are being generated.

[6] https://observablehq.com/@john-guerra/chi2024-papers

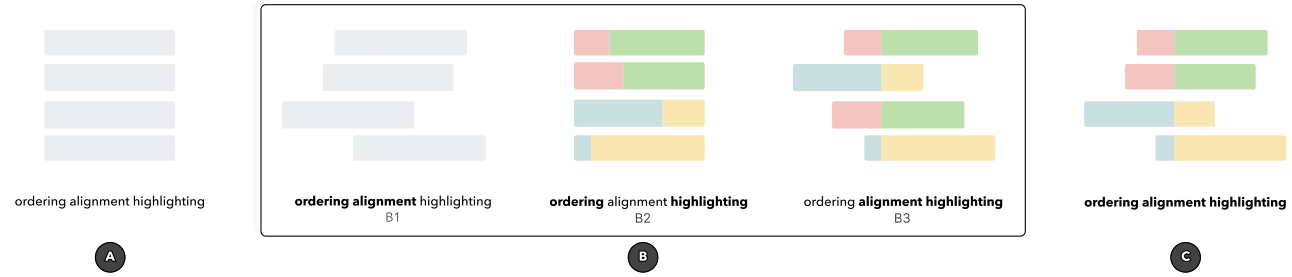


Figure 7: Conditions evaluated in the ablation study. Bolded feature names are enabled; regular-weight feature names are disabled. All participants saw sentences in conditions (A) and (C) and one condition among (B).

	Preferred				Least Cognitive Effort			
Interface	Count	Median	Mean	STD	Count	Median	Mean	STD
Baseline (A)	6	5.75	5.72	0.46	11	5.67	5.73	0.42
Without Highlighting (B1)	1	5.17	5.17	—	1	5.50	5.50	—
Without Alignment (B2)	5	5.33	5.37	0.22	0	—	—	—
Without Ordering (B3)	1	4.67	4.67	—	1	4.50	4.50	—
All-three-features (C)	11	5.50	5.18	1.10	11	5.33	5.00	0.97

Table 1: For each condition, we counted the number of participants who preferred it most, and then show the distribution of participants' NFC scores for that group. We do the same for reported lowest cognitive load.

5 Ablation Study

In this study, we allowed participants to experience views of same-aspect sentences (Section 4.1.1) with different combinations of highlighting, ordering, and alignment (as described in Section 4.1.2 and Section 4.1.4) enabled or not, in order to understand which and/or what combinations most effectively supported users' ability to skim and read laterally across documents.

Since reading is cognitively effortful, we consider how a reader's Need for Cognition (NFC) [4, 5]—defined as a personality trait that reflects one's tendency to rely on quick heuristics or engage in effortful cognition—affects their appreciation for the novel lateral reading/skimming that these features are intended to support.[7]

The specific research questions for this study were:

(1) How do highlighting, alignment, and ordering affect reading patterns, user experience, and cognitive load?
(2) How do participants' valuation of these features relate to their Need for Cognition?
(3) Does each feature provide value on its own, or only in conjunction with one or more of the other two features?

5.1 Participants

We recruited 24 participants (15 female, 7 male, and 2 non-binary; all undergraduate students) from Harvard University via mailing lists. Participants were fluent in English and over 18 years of age. Each study took 20–30 minutes and participants received $20 USD via digital payment as compensation.

[7]NFC has been taken into account in past HCI studies when examining objective performance and subjective experience in non-trivial cognitive tasks, e.g., [2, 74].

5.2 Procedure

5.2.1 Setup. We collected 80 sentences from our abstracts dataset labeled by our system as "Methodology/Contribution." Participants viewed the same 80 sentences in each condition—often with a different subset of sentences initially visible due to ordering changes—but only had two minutes to look at them in each condition.

To contrast participants' gaze patterns in each condition, we used a Tobii Pro Spark eye-tracker placed below the desktop monitor used by all subjects; Tobii Pro Lab software recorded each participant's gaze over time in each condition.

To avoid participant fatigue from viewing all combinations of feature settings, we used a mixed within- and between-subjects design. All participants experienced the baseline, i.e., all features turned off (Figure 7A), the condition with all three features enabled (Figure 7C), and one of the three feature-ablation conditions illustrated in Figure 7B. The study used a balanced design: each of the three ablation conditions—where one of the three features was disabled—was assigned to an equal number of participants. Additional details can be found in Appendix E.

5.2.2 Participant experience. All sessions were conducted in person. At the start, participants received a brief introduction, an informed consent form, and an opportunity to ask questions. After signing the consent form, they filled out a questionnaire to determine their Need for Cognition (NFC) levels via the NCS-6 survey [42]. As a final onboarding step, the eye tracker was calibrated to better track their eyes using its standard software.

In each condition, the participants' task was to "skim" the 80 sentences for two minutes and then verbally answer the question *What do you think this collection of text is about?* After each condition, they filled out the NASA-TLX questionnaire [27].

The study concluded with a 15-minute semi-structured interview. During the interview, participants saw screenshots from the three conditions and were asked which they preferred and disliked, why, what they wished the interface had, what influenced their skimming, and how they normally skimmed texts.

5.3 Results

5.3.1 The three features lose their effectiveness when not used together. The most popular condition had all three features enabled, i.e., 11 out of 24 participants (≈50%) preferred Figure 7C, as shown in the "Preferred" columns of Table 1. The remaining participants were roughly evenly split between the no-features baseline (6 participants) and the *without-alignment* ablation condition (5 participants). One participant each liked the *without-highlighting* and *without-ordering* ablation conditions most, respectively.

The most preferred condition (all three features enabled) was tied with the baseline no-features-enabled condition for lowest reported cognitive load. Specifically, 11 participants reported their lowest raw NASA-TLX scores[8] in the all-three-features condition, and a different 11 participants reported their lowest raw NASA-TLX scores in the baseline condition. (See Table 1's "Least Cognitive Effort" columns for statistical details.) Only one participant each reported their lowest raw NASA-TLX score when working with the *without-highlighting* and *without-ordering* ablation conditions, respectively. No participant reported their lowest Raw NASA-TLX score when using the *without-alignment* condition. These results suggest that these three features lose their effectiveness when not used together.

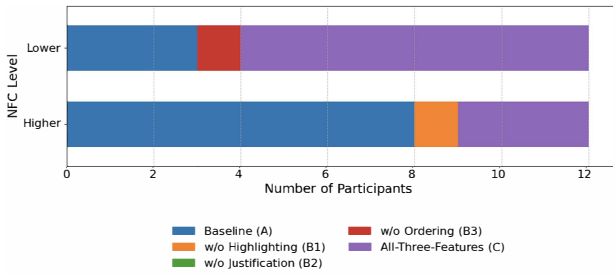


Figure 8: Stacked bar chart of conditions in which lower and higher NFC participants reported the least cognitive load, respectively.

5.3.2 Participants' Need for Cognition (NFC) explained some of participants' perceived cognitive load and preferences. For simplicity of analysis, we denote participants with NFC scores above the overall participants' median NFC of 5.42 (IQR = 0.583) as *higher NFC*, and *lower NFC* otherwise.[9] Figures in Appendix G show the distribution of participants' NFC scores as a function of the conditions in which they reported the least cognitive load (Figure 29) and which they preferred most (Figure 30). Participants with lower NFC more frequently preferred and experienced less cognitive load when skimming with all the features enabled.[10] Likewise, Figure 8 shows that most lower-NFC participants reported their lowest cognitive load when all features were enabled, while most higher-NFC participants reported their lowest cognitive load when no features were enabled (baseline).

5.3.3 Eye Tracking. Lower NFC participants were generally guided by emergent visual patterns created by the interactions between features, especially blocks of color spanning multiple sentences created when all three features are turned on. Figure 9 shows a typical lower-NFC participant reading *down, across documents* rather than left to right within a single document's sentence when all three features are working together; this is a radically different reading pattern. These participants also often verbally described in their interviews how they appreciated the colors guiding them in where and how to skim. Meanwhile, higher NFC participants often skimmed from left to right, line by line, regardless of what features were available to them (e.g., P16, whose gaze plots are shown in Figure 31 in Appendix G).

5.3.4 Lower NFC participants' skimming experiences. Both gaze data and the semi-structured interviews revealed that lower NFC participants were more willing to be guided by the three features and took advantage of them consciously. Specifically, many lower NFC participants used color to help them decide what to read, even without remembering what the colors meant: *"I think the colors definitely guide attention, even though I didn't know what the colors meant"* (P4) and *"I just read in groups. So, I did all the blue, all the orange, ... "* (P12). P13 used the features to help them decide what was important for them to read: *"When you're skimming something, you're trying to see what's the most important but given this legend, this coloring, this organization, you have a rough idea of what's the most important and that makes skimming much ... easier."*

As revealed in the semi-structured interviews, lower NFC participants seemed to enjoy and prefer how these features guided their attention, and found it cognitively less demanding to have all the features enabled, perhaps due to the guidance they believed it provided them while skimming. Many participants specifically pointed out how alignment helped them to filter out parts they thought were not important to skim: *"[vertical alignment] made it very easy to see which part is where ... I should focus on more"* (P13). P13 continued: *"It's so much easier to skim something when you're being guided."* When working with the baseline interface with no features enabled, many lower NFC users found themselves unsure about how to navigate through the text while skimming: *"everything blurred the same... I just felt kind of bored reading it"* (P24).

5.3.5 Higher NFC participants' skimming experiences. In contrast, higher NFC participants reported an easier and more pleasant experience reading in the baseline condition. Many said they appreciated that they could read as they normally would: *"[The baseline] was easiest to read just because maybe I'm used to this in general; this is how I usually read things"* (P16). P10 describes how they did not have to constantly *"restructure"* their eyes.

[8]The raw NASA-TLX score is the sum of all 6 NASA-TLX questions after reversing the appropriate questions.
[9]To compute a participant's NFC score, we averaged their response to the six questions, each ranging from 1 to 7, after reversing the appropriate questions.
[10]Using a two-tailed Mann-Whitney U Test, we found that participants who reported their lowest perceived cognitive load when all three features were enabled had significantly lower NFC than participants who reported their lowest cognitive load level when skimming with no features enabled—in the baseline interface (p=0.03).

(a) Ablation condition: without-highlighting **(b) All-three-features condition**

Figure 9: These two gaze plots show how the reading/skimming behavior of a lower NFC participant (P4) changed when role highlighting was added to the role ordering and alignment features: from more horizontal (within a sentence) to more vertical (across corresponding chunks of sentences across different abstracts).

When skimming in the all-three-features condition, many higher NFC participants felt compelled to discern patterns in the design and layout of the interface. If/when they could not find any patterns, they became increasingly frustrated. For example, P22 was distracted by thinking about *"the purpose of this design of the text"* rather than focusing *"attention on the task at hand."* P10 tried to connect the different topics covered in a block of color before realizing *"there's no connection between these two besides the overarching method."* Likely as a result, some higher NFC participants thought that the features *"made the text more dense"* (P16).

In general, higher NFC participants were annoyed by how the features guided their attention. For many, even when they were skimming, they still *"read it like a book, like, left to right"* (P23). With the presence of all three features, participants like P16 felt like *"something's trying to control how I'm reading... so, it feels a bit unnatural or confusing?"* Indeed, as shown in Figure 31 in Appendix G, in both non-baseline conditions, P16 still read from left to right. Visual elements rendered on the text likely impeded their typical, preferred reading strategy.

6 Summative User Study

After the ablation study validated the effectiveness of all three SMT-inspired features together (especially for lower NFC users), we completed the implementation of AbstractExplorer and evaluated its impact on researchers' reading and sensemaking of a corpus of all ~1000 paper abstracts from ACM CHI 2024. We did not conduct a comparative study because existing tools like the ACM Digital Library typically do not provide structured support for comparative close reading at scale, making direct comparison methodologically inappropriate and potentially misleading.

6.1 Participants

We recruited 16 participants (9 male, 7 female) from various universities across the USA through mailing lists and social media posts. Twelve of the 16 participants were between the ages of 25 and 34, while the remaining four were younger (18–24). The majority (nine) were PhD students, five were Master's students, one was an industry researcher, and one was academic research staff. The group had a roughly balanced mix of individuals with varying levels of experience in HCI, with eight participants having attended CHI or a similar HCI conference; three had specifically attended CHI 2024. Additionally, six participants were actively preparing manuscripts for submission to CHI 2025 or a similar HCI venue. Participants were roughly evenly split between lower and higher familiarity when asked to rate their familiarity with CHI or other major HCI confererences, i.e., UIST or CSCW, on a 1-7 scale. The average Need for Cognition (NCS-6) score [42] was 5.55, with a standard deviation of 0.59, indicating that participants generally had a moderate to high tendency to engage in and enjoy thinking, with relatively low variability across the group.

6.2 Study Procedure

All studies were conducted remotely via Zoom and facilitated by the first author. Each study took approximately 60-75 minutes, and participants received $25 (USD) via digital payment (Zelle or Venmo) as compensation.

6.2.1 Consent, Pre-Study Survey, and Tutorial. After providing verbal informed consent, participants completed a pre-study survey (Appendix J.1). They then accessed AbstractExplorer via a web link (no installation required) and watched a 3.5-min tutorial video demonstrating its features on the corpus of CHI 2024 paper abstracts. Participants were also provided with a reference sheet listing key terms and their definitions and annotated screenshots that explain different features. This reference sheet was available to participants throughout the study session. To help familiarize participants with AbstractExplorer, the study coordinator instructed them to practice three core interactions—comparing alignment options, toggling chunk highlighting, and creating a custom aspect—as warm-up exercises. This phase ensured that participants were aware of key features of AbstractExplorer.

6.2.2 Task 1: General Reading. Following the warm-up exercises, participants performed a general reading task using AbstractExplorer. They were presented with the following task: *imagine writing a survey on CHI 2024 papers, skim as many abstracts as possible, develop a mental model of the range of contributions, methodologies, problem domains, and study results, and identify three emergent patterns in content or style* (exact wording in Appendix I.1) and instructed to use all interface features except filters, to ensure exposure to a sufficient number of papers. They shared their screens (recorded by the research team) and were encouraged to think aloud while optionally taking notes in a provided document. The task had a 15-minute time limit, and participants received a reminder when 5 minutes remained.

6.2.3 Task 2: Focused Reading. In the second task, participants performed a focused reading task using AbstractExplorer. The task description and goals were similar to Task 1, but this time participants were instructed to target abstracts relevant to their own research or personal interests (exact wording in Appendix I.2). This task aimed to evaluate how users personalized the interface for work-related exploration, where deeper engagement might occur naturally. Participants used all interface features, especially filters and user-defined aspects, while screensharing (recorded by the research team as well) with optional thinking aloud and note taking. The task also lasted 15 minutes with a 5-minute reminder.

6.2.4 Interview and Post-Study Survey. Immediately after the focused reading task, we conducted a short interview asking participants to reflect on their experience with both tasks (Appendix J.2), followed by a post-study survey (Appendix J.3). Throughout the two tasks, we also collected detailed interaction logs including counts of user-defined aspects created, duration of highlighting usage, and time allocation across the three possible alignment options.

7 Results

In this section, we present findings on how AbstractExplorer supports comparative close reading at scale by integrating quantitative survey responses and log data with qualitative analysis of transcripts and open-ended responses. The qualitative analysis process is described in detail in Appendix H.

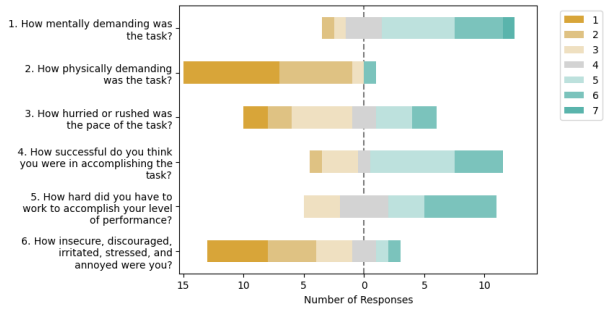


Figure 10: Distribution of participants' NASA-TLX responses after using AbstractExplorer. Responses are on a scale from 1 (very low) to 7 (very high), color-coded by score.

Overall, participants were positive about AbstractExplorer and its features, mentioning that they made the task easier or led to a better experience in one way or another (15 out of 16 participants), highlighting the benefit for more efficient skimming (9 participants) and comparison of abstracts (5 participants), which allowed them to read abstracts at scale (5 participants), although they also pointed out issues such as misclassified aspects and/or chunks (3 participants), readability issues (4 participants), and steeper learning curves with the user-defined feature (3 participants). According to NASA-TLX responses (Figure 10), participants perceived the reading task as mentally demanding (M=4.88, STD=1.26)—consistent with their reported effort levels (M=4.75, STD=1.18)—but they also reported comparable levels of perceived success in accomplishing the task (M=4.63, STD=1.26). Participants did not find the task too hurried or rushed (M=3.50, STD=1.59), and despite the novel presentation style and feature-rich design of the interface, reported low levels of negative effect such as insecurity, discouragement, irritation, stress, and annoyance (M=2.56, STD=1.55).

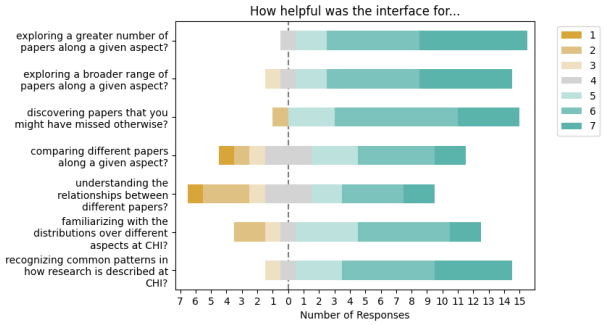


Figure 11: Distribution of AbstractExplorer's helpfulness for various tasks, as rated by participants; each row represents a task, with responses rated on a Likert scale from 1 (not useful) to 7 (very useful), color-coded by score. Participants consistently rated the interface as helpful for key literature exploration activities.

7.1 SMT-Inspired Chunk Ordering, Highlighting and Alignment

7.1.1 Skimming similarly structured sentences efficiently. More than half of the participants noted that AbstractExplorer increased their reading or skimming efficiency (P1, P4, P6, P9, P10, P11, P14, P15, P16). As for *how* it helped, many participants mentioned the two main features they noticed—highlighting and alignment. For instance, P15 said *"The color coding is quite effective for skimming; the alignment also helps and is flexible enough to accommodate different situations."* P9 further explained, *"It allows me to focus on analogous sentences across papers. It's easier to read the same type of sentences with similar sentence structures."* Similarly, P10 reported faster reading through what they described as *vertical reading*, which we also observed in the Ablation Study (Section 5). We observed this strategy among multiple summative study participants, i.e., skimming similarly colored chunks across sentences. Using this strategy, participants focused primarily on one highlighted cross-sentence block

of same-colored chunks and moved across sentences quickly, reading other chunks only when necessary. P3 and P6 also utilized this reading strategy and kept highlighting on for just their chosen focal chunk roles to minimize distractions.

Activity log data, which revealed how participants actually used the interface, echoed the above findings. According to the log data, participants spent most of their reading time (66.31%) with vertical alignment on the second element in structure pairs, followed by alignment on the first element (29.19%), and left-justified alignment (5.13%). Highlighting usage showed a similar preference: 91.13% of time with all chunks highlighted, 8.25% with partial highlighting, and minimal time (0.63%) without highlights. Notably, since participants were required to try all alignment and highlighting options during warmup exercises, their sustained use of SMT-inspired vertical alignment and chunk highlighting features after experimenting with alternatives provided strong evidence that they found these features helpful for their tasks.

As shown in Figure 12, participants' ratings of the usefulness of different features provides further evidence of their utility. Vertical alignment by the second element (M=6.19, STD=0.83) and chunk highlighting (M=6.06, STD=1.34) were the top two most useful features rated by participants. Vertical alignment by the first element (M=5.63, STD=1.09) received slightly lower but still favorable usefulness rating. Left-justified alignment received more mixed ratings (M=4.25, STD=1.84), demonstrating participants' ability to differentiate between different alignment options and preference for SMT-inspired vertical alignment.

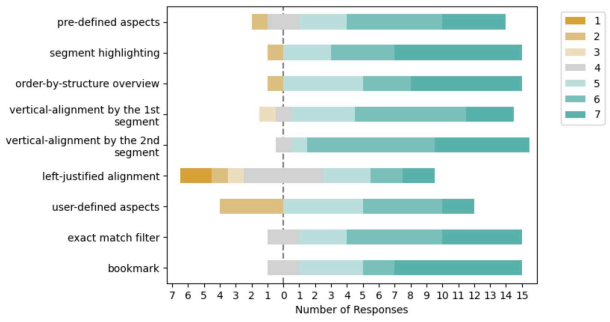


Figure 12: Distribution of participants' responses on the usefulness of different features. Each row represents a feature, with responses rated on a Likert scale from 1 (not useful) to 7 (very useful), color-coded by score.

7.1.2 Reading Sentences at Scale. As reflected in Figure 11, participants highlighted how the interface scaled up the number of papers they could review (P2, P4, P9, P11, P16), finding the interface to be highly helpful for exploring a greater number (M=6.19, STD=0.91) and broader range (M=5.94, STD=1.18) of paper abstracts along a given aspect than they typically would. For example, P2 said, *"It's very helpful in getting familiar with large numbers of articles/topics in a short time, esp. most papers have the similar structure."* Similarly, P16 wrote, *"The way it groups and renders sentences make it very easy to skim through many papers in this conference, something hard to do with traditional interfaces."* P9 also noted reading a greater number and diversity of abstracts in the same timeframe with AbstractExplorer. Participants also rated the interface as very helpful for discovering otherwise-missed papers (M=5.81, STD=1.22). Many mentioned serendipitous discoveries of new papers (P4, P8, P11) and a more comprehensive grasp of the collection (P7, P11, P15). As P4 remarked, *"I ran into many interesting papers that I will probably miss if using a traditional reader."*

7.1.3 Comparing sentences to find common patterns and outliers. Participants leveraged chunk highlighting and alignment for cross-paper comparison as well. For instance, P5 valued how the interface enabled them to *"compare a well-defined aspect across selected papers, such as methodology and study size."* Beyond identifying differences, participants rated the interface as very helpful for recognizing common patterns in CHI abstracts (M=5.81, STD=1.17) (Figure 11) and discovered interesting patterns in both content and writing style (P2, P4, P8, P14). For example, P2 noted a recurring sentence structure *"we developed xxx, using xxx, showing xxx"* when authors introduced novel systems, while P4 investigated typical participant counts for different study types. Two participants (P4, P14) reported that recognizing these common patterns helped them write their own abstracts for HCI venues.

7.1.4 Improving readability, context, and classification accuracy. Despite broad preference for the AbstractExplorer interface, participants identified several opportunities for improvement. Two noted that the need for horizontal scrolling interfered with skimming. P10 said that *"horizontal scroll makes switching to different lines a bit hard,"* and P6 added that *"some sentences are too long, requiring users to scroll to the end. Implementing a Line Break Mode or providing sentence breakdowns/summarization could improve readability."* While some found the interface "visually appealing" (P4), "intuitive," and "not too cluttered" (P7), others described it as "overwhelming" (P8, P16). P8 elaborated, *"There are so many colors, sometimes it distracted me or [made me] lose my attention."*

Lack of contextual information presented another challenge: while P14 appreciated the ability to access original abstracts by clicking sentences, P4 and P16 noted difficulties with limited context when parsing abstracts at the sentence level. P4 mentioned unclear pronoun references (e.g., "This system" or "to achieve that") resulting from splitting abstracts into single sentences, and P16 suggested the addition of *"a mechanism to see the previous and next sentences of a given sentence for better context."* Misclassification of sentences and sentence chunks also emerged as a potential concern, though it appeared infrequently and had minimal impact on the user experience. Only three out of sixteen participants reported such issues. P4 mentioned that *"While the labels are usually accurate, sometimes they are mislabeled."* P7 observed that *"The aspect types sometimes did not fully match the results shown for them,"* and P16 called for *"higher-quality labels,"* but still appreciated that the system made common patterns *"really clear."*

7.2 User-Defined Aspects

Each participant generated at least one user-defined aspect, averaging 2.25 (STD=1.57) per participant, with heavier use during the second task (focused reading), demonstrating active engagement

with this feature. Examples generated by participants included "topics in games" (P1), "methods in design" (P2), and "trends in AI for health" (P11). Though user-defined aspects received comparatively lower ratings (M=4.81, STD=1.80) than other features, they were still rated as useful by most participants (Figure 12). In qualitative responses, many participants described the user-defined aspect feature as "effective" (P2, P10) and helpful for literature searches (P4, P16). P12 praised the custom aspect results for their high quality, remarking, *"the system is very smart."* P16 noted, *"The user-defined feature is very cool and has a lot of potential, especially for researchers in their literature review process, when they want to find related papers from a niche aspect."* However, some participants reported concerns about slow result generation (P1, P15) and perceived overlap with the filter feature (P5, P11).

The feature's effectiveness seemed to vary by experience level. HCI-experienced participants appeared to derive more value from it, while those who were less experienced faced a steep learning curve—particularly in formulating meaningful domain or aspect names. P14, who lacked familiarity with HCI venues, explained, *"The user-defined function was a bit confusing, especially since I had little understanding of what the set of papers was specifically about, making it difficult for me to come up with helpful keywords that I can enter into the three boxes."* P8, another novice, confused "aspect" with "domain," entering topic nouns like "children" and "VR" into the aspect field, which produced less coherent outputs. These observations underscored the need for an improved UI design for authoring custom aspects, especially for those who are less familiar with the research corpus.

7.3 Real-World Usages

When asked about potential real-world usages of the interface, participants responded positively overall: 13 indicated they would use it regularly if made available, two would use it situationally, and only one preferred traditional interfaces. P11 noted, *"Probably—it really helps with reading papers, especially when I want to get an overview or write a survey about a conference."* Participants also envisioned specific use cases such as pre-conference applications and literature review. For instance, P3 stated they would use it *"to write abstracts and find aspects of abstracts that will influence my decision to attend a particular talk at the conference."* P16 explained, *"I would use it to go through a conference proceeding before attending it. I can also see myself using the user-defined feature when doing literature review."* P12 expressed conditional adoption, suggesting, *"Maybe, but I'd be more inclined to use it if there were options to export selected content or integrate the interface with other tools like Zotero, Notion, or reference managers."*

8 Discussion

Our ablation and summative studies verified the value of AbstractExplorer, specifically showing that all three components of the Structural Mapping Engine—color coding, sentence ordering, and vertical alignment—are crucial for facilitating comparative close reading at scale. Building on our findings, we now reflect on both the conceptual framework and empirical results of our design.

8.1 Rethinking Sensemaking and Information Seeking Paradigms

Like prior Structural Mapping Theory (SMT)-informed work in text corpora representation, AbstractExplorer's features have enabled some users to see more of both the overview and the details at the same time, facilitating abstraction without losing context. In other words, rather than be overwhelmed by a wall of text, precomputing and reifying cross-document *analogous* relationships make it psychologically possible for users to engage—if they are willing to be guided by it. (Lower NFC users are more likely to fall into this category.) With this augmented perception of the original text, users can notice both cross-document relationships already computed for them and beyond.

As a result, our approach can go beyond commonly used information seeking paradigms, including Overview First Paradigm ("Overview first, zoom and filter, then details on demand") [58] and Search First Paradigm ("Search, Show Context, Expand on Demand") [64]. While our work is not a critique of established paradigms, we align with prior works that demonstrate how alternative approaches are beneficial for seeking information from large corpora of spatial data where the notion of overview and details are poorly defined [44], global summaries are unimportant as they remove nuances [62, 64], and details on demand is impractical or inefficient due to the demand of repetitive interactions [7].

In this work, we introduce a new paradigm for exploring a large corpus of small documents by *identifying roles at the phrasal and sentence levels, then slice on, reify, group, and/or align the text itself on those roles, with sentences left intact.* We demonstrate how *slicing* sentences according to roles and visually *aligning* them can help readers perceive cross-document relationships in a coherent manner. We extend existing approaches through automated role annotation, establishing alignments using grammatical chunk boundaries, and preserving sentences in their entirety, instead of relying on abstract meta-data. In the context of close reading of research paper abstracts at scale, our findings suggest AbstractExplorer enabled participants to scale up the number of papers they could review through efficient skimming and find common patterns and outliers through sentence comparison, resulting in a rich synthesis of ideas and connections to foster deeper engagement with scholarly articles. We posit that our approach can generalize to other domains such as journalism [21], code synthesis [20, 22], and social media analytics [34] where visual alignment of text can enable meaningful comparisons of underlying patterns to identify relational clarity.

8.2 Facilitating Structured Variation Seeking to Invite Cognitive Engagement

Our work introduces novel human cognition-informed affordances that facilitate and invite users to make use of variation present in a corpus of abstracts; participants found value in these affordances and used them to engage with the revealed variation. We will call them *variation affordances*. AbstractExplorer used variation affordances present in prior systems, e.g., color-coordinated highlighting of analogous text in Gero et al. [18], and introduced new ones, such as alignment of sentences based on analogous chunks within them, which had only been hypothesized in prior work [21].

By definition, sensemaking and other dialectical activities [73] necessitate engagement. Our work demonstrates that designs informed by Structure-Mapping Theory can support users in navigating, making use of, and engaging with variation present in information. In this sense, AbstractExplorer enables dialectical activities that users may otherwise have found to be too tedious or difficult to engage with. Dialectical activities cannot be done on a user's behalf by AI; with variation affordances, AI is supporting the user's engagement with the data themselves.

8.3 Limitations and Future Work

We chose to test AbstractExplorer using CHI abstracts as they are a corpus of diverse short documents with implicit norms for content and style. However, AbstractExplorer could be generalized to abstracts from other CS conferences, journal articles, or even beyond abstracts to other types of implicitly or explicitly structured short documents. According to SMT, this generalization depends on most documents having some shared implicit *structure*. Supporting a new corpus would require defining new corpus-appropriate predefined aspects and chunk role labels. Other components would likely be able to remain unchanged.

Reading abstracts along one aspect at a time offers both advantages and limitations. On one hand, it enables users to quickly skim similar sentences across papers at scale. However, such sentence-level analysis may oversimplify the more complex arguments that span across several sentences. Abstracts, while concise, often contain intricate reasoning that may be split across several lines, and focusing on sentence-level patterns may lead to a loss of context. While our system includes linkages that allow users to quickly access the full abstract by clicking on any selected sentence, the resulting context-switching can be cognitively demanding and disruptive. Future work could explore more seamless ways of preserving context, such as allowing users to navigate through every sentence of an abstract directly within the Cross-Sentence Relationship pane, fostering a more cohesive understanding of the content.

Acknowledgments

We appreciate the engagement of all our participants. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391, IIS-1955699, and CCF-2123965. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. It was also conducted during the tenure of an Alfred P. Sloan Research Fellowship (Grant No. FG-2023-19960). The authors acknowledge the Alfred P. Sloan Foundation for its support. The Foundation had no role in the design, execution, or interpretation of the research. This material is also based on work that is partially funded by an unrestricted gift from Google.

References

[1] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[2] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[3] Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi (Eds.). Association for Computational Linguistics, Berlin, Germany, 554–563. doi:10.18653/v1/W16-2347

[4] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[5] John T Cacioppo, Richard E Petty, Jeffrey A Feinstein, and W Blair G Jarvis. 1996. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological bulletin* 119, 2 (1996), 197.

[6] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–21.

[7] Min Chen, Miquel Feixas, Ivan Viola, Anton Bardera, Han-Wei Shen, and Mateu Sbert. 2016. *Information theory tools for visualization.* AK Peters/CRC Press.

[8] Juliet Corbin et al. 1990. Basics of qualitative research grounded theory procedures and techniques. (1990).

[9] Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena L. Glassman. 2025. CorpusStudio: Surfacing Emergent Patterns in a Corpus of Prior Work while Writing. doi:10.1145/3706598.3713974

[10] R Haentjens Dekker and Gregor Middell. 2011. Computer-supported collation with CollateX: managing textual variance in an environment with varying requirements. In *Supporting Digital Humanities 2011: Answering the unaskable.*

[11] Nicholas Diakopoulos, Dag Elgesem, Andrew Salway, Amy Zhang, and Knut Hofland. 2015. Compare clouds: Visualizing text corpora to compare media frames. In *Proceedings of IUI Workshop on Visual Text Analytics*. Citeseer, 193–202.

[12] Tarek Eid, Eric vanSonnenberg, Antoine Azar, Porus Mistry, Kareem Eid, and Paul Kang. 2018. Analysis of the variability of abstract structures in medical journals. *Journal of general internal medicine* 33 (2018), 1013–1014.

[13] Weiqi Feng and Dong Deng. 2021. Allign: Aligning All-Pair Near-Duplicate Passages in Long Texts. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 541–553. doi:10.1145/3448016.3457548

[14] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *arXiv preprint arXiv:2310.07581* (2023).

[15] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 476–490.

[16] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. ACM, Hamburg Germany, 1–19.

[17] Dedre Gentner and Christian Hoyos. 2017. Analogy and abstraction. *Topics in cognitive science* 9, 3 (2017), 672–693.

[18] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K Kummerfeld, and Elena L Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[19] Elena L Glassman, Lyla Fischer, Jeremy Scott, and Robert C Miller. 2015. Foobaz: Variable name feedback for student code at scale. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 609–617.

[20] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. 2015. OverCode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 2 (2015), 1–35.

[21] Elena L. Glassman, Janet Sung, Katherine Qian, Yuri Vishnevsky, and Amy X. Zhang. [n. d.]. Triangulating the News: Visualizing Commonality and Variation Across Many News Stories on the Same Event. https://bpb-us-e1.wpmucdn.com/sites.northeastern.edu/dist/0/367/files/2019/11/CJ_2020_paper_67.pdf

[22] Elena L Glassman, Tianyi Zhang, Björn Hartmann, and Miryung Kim. 2018. Visualizing API usage examples at scale. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[23] Pascal Goffin, Tanja Blascheck, Petra Isenberg, and Wesley Willett. 2020. Interaction techniques for visual exploration using embedded word-scale visualizations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[24] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K Kummerfeld, and Elena L Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[25] Ronald Haentjens Dekker, Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris Van Zundert. 2015. Computer-supported collation of modern manuscripts:

CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities* 30, 3 (2015), 452–470.
[26] Abram Handler, Narges Mahyar, and Brendan O'Connor. 2022. ClioQuery: Interactive query-oriented text analytics for comprehensive investigation of historical news archives. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–49.
[27] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[28] Aaron Hoffman, Bradley Love, and Arthur Markman. 2010. Selective Attention by Structural Alignment: An Eyetracking Study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 32.
[29] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling creative inspiration with fine-grained functional aspects of ideas. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
[30] Ting-Hao'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. 2020. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367* (2020).
[31] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
[32] Stefan Jänicke, Thomas Efer, Marco Büchler, and Gerik Scheuermann. 2014. Designing close and distant reading visualizations for text re-use. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 153–171.
[33] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. 2022. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.
[34] Mahmood Jasim, Mumtaz Fatima, Sagarika Ramesh Sonni, and Narges Mahyar. 2023. Bridging the Divide: Promoting Serendipitous Discovery of Opposing Viewpoints with Visual Analytics in Social Media. 26–30.
[35] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, Vol. 12. Springer, 1–10.
[36] Steffen Koch, Markus John, Michael Wörner, Andreas Müller, and Thomas Ertl. 2014. VarifocalReader—in-depth visual analysis of large text documents. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1723–1732.
[37] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.
[38] Kenneth J Kurtz and Dedre Gentner. 2013. Detecting anomalous features in complex stimuli: the role of structured comparison. *Journal of Experimental Psychology: Applied* 19, 3 (2013), 219.
[39] Jing Li, Jean-Bernard Martens, and Jarke J van Wijk. 2010. A model of symbol size discrimination in scatterplots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2553–2562.
[40] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28, 7 (2012), 991–1000.
[41] Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. (2010).
[42] Gabriel Lins de Holanda Coelho, Paul HP Hanel, and Lukas J. Wolf. 2020. The very efficient assessment of need for cognition: Developing a six-item version. *Assessment* 27, 8 (2020), 1870–1885.
[43] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, et al. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *arXiv preprint arXiv:2303.14334* (2023).
[44] Timothy Luciani, Andrew Burks, Cassiano Sugiyama, Jonathan Komperda, and G Elisabeta Marai. 2018. Details-first, show context, overview last: supporting exploration of viscous fingers in large-scale ensemble simulations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 1225–1235.
[45] Ference Marton. 2014. *Necessary conditions of learning*. Routledge.
[46] Justin Matejka, Tovi Grossman, and George Fitzmaurice. [n. d.]. Paper forager: Supporting the rapid exploration of research document collections. In *Graphics Interface 2021*.
[47] Bryan J Matlen, Dedre Gentner, and Steven L Franconeri. 2020. Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance* 46, 5 (2020), 443.
[48] Frank Meng, Craig A Morioka, and Suzie El-Saden. 2011. Determining word sequence variation patterns in clinical documents using multiple sequence alignment. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 934.
[49] Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.
[50] Aditi Muralidharan and Marti A. Hearst. 2014. Improving the Recognizability of Syntactic Relations Using Contextualized Examples. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Kristina Toutanova and Hua Wu (Eds.). Association for Computational Linguistics, Baltimore, Maryland, 272–277. doi:10.3115/v1/P14-2045
[51] Aditi Muralidharan, Marti A Hearst, Christopher Fan, and Exequiel Ganding III. [n. d.]. Text Sliding: Information Discovery with Intensely Integrated Text Analysis. ([n. d.]).
[52] Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube–A Test Collection of Computer Science Research Articles for Faceted Query by Example. *arXiv preprint arXiv:2103.12906* (2021).
[53] Mahin Naderifar, Hamideh Goli, and Fereshteh Ghaljaie. 2017. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education* 14, 3 (2017).
[54] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[55] Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.
[56] André Santos, José João Almeida, and Nuno Carvalho. 2012. Structural alignment of plain text books. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 2069–2074. https://aclanthology.org/L12-1576/
[57] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.
[58] Ben Shneiderman. [n. d.]. Visual Information-Seeking Mantra. https://infoviswiki.net/wiki/Visual_Information-Seeking_Mantra. Accessed: 2024-09-08.
[59] Aditi Shrikumar. 2013. *Designing an Exploratory Text Analysis Tool for Humanities and Social Sciences Research*. Ph. D. Dissertation. UC Berkeley.
[60] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of business research* 104 (2019), 333–339.
[61] Jänicke Stefan, Greta Franzini, Muhammad Faisal Cheema, and Scheuermann Gerik. 2015. On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)(2015)*. R. Borgo, F. Ganovelli, I. Viola, N–A.
[62] Nicole Sultanum, Michael Brudno, Daniel Wigdor, and Fanny Chevalier. 2018. More text please! understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[63] Mariia Tytarenko, Lin Shao, Tobias Walter Rutar, Michael A Bedek, Cornelia Krenn, Stefan Lengauer, and Tobias Schreck. 2024. Hierarchical Topic Maps for Visual Exploration and Comparison of Documents. (2024).
[64] Frank Van Ham and Adam Perer. 2009. "Search, show context, expand on demand": Supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 953–960.
[65] Martin Wattenberg and Fernanda B Viégas. 2008. The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics* 14, 6 (2008), 1221–1228.
[66] Yating Wei, Honghui Mei, Ying Zhao, Shuyue Zhou, Bingru Lin, Haojing Jiang, and Wei Chen. 2019. Evaluating perceptual bias during geometric scaling of scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 321–331.
[67] Sam Wineburg and Sarah McGrew. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record* 121, 11 (2019), 1–40. doi:10.1177/016146811912101102 arXiv:https://doi.org/10.1177/016146811912101102
[68] Tongshuang Wu, Kanit Wongsuphasawat, Donghao Ren, Kayur Patel, and Chris DuBois. 2020. Tempura: Query analysis with structural templates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[69] Litao Yan, Miryung Kim, Bjoern Hartmann, Tianyi Zhang, and Elena L. Glassman. 2022. Concept-Annotated Examples for Library Comparison. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 1–16. doi:10.1145/3526113.3545647
[70] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N Truong. 2011. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1541–1550.
[71] Vahan Yoghourdjian, Yalong Yang, Tim Dwyer, Lee Lawrence, Michael Wybrow, and Kim Marriott. 2020. Scalability of network visualisation from a cognitive load perspective. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 1677–1687.

[72] Tariq Yousef and Stefan Janicke. 2020. A survey of text alignment visualization. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 1149–1159.
[73] Haoqi Zhang. 2024. Searching for the Non-Consequential: Dialectical Activities in HCI and the Limits of Computers. Association for Computing Machinery. doi:10.1145/3613904.3641945
[74] Tianyi Zhang, Zhiyang Chen, Yuanli Zhu, Priyan Vaithilingam, Xinyu Wang, and Elena L Glassman. 2021. Interpretable program synthesis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[75] Yinyuan Zheng, Bryan Matlen, and Dedre Gentner. 2022. Spatial Alignment Facilitates Visual Comparison in Children. *Cognitive Science* 46, 8 (Aug. 2022), e13182. doi:10.1111/cogs.13182
[76] Joyce Zhou, Elena Glassman, and Daniel S. Weld. 2023. An Interactive UI to Support Sensemaking over Collections of Parallel Texts.

A Formative Interview Study Participants

We recruited 12 active researchers by word of mouth followed by snowball sampling [53]. Demographically, seven with he/him pronouns, four with she/her pronouns, and one who did not disclose. Two were 18–24 years old, 7 were 26–35, and 3 were 36–45. Participants were in varying stages of their research careers. Three participants had been active researchers for 1-5 years, six participants had been active for 6-10 years, and three participants had been active researchers for more than 10 years. Their research interests included human-computer interaction, software engineering, web security, music, nanophotonics, and computational economics, among others. All of them except one had published manuscripts in conferences or journals in their respective research fields, and the remaining participant was finalizing a manuscript for submission. While English was the primary language of research for all participants, three mentioned being most comfortable with languages other than English.

B Formative Interview Study Data Collection and Analysis

Interviews were video and audio recorded. We transcribed the audio using OpenAI's Whisper automatic speech recognition system and anonymized the transcript before analysis. We analyzed the interview data using thematic analysis [1]. First, two members of the research team independently coded four (25% of collected data) randomly chosen participant data to generate low-level codes. The inter-coder reliability between the coders was 0.88 using Krippendorff's alpha [37]. The two coders then met together to cross-check, resolve coding conflicts, and consolidate the codes into a codebook across two sessions. Using the codebook, the two coders analyzed six randomly selected participant data each. The research team then met, discussed the analysis outcomes, and finalized themes over three sessions.

C Formative Interview Study Design Probes

C.1 Cross-document relationship design probes

Figures 13, 14, 15, 16, 17, 18, 19, 20, and 21 are the full collection of cross-document relationship visualization design probes used in the formative study.

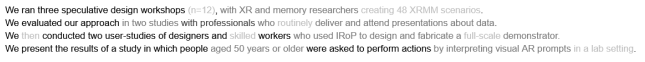


Figure 13: CDR design probe featuring varying greyscale emphasis depending on sentence subsection importance

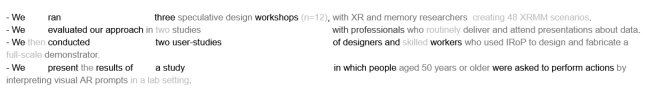


Figure 14: CDR design probe featuring both greyscale emphasis and multiple subsection alignment

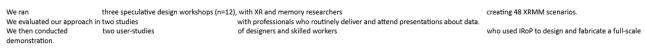


Figure 15: CDR design probe featuring multiple subsection alignment

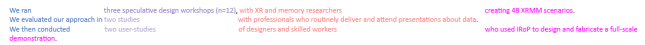


Figure 16: CDR design probe featuring colored text with multiple subsection alignment

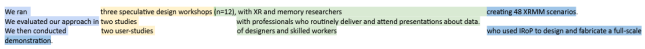


Figure 17: CDR design probe featuring highlighted text with multiple subsection alignment

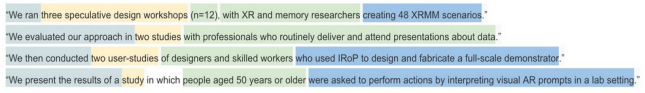


Figure 18: CDR design probe featuring highlighted text

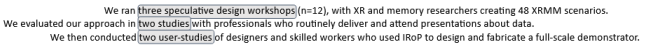


Figure 19: CDR design probe featuring two-column alignment on a user-selected attribute with highlights for emphasis

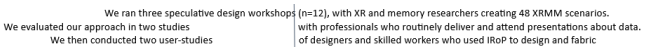


Figure 20: CDR design probe featuring three-column alignment on a user-selected attribute

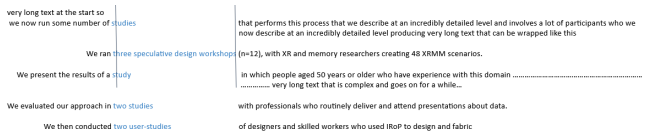


Figure 21: CDR design probe featuring three-column alignment on a user-selected attribute with colored text for emphasis

C.2 Custom aspect creation design probes

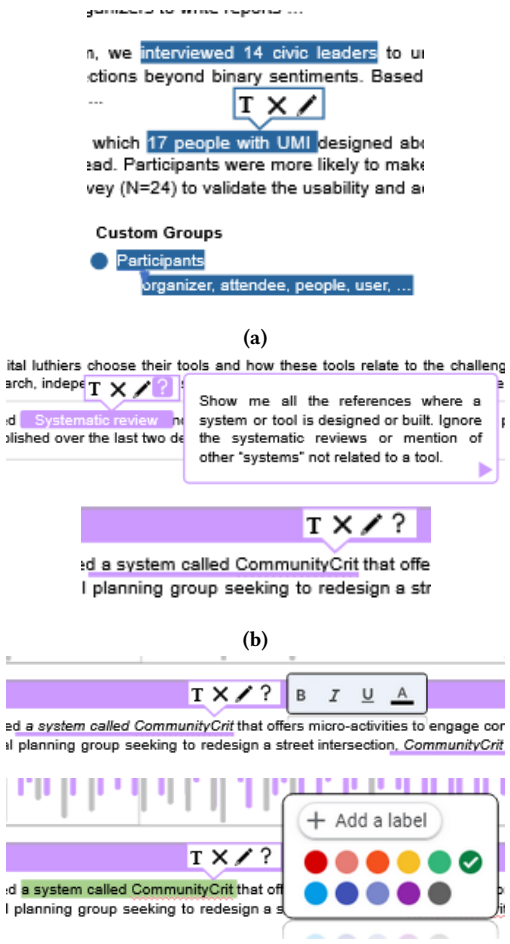


Figure 22: Example mock-ups of custom aspect creation tools and visualizers we created for the formative study inspired by PaTAT [16]: (a) shows custom aspect creation via entering a description or manual annotations. (b) shows editing a LLM-generated description or marking feedback on suggested results. (c) shows potential visualization controls.

The following custom aspect design probes used in the formative study were very lightly animated to indicate user interactivity. Figures 23, 24, 25, 26, and 27 are notable key frames from the probe.

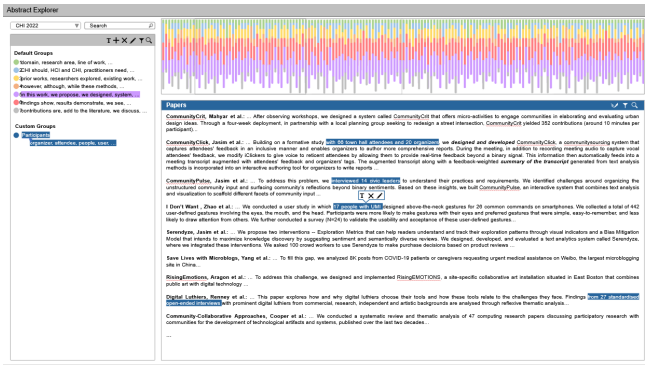


Figure 23: Custom aspect design probe: allowing users to highlight or edit relevant texts to create or edit a custom aspect

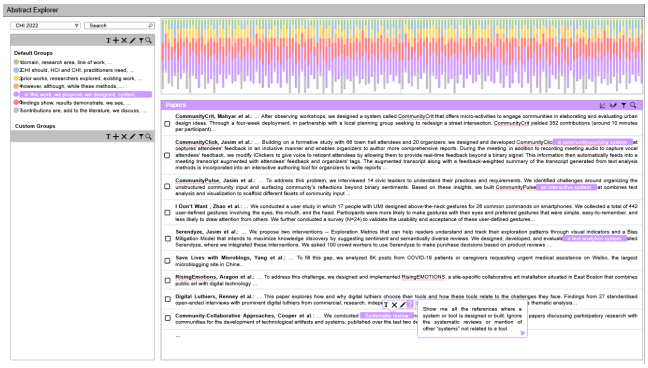


Figure 24: Custom aspect design probe: using user-written description to create, edit, or reflect on a custom aspect

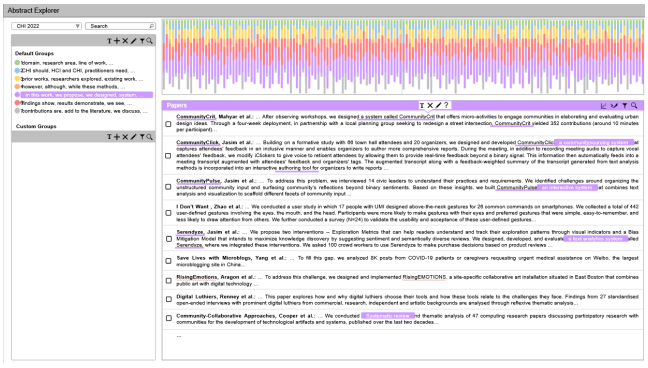


Figure 25: Custom aspect design probe: The system could provide suggestions for the user to accept or reject

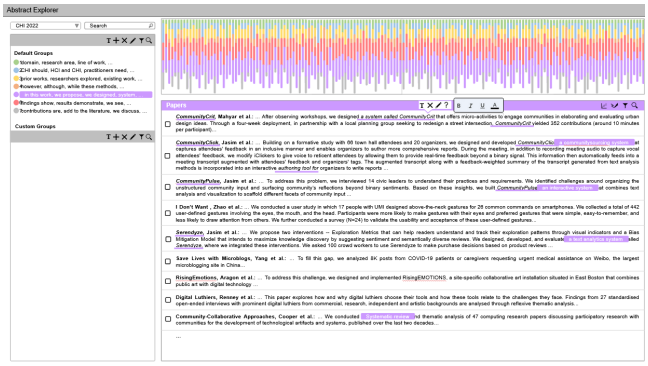


Figure 26: Custom aspect design probe: user could customize how the aspect is visualized, using italics

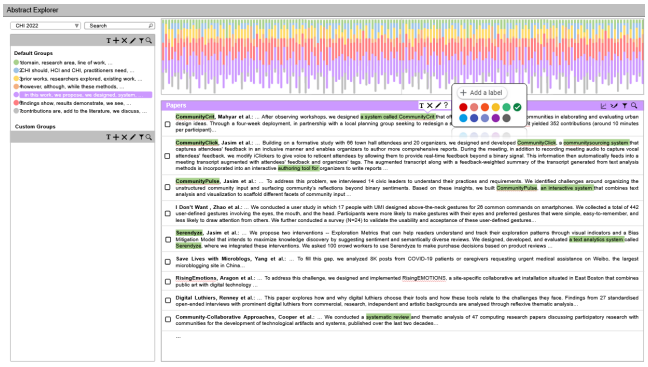


Figure 27: Custom aspect design probe: user could customize how the aspect is visualized, using highlights

D LLM Prompts

Here, we provide the detailed prompts used in our implementation of AbstractExplorer.

D.1 Sentence Categorization

Table 2 contains the prompt used to categorize each sentence into one of the pre-defined aspects.

D.2 Chunk Segmentation

Table 3 contains the prompt used to segment sentences into chunks.

D.3 Chunk Annotation

Table 4 contains the prompt used to annotate each chunk with the relevant facet type.

D.4 Custom Aspects

Table 5 contains the prompt used to generate an internal aspect description for downstream processing. Table 6 contains the prompt used to identify relevant sentences based on the initial and internal aspect description.

Table 2: Prompt used to classify each sentence into one of the pre-defined aspects.

```
The following sentence "${sentence}" is from the abstract of a CHI 2024 paper.
The full abstract is "${abstract}".
Please classify the sentence into one of the five categories below.

Categories:
0 Problem Domain (Introduction of the problem area),
1 Gaps in Prior Work,
2 Methodology (Work done by the authors),
3 Results & Findings,
4 Conclusion (Or implications for future work)

Please answer only the category number.
```

Table 3: Prompt used to identify chunk boundaries (segmenting).

```
Does the following sentence end properly?.
"${sentence}"

Please answer only Yes or No.
```

Table 4: Prompt used to annotate or classify each sentence chunk.

```
A sentence (from a CHI 2024 paper abstract) was splitted into several segments, put into the following list. For each list element, please classify it into one of the 9 categories below, based on what it describes.
"${sentence}"

Categories:
0 Status Quo/Context (the particular context or existing work)
1 Challenge/Problem/Obstacle
2 Contribution (what the authors did)
3 Purpose/Goal/Focus (why the work was done)
4 Methodology (how the work was done)
5 Participants (who were involved)
6 System Description (of a system the authors developed or proposed)
7 Findings (what are the results)
8 Enumeration (a list of things)

Here are some examples for your reference: "${examples}"

Please return a python list of the category numbers only.
The length of that list must be the same as that of the input list.
If the task is impossible, return an empty list.
```

These prompts were tuned manually. We identified multiple example aspects by sampling from comments that participants made in the formative interviews regarding possible custom aspects they

might be interested in, and iterated over system prompts while manually reviewing the quality of each prompt result for these example aspects.

Table 5: Prompt used to generate aspect description. Note that the example lists support a variable number of quotes. If there were no examples, a different sentence was used.

Here is the user's query:
"""
${query text}
"""

Here are some examples of sentences or sentence quotes that MATCHED the query:
- "${sentence quote 1}"
- "${sentence quote 2}"
- "${sentence quote 3}"

Here are some examples of sentences or sentence quotes that DID NOT MATCH the query or are otherwise NOT IDEAL quotes for this query:
- "${sentence quote 1}"
- "${sentence quote 2}"
- "${sentence quote 3}"

${if there are no examples at all, remove the two above texts and instead only include: "The user has not given any example matching or non-matching quotes yet."}

There are a few ways a paper abstract might match the query.
CONTENT: The query describes some focus, methodology, or some other holistic attribute of a paper, and the paper matches that filter.
TIDBIT: The query describes some type of information about a paper that the user is looking for, and the abstract provides that information.

Answer each of these questions with 1-3 sentences:
A. What type of query is this?
B. What information in a paper abstract would indicate whether that abstract matches the query?
C. The user wants to get one short quoted phrase from each abstract that matches the query. What information should be included in the quote? What information does not need to be in the quote? You may use any examples from the user when reasoning about this. Remember that we need quotes to be as concise as possible.

E Ablation Study Setup

A number of simplifications were made to accommodate eye-tracking software limitations and reduce confounding factors. To just focus on reading-based insights and behavior and eliminate the complications of interaction, e.g., toggling a feature on and off, we removed everything but the sentences themselves rendered with each condition's enabled features, i.e., as a static webpage. The constraints of our screen-based eye tracking software also dictated some modifications: Given the eye-tracker's resolution, sentences were rendered

Table 6: Prompt used to identify relevant abstracts and select abstract quotes.

Here is the user's query:
"""
${query text}
"""

There are a few ways a paper abstract might match the query.
CONTENT: The query describes some focus, methodology, or some other holistic attribute of a paper, and the paper matches that filter.
TIDBIT: The query describes some type of information about a paper that the user is looking for, and the abstract provides that information.

Here are some additional notes about what the query is looking for and how quotes should be selected if an abstract does match:
${A/B/C list text of abstract description responses}

Here is a numbered list of sentences from one paper abstract: (note that there are divider strings "<>" to mark quotable segments in the sentences)
1. ${sentence 1 part 1 <> sentence 1 part 2 <> ...}
2. ${sentence 2 part 1 <> sentence 2 part 2 <> ...}
3. ${sentence 3 part 1 <> sentence 3 part 2 <> ...}

Answer these questions:
A. Does this paper abstract match the query?
B. Why does it match or not match?
C. If it matches, which sentence number would provide the best quote that explains why the query matches the abstract or give the info the query is looking for? Write this as a single digit. If the query did not match, write NONE.
D. Write the quote from that sentence (the most relevant substring of the sentence) in double quotes. A quote must start and end at dividers. Do not include any divider strings in the quote. Keep the quote short. If the query did not match. write NONE.

in a font that was large enough on the screen that a finger held horizontally at arms length approximately covered only one row of text (one sentence). Since tracking eye gaze across users' vertical scrolling was possible while tracking eye gaze during horizontal user scrolling was not, participants were not allowed to scroll horizontally. As a result, some sentences were truncated by virtue of extending past the end of the screen. To simplify participants' potential visual recognition of the effects of the ordering feature, we only showed the first two chunks of sentences, which still form grammatically valid sentences by construction, as described in Section 4.1.2. We recognize this reduces the ecological validity of the experiment, but believe the results are still informative at the conceptual level.

F AbstractExplorer System Components

(a)

(b)

(c)

Figure 28: `Methodology` + `Participants` sentences as viewed in the Cross-Sentence Relationship pane. (a) vertically aligns in the middle of the chunk type tuple. (b) aligns to the left of the chunk tuple. (c) aligns on the beginning of the sentence.

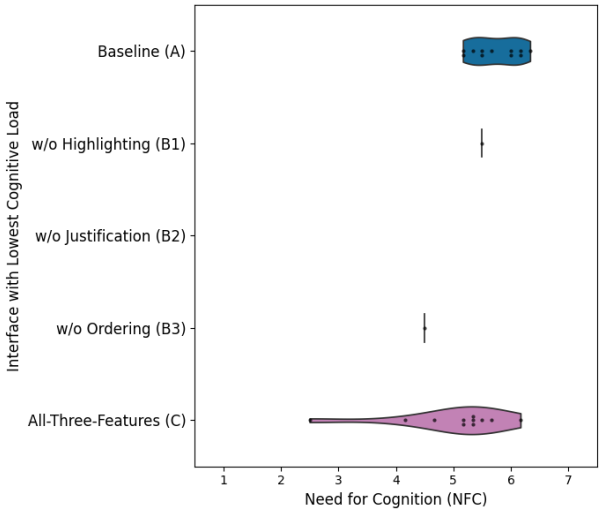


Figure 29: Distribution of participants' NFC scores by condition that gave the participant the least cognitive load. Note that B2 did not provide any participant with the lowest cognitive load, and so there is no data to show for it.

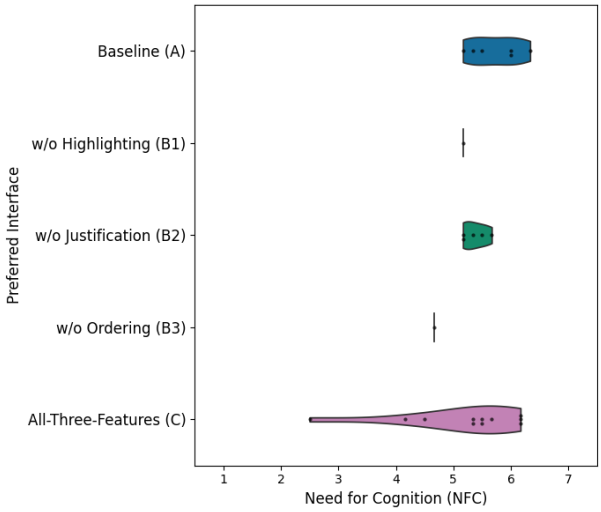


Figure 30: Distribution of participants' NFC scores by preferred condition

G Ablation Study Results

(a) Ablation Condition: Without Ordering (B3)

(b) All-three-features (C)

Figure 31: Gaze plots showing eye movements of P16, a higher NFC participant, across two interface conditions

H Summative Study Qualitative Data Analysis Procedure

We conducted a qualitative analysis of user study transcripts and survey responses using a Grounded Theory approach [8]. First, the lead researcher collected a list of participants' behaviors, approaches, reflections on their experience, and feedback about the interface. The researcher then systematically coded this data, revisiting the data multiples times and refining the codes to ensure consistency and coherence. Through this process, high-level themes were identified and organized using affinity diagramming. Once the thematic structure was finalized, the researcher gathered supporting evidence for each theme and synthesized the findings, which were reviewed by the research team to ensure agreement on the results.

I Summative User Study Tasks

I.1 Task 1: General Reading

Imagine you are writing a survey about the papers published at CHI 2024. You'll need to develop a comprehensive understanding of as much of the research published at CHI 2024 as possible. Given the time constraints, you've decided to focus solely on paper abstracts.

Goal:

(1) Skim as many paper abstracts as possible
(2) Develop a mental model of the variety of
   (a) contributions,
   (b) methodologies,
   (c) problem domains, and
   (d) study results

   within the papers at the conference.
(3) Bookmark sentences you find relevant or interesting.
(4) Look for 3 emergent patterns—both in content and style—over one or more of these aspects in the CHI 2024 paper abstracts (We understand you may not be able to address every aspect).

Feel free to use any feature of the interface except the "filter" feature. Please share your screen, and you may think aloud if it helps you verbalize your observations at the end. You can optionally use this doc as a place to take notes.

I.2 Task 2: Focused Reading

This part is similar to Part 1, ***but this time, focus on a subset of paper abstracts that are related to your work or of personal interest.***

Goal:

(1) Skim as many paper abstracts as possible. Feel free to engage more closely with abstracts that capture your attention.
(2) Develop a deeper understanding of the distribution of
   (a) contributions,
   (b) methodologies,
   (c) problem domains, and
   (d) study results

   within this subset of papers.
(3) Bookmark sentences you find relevant or interesting.
(4) Look for 3 emergent patterns—both in content and style—over one or more of these aspects in the paper abstracts that you choose to read (We understand you may not be able to address every aspect).

Again, feel free to use any feature of the interface. Please share your screen, and you may think aloud if it helps you verbalize your observations at the end. You can optionally use this doc as a place to take notes.

J Summative User Study Surveys

J.1 Pre-study Survey

J.1.1 Demographics.

(1) What is your Participant ID
   (Given to participant by study coordinator)
(2) What is your gender?
   – Male
   – Female
   – Non-binary
   – Prefer not to disclose
(3) What is your age?
   – Under 18 years old
   – 18-24 years old
   – 25-34 years old
   – 35-44 years old
   – 45-54 years old
   – 55-64 years old
   – 65-74 years old
   – 75 years or older
   – Prefer not to disclose
(4) On a scale from zero to ten, please select your level of proficiency in reading English.
   – 0 - none
   – 1 - very low
   – 2 - low
   – 3 - fair
   – 4 - slightly less than adequate for current role
   – 5 - adequate for current role
   – 6 - slightly more than adequate for current role
   – 7 - good
   – 8 - very good
   – 9 - excellent
   – 10 - perfect
(5) What is the highest degree or level of school you have completed?
   – High school graduate
   – Bachelor's degree
   – Masters' degree
   – Doctorate degree
   – Other...
(6) How would you describe your research position?
   – Undergraduate student
   – Master student
   – PhD student
   – Post-doctoral researcher
   – Faculty
   – Academic research staff
   – Industry researcher
   – Other…

J.1.2 Familiarity with CHI.

(1) How many CHI (or UIST/CSCW) conferences have you attended?
   – 0
   – 1
   – 2
   – 3
   – 4
   – 5+
(2) Did you attend CHI 2024?
   – Yes
   – No
(3) Are you actively preparing a manuscript for CHI (or UIST/CSCW) or a similar venue?
   – Yes
   – No
(4) How would you rate your familiarity with CHI (or UIST/CSCW)?
   – Scale: 1 (Not familiar at all) to 7 (Very familiar)
(5) How would you rate your knowledge about what kind of writing patterns are present in abstracts in CHI (or UIST/CSCW)?
   – Scale: 1 (Not knowledgeable at all) to 7 (Very knowledgeable)
(6) How would you rate your confidence about being able to write an abstract for CHI (or UIST/CSCW)?
   – Scale: 1 (Not confident at all) to 7 (Very confident)

J.1.3 Need for Cognition (NCS-6). The following questions are on a 7 point likert scale from "not at all like me" to "very much like me."

(1) I would prefer complex to simple problems.
(2) I like to have the responsibility of handling a situation that requires a lot of thinking.
(3) Thinking is not my idea of fun.
(4) I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
(5) I really enjoy a task that involves coming up with new solutions to problems.
(6) I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

J.2 Post-Study Interview Questions

(1) How did it go? Can you tell me more about your experience?
(2) Is there anything else you want to share? Any other feedback on the interface/tasks/ anything?

J.3 Post-study Survey

J.3.1 NASA-TLX.

(1) How mentally demanding was the task?
   – Scale: 1 (Low mental demand) to 7 (High mental demand)
(2) How physically demanding was the task?
   – Scale: 1 (Low physical demand) to 7 (High physical demand)
(3) How hurried or rushed was the pace of the task?
   – Scale: 1 (Not rushed at all) to 7 (Very rushed)
(4) How successful do you think you were in accomplishing the task?
   – Scale: 1 (Failure) to 7 (Perfect)
(5) How hard did you have to work to accomplish your level of performance?
   – Scale: 1 (Not hard at all) to 7 (Very hard)
(6) How insecure, discouraged, irritated, stressed, and annoyed were you when accomplishing the task?
   – Scale: 1 (Not really) to 7 (Highly)

J.3.2 Familiarity with CHI.

(1) After using the interface, how would you rate your familiarity with the conference venue?
   – Scale: 1 (Not familiar at all) to 7 (Very familiar)
(2) After using the interface, how would you rate your knowledge about what kind of writing patterns are present in abstracts in this venue?
   – Scale: 1 (Not knowledgeable at all) to 7 (Very knowledgeable)
(3) After using the interface, how would you rate your confidence about being able to write an abstract for this venue?
   – Scale: 1 (Not confident at all) to 7 (Very confident)

J.3.3 Interface Helpfulness.

(1) How helpful was the interface for exploring a ***greater number*** of papers along a given aspect than you typically would?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(2) How helpful was the interface for exploring a ***broader range*** of papers along a given aspect than you typically would?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(3) How helpful was the interface for ***discovering*** papers that you might have missed otherwise?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(4) How helpful was the interface for ***comparing*** different papers along a given aspect?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(5) How helpful was the interface for forming a more clear understanding of the ***relationships*** between different papers along a given aspect?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(6) How helpful was the interface for familiarizing with the different ***problem domains*** studied at CHI?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(7) How helpful was the interface for familiarizing with the different kinds of ***contributions*** presented at CHI?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(8) How helpful was the interface for familiarizing with the different ***methodologies*** used at CHI?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(9) How helpful was the interface for familiarizing with the ***distributions*** over problem domains, methodologies, contributions, etc. at CHI?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)
(10) How helpful was the interface for recognizing ***common patterns*** in how research is described at CHI?
   – Scale: 1 (Not helpful at all) to 7 (Very helpful)

J.3.4 Feature Usefulness.

(1) How would you rate the usefulness of the Pre-Defined Aspects feature (or the ability to see a particular slice of all paper abstracts)?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(2) How would you rate the usefulness of the coloring of sentence segments based on their roles in the sentence?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(3) How would you rate the usefulness of the order-by-structure overview?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(4) How would you rate the usefulness of the vertical-alignment by the first segment in the pattern pair?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(5) How would you rate the usefulness of the vertical-alignment by the second segment in the pattern pair?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(6) How would you rate the usefulness of the left-justified alignment?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(7) How would you rate the usefulness of the User-defined Aspects feature?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(8) How would you rate the usefulness of the search feature?
   – Scale: 1 (Not useful at all) to 7 (Very useful)
(9) How would you rate the usefulness of the bookmark feature?
   – Scale: 1 (Not useful at all) to 7 (Very useful)

J.3.5 Open-Ended Questions.

(1) What did you like about this interface?
(2) What did you not like about this interface?
(3) What did you wish you had in this interface?
(4) Would you see yourself using this interface in a real-world setting? Why or why not?
(5) What would need to change or improve for this interface to be useful in your daily work/life?