Accelerating Scientific Paper Skimming with Augmented Intelligence Through Customizable Faceted Highlights

RAYMOND FOK, University of Washington, USA
LUCA SOLDAINI, Allen Institute for AI, USA
CASSIDY TRIER, Allen Institute for AI, USA
ERIN BRANSOM, Allen Institute for AI, USA
KELSEY MACMILLAN, Allen Institute for AI, USA
EVIE CHENG, Allen Institute for AI, USA
HITA KAMBHAMETTU, University of Pennsylvania, USA
JONATHAN BRAGG, Allen Institute for AI, USA
KYLE LO, Allen Institute for AI, USA
MARTI A. HEARST, UC Berkeley, USA
ANDREW HEAD, University of Pennsylvania, USA
DANIEL S. WELD, Allen Institute for AI & University of Washington, USA

Scholars need to keep up with an exponentially increasing flood of scientific papers. To aid this challenge, we introduce Scim, a novel intelligent interface that helps scholars skim papers to rapidly review and gain a cursory understanding of its contents. Scim supports the skimming process by highlighting salient content within a paper, directing a scholar's attention. These automatically-extracted highlights are faceted by content type, evenly distributed across a paper, and have a density configurable by scholars. We evaluate Scim with an in-lab usability study and a longitudinal diary study, revealing how its highlights facilitate the more efficient construction of a conceptualization of a paper. Finally, we describe the process of scaling highlights from their conception within Scim, a research prototype, to production on over 521,000 papers within the Semantic Reader, a publicly-available augmented reading interface for scientific papers. We conclude by discussing design considerations and tensions for the design of future skimming tools with augmented intelligence.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Intelligent reading interfaces, skimming, highlights, scientific papers

Authors' addresses: Raymond Fok, rayfok@cs.washington.edu, University of Washington, USA; Luca Soldaini, lucas@allenai.org, Allen Institute for AI, USA; Cassidy Trier, cassidyt@allenai.org, Allen Institute for AI, USA; Erin Bransom, erinbransom@allenai.org, Allen Institute for AI, USA; Kelsey MacMillan, kelseym@allenai.org, Allen Institute for AI, USA; Evie Cheng, eviec@allenai.org, Allen Institute for AI, USA; Hita Kambhamettu, hitakam@seas.upenn.edu, University of Pennsylvania, USA; Jonathan Bragg, jbragg@allenai.org, Allen Institute for AI, USA; Kyle Lo, kylel@allenai.org, Allen Institute for AI, USA; Marti A. Hearst, hearst@berkeley.edu, UC Berkeley, USA; Andrew Head, head@seas.upenn.edu, University of Pennsylvania, USA; Daniel S. Weld, danw@allenai.org, Allen Institute for AI & University of Washington, USA.
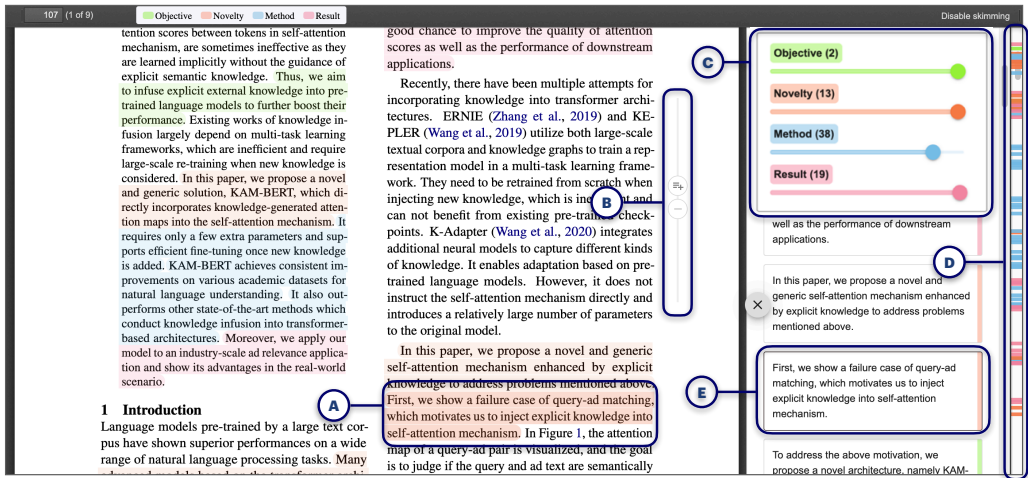
Fig. 1. Scim is an intelligent reading interface for skimming scientific papers. To help readers develop a broad overview of content in a paper, Scim intelligently highlights passages (A). The passages are colorized to indicate the rhetorical role of the passage, i.e., whether it describes the research's objectives, novelty, methods, and results. Highlights are distributed throughout the text to support a holistic skim. Readers can request additional (or fewer) highlights by using paragraph-local (B) and paper-wide (C) controls. To understand where to find information of a certain kind, readers can glance at highlight markers in the scroll bar (D). Readers can also collect an overview of the paper by reviewing highlighted passages in a sidebar (E).

1 INTRODUCTION

With the rise of knowledge work and a contemporaneous explosion of information, experts are expected to sift through and make sense of large volumes of rapidly evolving information. One domain where this trend is particularly pronounced is scientific research. Researchers spend a tremendous amount of effort staying up to date with the literature. They do so by regularly undertaking the tasks of foraging for papers, skimming or reading those deemed most relevant, and integrating knowledge gained from reading into their personal records.

Skimming is a critical task, and requires researchers to quickly review the contents of a paper to develop a cursory understanding of its contents. While faster than reading, skimming achieves a coarser view of papers' contents. With the shift of scientific publishing from paper to digital online publications, the practice of skimming has become yet more widespread [46, 76]. Despite the pervasiveness of skimming as a practice for reviewing papers [64], skimming is not easy [20, 54]. Skimming may devolve into reading should a reader find themselves drawn into the details of a passage. Even for experienced readers, skimming requires attention to make strategic choices of what to read, where, and when to stop reading.

In this paper, we investigate how an intelligent user interface can help both novice and expert researchers skim scientific papers more efficiently. Today, AI-powered techniques are increasingly used in tools for searching over the scholarly literature (e.g., [2, 5]) and in scientific reading applications (e.g., [3, 27, 47]). This paper explores how intelligent tools can facilitate the task of skimming, the seam between searching and reading.

As a starting point, we ask how judicious use of automatic highlighting can be presented in a tool to help readers direct their attention while they skim. To gain inspiration for designing such a tool, we conducted formative studies with researchers, including interviews, observations, and pilot studies of prototype highlighting tools. In these studies, we found that readers desired highlights that cover diverse content, are evenly distributed across a paper, and capture important paper content. These studies reveal a tension between reader expectations and system design, because it is not always possible to highlight according to passage importance while achieving a desirable distribution of highlights. Readers also desired some influence over the quantity and distribution of highlights within a paper.

We incorporated these insights into the design of Scim,[1] a prototype, intelligent user interface for skimming scientific papers (Figure 1) and subsequently into skimming features within Semantic Reader, a production-grade reading tool (Figure 6). To address readers' needs around highlighting, Scim highlights passages in the following ways. First, passages are highlighted with distinct colors for each of four diverse kinds of content sought by readers: research objectives, novel aspects of the research, methodology, and results. Second, Scim aims to support an evenly-distributed skim of a paper, highlighting passages in a way that most paragraphs contain at least one highlighted sentence. Finally, Scim lets readers customize the number of highlights in a paper, both across an entire paper and within individual paragraphs.

We conducted a sequence of studies to evaluate Scim's utility as an intelligent skimming tool. First, we performed a lab usability study to understand how Scim affects readers' ability to search for specific kinds of information in a paper. When using Scim, readers located the desired information in significantly less time compared to a standard document reader, with comparable effort and accuracy. Second, to understand more realistic usage, we conducted a two-week-long diary study. In this study, readers found Scim particularly useful when skimming text-dense passages with few visuals, or when skimming a paper that fell outside their area of expertise. Scim became more usable over time as readers became accustomed to the highlights. The study also suggests how skimming assistants could be improved in the future, for instance by highlighting passages that provide background for later highlighted passages, and integrating highlights with the typographical emphases authors may have already provided, such as boldface font and text formatting.

Before deploying skimming capabilities to an even larger population of users, we conducted another round of iterative design with several aims. First, we sought to make the skimming features more inclusive for blind and low vision users with marginal flags, user-controllable color intensities, and a screen-reader compatible side bar. Second, we evaluated and improved the quality of the salient passage classifiers, while re-engineering them to scale to process millions of documents. We then conducted a large-scale randomized controlled trial to understand skimming usage and check if skimming might decrease important metrics such as the number of papers read or the number of return visits. We found that 19% of users exposed to skimming capabilities used the features, and with no significant effect on guardrail metrics. As a result, skimming was enabled for all users in Semantic Reader.

The Scim system and its accompanying studies offer a vision of how applications can help readers skim texts like scientific papers with intelligent highlighting. Altogether, this paper contributes:

- Design goals to guide the design of intelligent, highlight-based skimming user interfaces, based on formative interviews and preliminary usability studies of a prototype tool.
- Scim, an augmented reading interface that aims to support scholars in skimming scientific papers by highlighting passages in a way that balances importance, diversity, and distribution of content, and affording control at both the paper and paragraph level.

---

[1]Code available at https://github.com/rayfok/scim

- A reference implementation of SCIM's end-to-end paper processing pipeline, including a language model for classifying salient sentences fine-tuned using a data programming approach, and heuristics for improving accuracy and achieving well-distributed highlights.
- Insights into the strengths and limitations of SCIM and AI-powered highlights based on a lab usability study, longitudinal diary study, and production-scale deployment.

2 RELATED WORK

In this section, we first introduce motivating insights about the process of skimming, and then we review tools and techniques that have been introduced to support skimming.

2.1 The Skimming Process

In the literature, skimming is characterized as a form of rapid reading in which the goal is to get a general idea of the text or visual content, typically by focusing on information relevant to one's goals and skipping over irrelevant information [53, 64]. Skimming is a necessary and useful skill for researchers. As the number of published papers increases year over year and papers have moved from print into digital media, scholars have tended toward reading more papers and spending less time on each, likely doing so by skimming [46, 76].

The psychology literature describes skimming as a cognitively demanding task. In this task, readers incrementally build a mental model of the text and integrate information across sentences as they read [61, 64, 75]. Generally, readers are not accurate at identifying goal-relevant information within text. Skimming is also physically demanding—limitations in the oculomotor system, which is responsible for controlling eye movements, preclude rapid, accurate placements of eye gaze for extended periods of time, such as when a reader skims a long document [52, 53].

Amidst the challenges of skimming, success is often determined by a reader's ability to "satisfice" [20, 21, 65]. Satisficing is a skim reading strategy where a reader sets a threshold of how useful information should be to deserve their attention, and if a unit of text falls below that threshold, they skip to the next unit of text. Studies have found that readers tend to spend more time at the beginning of paragraphs, the top of pages, and the beginning of documents [20], perhaps in part because this information is often believed to have high relevance. When grounded in particular information needs, skimming for rapid high-level comprehension is also often complemented by scanning, comprised of saccadic movements through a text in search of specific pieces of information.

One study of skimming for scientific document triage found readers were hasty and incomplete, with readers scrolling through documents quickly and paying attention to highly visual content and section headers [49]. Scientific documents are laden with visual content, typographical cues (e.g., italicized, bold, or colored text), and structural information. Readers draw on document features to support rapid comprehension via these macro- and micro-structures [12, 40, 50] and visual content [33, 86]. In this paper, we explore how automated assistance may support skimming by cueing readers towards significant sentences that might otherwise be missed. SCIM's use of highlighting lets readers continue to pay attention to traditional visual and structural landmarks, while also heeding the passages highlighted by the skimming assistant.

2.2 Tools for Reading and Skimming

Researchers have long sought to equip readers with tools to support and augment their cognition while reading documents. The nascent days of human-computer interaction saw the introduction of augmented reading interfaces to support the reading process, including fluid documents that provided contextual access to supplemental information between lines of text [14], fluid hypertext [88], visualizations for social annotations within papers [28], and affordances for annotating papers and

jumping readers to passages of interest [25, 68]. Since then, several approaches have been proposed to support the various aspects of reading, such as document navigation and comprehension.

2.2.1 Modified Scrolling Interactions. One line of research sought to facilitate the rapid exploration of long documents by modifying the behavior of reading interfaces during scrolling. Applications of content-aware scrolling were used to redefine the presentation order of content within a document [32], provide pseudo-haptic feedback when scrolling past relevant information [35], and dynamically resize document headings within paper thumbnails in a document viewer [10]. The Spotlights project implemented an attention allocation technique which pinned headings and figures as static overlays to a document as it was continuously scrolled [41].

2.2.2 Typographical Cueing. Another approach involved augmenting reading interfaces with typographical cues, e.g., highlighting. Studies in cognitive psychology have found visual cueing mechanisms can be effective in focusing reader attention [16] and improving retention of material [23, 67]. The Semantize system used highlights to visualize sentiment within a document, and underlined words with positive or negative sentiment in different colors [82]. The ScentHighlights system used highlights to identify conceptually relevant text based on a user's query [17]. The HiText technique introduced dynamic graded highlighting of sentences within a document in accordance with their salience [85]. Modern reading interfaces also commonly support readers in marking regions of interest with a document with highlights or free-text annotations. The pervasiveness of highlighting as a technique for drawing readers' attention can be attributed to the von Restorff isolation effect, which states an item isolated against a homogeneous background will be more likely to be attended to and remembered [30, 79]. Studies have since found evidence of this effect on the visual foraging behavior of readers, finding that highlights attract about half of the total number of fixations within a document, and readers' eyes are often drawn to them [16]. In this paper, we explore how typographical cues such as highlighting can augment the paper skimming process. In contrast to existing systems, we introduce an approach to generate these highlights automatically, and further organize the suggested highlights into semantically meaningful facets to facilitate machine-supported navigation throughout a paper.

2.2.3 Document Augmentations. Beyond typographical cues, other reading interface augmentations exist to specifically support the reading of scientific papers. For instance, online paper providers like Springer, PubMed, and Semantic Scholar provide readers with in-context citation information. Experimental systems have linked document text to marks within charts [38] and cells within tables [34], generated on-demand visualizations based on text within the paper [4], transformed static visualizations with animated [26] or interactive [51] overlays, augmented citations with cited paper [60] and personalized contexts [15], integrated segments of authors' talk videos into corresponding paper passages [36], and provided in-context definitions for nonce words [27]. We design Scim with inspiration from many of these prior reading interfaces, presenting a novel augmentation technique to support the process of skimming papers by visualizing information scent through the identification and highlighting of salient sentences.

2.2.4 Summarization. An alternative method to skimming a full paper is to read a shortened representation of the paper's content in the form of a summary. An author-provided summary is de facto included with each paper as an abstract, which researchers often read before continuing to the rest of the paper. Automated summarization has garnered significant interest from the natural language processing community, and many extractive and abstractive methods for generating summaries from long-form documents have been developed over the years (e.g., see [58, 69] for a survey). Some methods have even been proposed for generating extreme (single sentence) summaries, or TLDRs, from full papers [11].

However, summaries are often unsatisfactory. Despite recent advances, automated summaries remain error-prone, susceptible to hallucination [89], and unreliable as a standalone replacement for reading a paper itself. Furthermore, summaries do not provide readers with the ability to quickly interact with the full paper. As readers' goals and interests change while reading, they may wish to explore certain sections in further detail. Unlike summaries, augmented reading interfaces naturally retain the context of the paper. In this work, we present automatically extracted salient paper content as faceted highlights within a carefully-designed augmented reading interface to provide the interactivity and context lacking in standalone summaries.

3 DESIGN GOALS

To better understand how to design usable, intelligent skimming interfaces, we used an iterative design process that began with interviews and observations of academic researchers (referred to as *readers*), and continued into an evaluation of an early prototype of Scim. In this section, we describe that design process (Section 3.1). We then distill the lessons learned from this formative research into a set of design goals (Section 3.2) to guide the design and implementation of intelligent, highlighting-based skimming support tools.

3.1 Design Methodology

3.1.1 Formative interviews and observations. We conducted formative study sessions with eight readers (F1–8) to better understand how they skim scientific papers. All readers belonged to the target user group for Scim, and were either graduate students or academic faculty. Readers were first observed as they skimmed a paper of their choice. Then, readers were asked to describe their skimming process, including goals they held while skimming, strategies they employed, and any aspects of skimming they found difficult or tedious.

3.1.2 Prototype development and evaluation. A prototype of Scim was iteratively designed and developed based on our formative interviews and observations. While many kinds of tools could support skimming, our design exploration focused specifically on skimming aids which incorporate intelligent highlights.

The prototype was similar to the version of the Scim system described in Section 4, with a few differences. First, the prototype's highlighting policy was different, resulting in fewer highlighted passages, and a less uniform distribution of highlights. Second, the prototype had no paragraph-level or facet-specific controls for the number of highlights, but rather only global-level controls on the number of highlights and switches to turn on or off individual facets.

Two preliminary usability studies were conducted with this prototype. Thirteen readers (E1–13) were recruited from university mailing lists, and via direct outreach following purposive and snowball sampling approaches. Sessions in both studies were one hour in length and conducted on the Zoom platform. In both studies, readers skimmed papers with Scim for a limited amount of time and completed a task demonstrating their understanding of the paper, for instance outlining the paper or answering questions about the paper. Afterward, readers were asked to comment on their interactions with Scim and what aspects of the system required improvement.

3.1.3 Synthesis. One author conducted analyzed data from the formative study and preliminary evaluations following a thematic analysis methodology [6, Ch. 5]. Notes and transcripts from study sessions were analyzed for themes and supporting evidence. Themes were validated through discussion and review with a second author. Those themes that provided actionable guidance for design are reported in the next section.

3.2 Design Goals

We introduce seven design goals for intelligent highlight-based skimming interfaces, based our formative research.

D1. Augment readers' skimming practices. Readers described myriad strategies they already used to skim papers. One common strategy was to first read the abstract and introduction of a paper. Then, readers consulted other key material in the paper, including bulleted lists of contributions (F1, F4, F6), summaries of results (F1–3), and conclusions (F1, F3, F6, F7). Readers also employed strategies particular to their goals, the paper, or their level of comfort with the paper. Readers relied on various visible cues in the text to help them identify important information, including typographical cues (e.g., italics, boldface) (F3, F6), structural cues (e.g., section headers) (F2, F6), visuals (e.g., figures and tables) (F1, F2, F4, F6, F8), and text position (e.g., inspecting the first sentences of paragraphs) (F2, F3, F6). We believe skimming interfaces should not impede or replace these reading strategies.

D2. Highlight diverse kinds of content. Readers' skimming goals were diverse. For instance, some readers sought to learn specific techniques introduced in a paper (F1), and others wished to understand a paper's relationship to prior research, or discover new research directions (F2–4, F7). Some desired a high-level understanding suitable for discussing the paper with colleagues (F3, F7). These goals influenced readers' skimming strategies, leading them to look for answers to different sorts of questions. We suggest skimming tools should support readers' diverse goals by enabling review of varied aspects of paper contents.

D3. Support skimming in the lengthy middle sections of the paper. Readers noted that while one recommended strategy for skimming is to read the beginning and ends of paragraphs, important content may reside in the middle of paragraphs. When asked to skim, we often observed readers transitioning into a deep read of some passages in the paper (F1, F3, F5). We propose that skimming tools should help readers identify important passages which conventional strategies do not reach, such as content in the middle of paragraphs and in the middle of the paper.

D4. Minimize distraction. Without careful visual design, an augmented reading tool can occlude text or misdirect readers' attention. Our early prototypes incorporated a variety of text highlighting techniques, including underlines, lowlighting unimportant paper contents (inspired by ScholarPhi [27]), and highlighting text by setting its background color. Underlining was too subtle to consistently catch the reader's eye. Lowlighting tended to distract readers, requiring additional effort to read lowlighted content. Highlighting was chosen for its familiar use in documents, with the colors tuned to distinguish the categories of text and minimal contrast to avoid an unpleasant visual pop-out effect. We suggest that other designers similarly aim to minimize the visual distraction introduced by design interventions.

D5. Supply enough highlights. In our preliminary usability studies, readers often felt uncomfortable when they saw long, unhighlighted passages where they thought important information likely could be found. Some readers wanted to see highlights distributed more uniformly throughout the paper (as opposed to highlights concentrated primarily in an introduction or conclusion). We suggest the rule of thumb that a highlight should be provided around once per paragraph, and that readers should be able to request additional highlights in particularly dense passages.

D6. Provide accurate highlights. A side effect of introducing faceted highlights (where highlights are color-coded by their predicted rhetorical category) was that classification errors became obvious to readers, such as when a passage about results was labeled as being about methods. Readers found themselves distracted when the classification of a passage clashed with their expectations and became skeptical of the tool's capabilities (E11, E12). If skimming tools provide faceted highlights, it is especially important to classify these categories correctly.
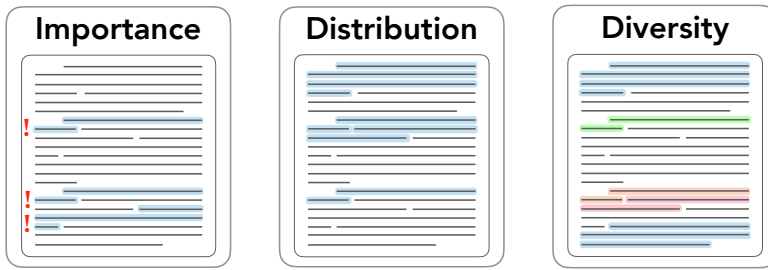
Fig. 2. Our formative research revealed that intelligent highlights need to do more than pointing readers to important content. They should also be *well-distributed* throughout a paper (D3, D5) and steer readers towards *diverse* content types (D2).

D7. Support user control and personalization. Readers desired more control over the amount of highlights shown by the prototype. Many suggested that the tool could help them fine-tune what was highlighted, either through manual adjustments, or with adaptive personalization of the highlights (i.e., responding to passages a reader has highlighted themselves or highlights they have deleted) (E5, E7, E8, E12).

A final takeaway from our formative research was that readers believed their comfort using intelligent highlights would change over time, as they became more familiar with the features, the colors associated with the highlights, and the accuracy of the highlights. One reader described this as the issue of "getting used to seeing highlights that aren't my own" (E13). This observation motivated our choice of a longitudinal diary study (Section 7) and production deployment (Section 8.3) as summative evaluation methods for Scim.

4 SCIM

We now describe the design of Scim, an interface that provides intelligent support for skimming scientific papers, and explain how particular aspects of the system address the design goals (D1–7) introduced previously.

4.1 Overview

A reader interacts with Scim as a tool that supports and augments their typical skimming process (D1). One common strategy for readers is to begin with a paper's title and abstract, followed by a piecemeal review of the paper. A reader can employ this strategy, and at the same time follow the highlights offered by Scim, which extend into parts of the paragraph that a reader may not notice otherwise (D3).

4.2 Faceted Highlights

Scim intelligently highlights a paper to direct a reader's attention to key passages (Figure 1.A). These highlights were tailored in three ways to support skimming.

Faceted. Because readers have different goals when skimming, Scim colorizes highlights according to facets of information (D2). To promote memorability we limit the number of facets to four. The specific set of facets was selected to encompass the kinds of information participants described in the formative study, balanced by the requirement that we could detect them reliably (D6), as described in the implementation section.

Numerous schemes exist for sentence-level classification of scientific literature into facets. Coarse-grained schemes classify sentences according to common section names from scientific
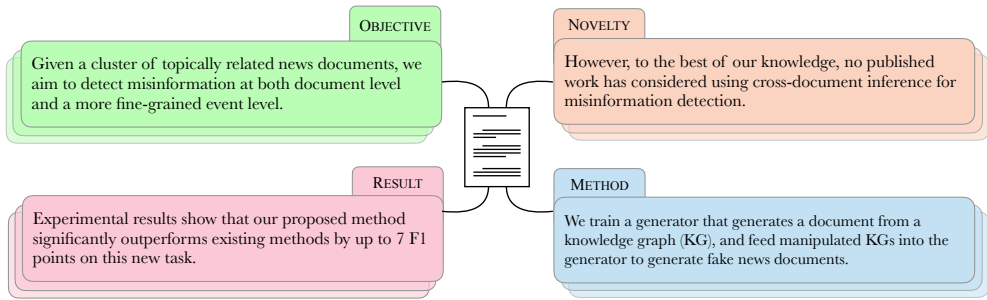
Fig. 3. Scim classifies and highlights four facets of information commonly found in papers: Objective, Novelty, Method, and Result. These facets aim to surface specific kinds of paper content that align with common skimming goals identified in formative research, reflecting design guideline *D2*. Above, we show example passages matching each of the four facets. The passages appear in Wu et al.'s scientific paper, "Cross-document Misinformation Detection based on Event Graph Reasoning" [83].

papers (e.g., [18, 29]) and consist of a small number of facets. Other fine-grained schemes rely on argumentative zones and conceptual structure (e.g., [43, 44, 77, 78]).

We derived a taxonomy of four facets by augmenting facets from of one coarse-grained schema for classifying scientific abstracts [18] with the "NOV_ADV" category (i.e., corresponding to sentences describing the novelty of a paper) from Argumentative Zoning [78]. As shown in Figure 3, Scim's four facets are: Objective, Novelty, Method, and Result, each of which is represented in Scim with its own color.

Low distraction. Text is highlighted using the familiar paradigm of a solid rectangular box behind the text, since this was observed in our evaluations of prototypes to be noticeable yet minimally distracting (D4). By using the same facet color mapping across papers, we hoped to foster a learned association device for each facet (D4). To help readers develop familiarity with highlight colors, the Scim interface header includes a legend mapping colors to facets.

Distributed. Since users of initial prototypes were concerned when they saw passages without any highlights (D5), we post-processed model predictions to distribute highlights approximately evenly throughout the paper.

4.3 Controls

Different readers may have very different goals in skimming, and even a single reader's goals may vary from one passage to the next. To provide flexibility in the skimming experience (D7), Scim provides two kinds of controls:

Paper-level controls. If a reader wishes to perform a cursory high-level skim of a paper, they can reduce the density of highlights, or to inspect a paper more closely, they can increase the density. If a reader does not wish to review a particular kind of content as they skim (e.g., they want to learn about the results of a study but not its methodology), the reader can disable highlights of a certain facet. Readers can control the density of highlights using facet-specific sliders found in Scim's side bar (Figure 1.C). As a reader drags a slider, they can see the effect on highlight density as highlights appear and disappear in the paper, markers appear and disappear in the scrollbar (Section 4.4), and a count of highlights change next to the slider.

Paragraph-level controls. If a reader desires additional highlights (e.g., if they have encountered a long paragraph of results they wish to skim more closely), Scim provides paragraph-level controls allowing them to rapidly access additional highlights. Readers can request more or fewer highlights
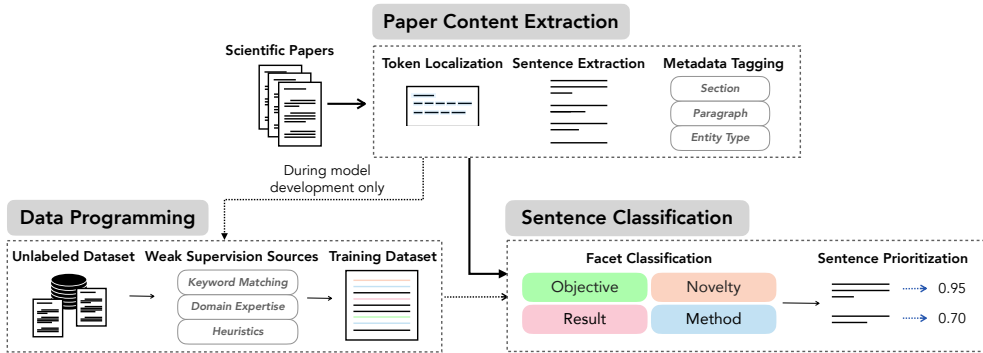
Fig. 4. Overview of Scim's paper processing pipeline. Scim takes as input a scientific paper in PDF format and then parses it into sentences with bounding boxes and other accompanying metadata. It then classifies sentences into one of four facets using a large language model fine-tuned via a data programming approach. Scim chooses which highlights to show by reconciling prediction weights with heuristics controlling highlight distribution and readers' preferences.

by hovering their mouse over a block of text, and then clicking on "+" and "-" buttons that appear in the margin (Figure 1.B). This feature provides quick and flexible control to complement paper-level controls, allowing a reader to request highlights precisely where they need them. For both paper- and paragraph-level controls, highlights are added and removed using a sentence prioritization score assigned during the document processing phase, as described in Section 5.

4.4 Scrollbar Annotations

A reader can discover where to skim in a paper by viewing highlight annotations in the scrollbar (Figure 1.D). This feature is inspired by edit wear and read wear affordances [28] and scrollbar annotations in code editors (e.g., [55]). When viewed together, these annotations can suggest paper structure, for instance implying if a paper has a particularly lengthy methods or results section, and where to find that information. The annotations also offer feedback to readers as they configure highlight density with Scim's controls.

4.5 Side Bar Display of Faceted Highlights

A reader may also review a paper's key passages by opening a side bar, which shows a compact list of all highlighted passages in order, grouped by paper section (Figure 1.E). This display updates dynamically as a reader configures the highlight density. A vertical colored bar appears next to each passage, providing a subtle and compact indication of the passage's facet. A reader desiring more context for a passage can click on it to scroll Scim to the passage's position in the paper, an interaction we call *context linking*.

5 IMPLEMENTATION

Scim was developed with an end-to-end document processing pipeline that supports intelligent highlighting. The main component of this pipeline is a pretrained language model, fine-tuned via weak supervision to identify and classify salient sentences within papers. An overview of the pipeline is presented in Figure 4.

5.1 Paper Content Extraction

Given an input PDF document, Scim uses VILA [70] and the open-source Multimodal Document Analysis (MMDA) library [22] to extract textual tokens, mathematical symbols, section headers, and metadata. Scim then segments the tokens into sentences, simultaneously merging bounding boxes for tokens into bounding boxes for sentences. Each sentence is labeled with its corresponding section header and paragraph index, attributes which are later used in the prioritization of sentences for which highlights should be shown.

5.2 Sentence Classification

To classify sentences into facets, we adapted the sequential sentence classification model from Cohan et al. [18], replacing the base BERT model with a pretrained MiniLM model [80, 81]. The MiniLM model considers surrounding context—up to a combined sequence length of 512 words or 10 sentences—when classifying a target sentence. We fine-tuned the model with the CSAbstruct dataset [18], a corpus of abstracts from computer science papers with manually-curated "gold" labels. Since sentences only came from paper abstracts, we ultimately found the model insufficient for classifying sentences from the body of papers, so we pursued additional fine-tuning as we describe in the next section.

5.2.1 Data Programming. We initially attempted to create manually-curated datasets of "gold" facet labels for sentences from full papers. However, this task was difficult to define, time-consuming, and expensive to execute during our pilot runs of the data collection process. As a result, we decided to extend our dataset with weak supervision following a data programming approach [63] to further fine-tune the model. Weak supervision provides a model-agnostic way to incorporate domain expertise into a model, and is sometimes a satisficing alternative to costly manual annotation. Weak supervision assumes access to a large unlabeled dataset and one or more labeling functions (e.g., heuristics encapsulating domain expertise, crowdsourcing, or knowledge bases), which are used to generate noisy labels for the dataset. While a collection of labeling functions can on their own serve as a classifier, we sought generalization beyond precise but potentially brittle labeling rules. We therefore employed a data programming paradigm to unify and de-noise the labeling functions, creating a weakly-labeled training set of sentences for downstream fine-tuning.

To build an unlabeled dataset for weak supervision, we extracted full paper sentences from the proceedings of NAACL 2018, 2019, and 2021, and ACL 2020–2022. In total, the dataset consisted of 3,051 papers with 606,400 unlabeled sentences. We then created weak supervision labeling functions consisting of heuristic rules and keyword matches to provide noisy facet labels for sentences in the dataset. For example, one rule-based supervision function detected sentence salience based on the presence of author intent via keywords such as "we", "our", "this paper," and their aliases. Other labeling functions relied on keyword matches to perform facet labeling. For example, sentences were weakly labeled as Novelty if any relevant keywords (e.g., "novel", "propose", "differ," and their aliases) could be found. We used Snorkel [62] to unify these labeling functions and output a dataset of weakly labeled sentences.

The dataset was further improved by incorporating weakly labeled negative sentences, selected from the full papers associated with the CSAbstruct abstracts used during the first round of training. We selected novel sentences by using the all-mpnet-base-v2 model [72] from the Sentence Transformers library [66] to score sentence similarity between full text and the abstract, and then labeling the most dissimilar sentences to the abstract, and which were not labeled with a facet in the prior phase, as not relevant for any facet (using an empirically chosen threshold cosine similarity of 0.25). Model fine-tuning was done on an NVIDIA A6000 GPU, using 0.1 dropout rate

and Adam optimizer [37] over 5 epochs, and 5e-5 learning rate. All parameters were determined using the CSAbstruct validation split.

5.2.2 Evaluation. We conducted a preliminary evaluation of Scim's sentence classification model over a set of 20 NLP papers. We recruited annotators from Upwork, an online freelancing marketplace. All hired annotators were required to have experience with NLP and scientific writing. Detailed instructions asked annotators to role play as a reviewer for a scientific communication magazine, tasked with creating abridged versions of scientific papers. Annotators were asked to identify significant, complete sentences within each of the 20 papers, and were paid $20 USD/hr.

Each paper took on average 20 minutes to annotate, and was annotated by three Upworkers using the PAWLS PDF annotation tool [59]. Sentences selected by at least two of the three annotators were considered as ground truth "significant sentences," and collected into a test set. On this test set, our classification model achieved an F1 score of 0.533, compared to an annotator-annotator F1 score of 0.725 (which we consider as a gold-standard, i.e., a performance ceiling, since there is inherent variability in which sentences annotators believe are significant for skimming). Our goal with this preliminary evaluation was not to evaluate whether we advanced the state-of-the-art in NLP, but rather to verify that the model reliably identified meaningful highlights for use in Scim.

5.3 Cleaning and Prioritizing Highlights

Scim's user interface selects which highlights to show using the predicted facet label, probability score, and other heuristics. One heuristic enforced consistency between facet labels and the section in which a sentence appeared (e.g., if a highlight appeared within a methods section, it had to be tagged with the "Method" facet; similar constraints were imposed for the "Novelty" and "Results" facets). Another heuristic prompted a more uniform distribution of highlights throughout a paper, prioritizing sentences within paragraphs which did not already contain other highlights.

5.4 User Interface Implementation

Scim is implemented as a web application built atop the PDF rendering platform pdf.js [56]. The system retains text markup already present in the paper which may support skimming, such as hyperlinks, clickable citations, bold and italicized text, and other visual cues provided by the authors. Scim's features including highlights, side bars, and controls were implemented as interactive React components incorporating widgets from the Material UI library [57].

6 STUDY 1: IN-LAB USABILITY STUDY

To assess Scim's usability and effectiveness for skimming, we conducted a three-phrase sequence of studies. This section describes the first phase, an in-lab usability study aiming to evaluate how Scim affects readers' speed of skimming and their ability to identify relevant information throughout a paper, in comparison to a baseline of unaided paper navigation. In Section 7, we describe the second phase, a longitudinal diary study building on insights from the usability study, aimed at understanding the real-world value and applicability of Scim over a two-week period. Finally, in Section 8.3, we detail the third phase, a deployment-scale evaluation with Scim's features integrated into the Semantic Reader, a publicly available augmented reading interface.

6.1 Study Design

For our usability study, participants were asked to complete a series of short tasks using both Scim and a standard document reader. This study sought to answer two research questions:

RQ1. *Does Scim enable readers to skim papers more quickly?*
RQ2. *How does Scim affect readers' ability to identify relevant information after a skim?*

6.1.1 Participants. We recruited 19 participants (8 male, 10 female, 1 non-binary) via university-affiliated mailing lists and Slack channels. We also conducted pilot studies with three additional participants, results of which we do not include in our analysis. Participants were required to have experience reading NLP papers, since they would be required to do so during the study. They ranged from 21 to 30 years of age, and included 11 PhD students, 5 master's students, 2 software engineers, and 1 industry researcher. Participants self-reported an average of 3.78 (on a 5-point Likert scale) for comfort with reading NLP papers, suggesting they were generally familiar with the type of literature used in the study. Participants were compensated $25 USD for their time.

6.1.2 Procedure. Participants first provided consent and then were led through a tutorial of Scim's features. The study used a within-subjects design, and consisted of three tasks, each with two sub-tasks, one for each of the two reading interface conditions—Scim and a baseline consisting of a standard document reader, specifically the built-in PDF viewer in the Chrome web browser. We designed the study to be completed in under one hour to limit participant fatigue. The studies were conducted remotely via Zoom. To minimize biases, we counterbalanced the order of the reading interfaces and papers used in each task. Below, we describe the three tasks.

- *Task 1*: Participants skimmed a paper and identified a passage in the paper that described a key feature (e.g., dataset creation or evaluation) of the paper. This task was intended to familiarize participants with the two interfaces, so we did not include any measures from this task in our analysis.
- *Task 2*: Participants skimmed a paper and answered two multiple-choice questions based on information found in the paper. Answers to these questions *could* be found in text highlighted by Scim. We hypothesized the main points highlighted by Scim should be easier to locate, and this task was designed to test that hypothesis.
- *Task 3*: Participants skimmed a paper and answered two multiple-choice questions based on information found in the paper. Answers to these questions *could not* be found in text highlighted by Scim. In contrast to Task 2, we hypothesized that Scim might prove a hindrance when finding information outside of the highlights, and this task was designed to check this concern.

Participants skimmed a different paper for each of the sub-tasks. The six papers for these tasks [1, 31, 42, 73, 74, 84] were selected from the proceedings of NAACL 2022, and corresponded to the following types: (1) technical papers introducing new datasets or metrics, (2) exploratory papers investigating the effectiveness of current tools and proposing new design guidelines, and (3) technical papers proposing novel language models for specific applications. Questions in Tasks 2 and 3 focused on aspects of a paper a reader might be interested in while skimming, such as evaluation metrics or the motivation behind a proposed method. Regardless of interface, participants were given multiple attempts and asked to skim until they answered correctly.

For each question, we used the following quantitative metrics:

- *Time* — The number of seconds taken by the participant to answer the question, from when the paper was first opened to when the correct answer choice was selected.
- *Accuracy* — A binary variable indicating whether the participant's first response to the question was correct.
- *Difficulty* — A five-point Likert scale variable indicating the participant's self-assessment of the following prompt: "I found the task difficult."

After completing all sub-tasks, participants were also asked to self-assess on a five-point Likert scale whether they found the tasks overall easier to complete with Scim and whether Scim's highlights were distracting during skimming.
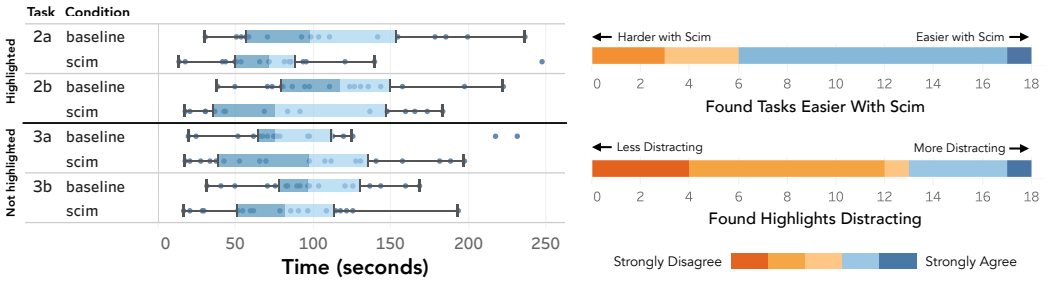
Fig. 5. **(Left)** Time taken by participants to complete each information seeking question in Tasks 2 and 3 of the in-lab usability study. Overall, participants answered questions more quickly with Scim than with a standard (baseline) document reader. **(Right)** Participants' subjective responses regarding the ease of completing information seeking tasks with Scim compared to a standard document reader, and whether they found Scim's highlights distracting.

6.1.3 Analysis. We compared readers' time, accuracy, and perceived task difficulty using linear mixed-effects models [45] with reading interface as a fixed effect, task and question number as nested fixed effects, and participant as a random effect. We first conducted *F*-tests for any differences across the interface conditions, and then we conducted post-hoc *t*-tests when appropriate for differences in the estimated fixed-effects between conditions.

6.2 Results

Participants answered questions more quickly with Scim (μ = 94.3s, σ = 74.9s) than with a standard document reader (μ = 117.7s, σ = 76.4s), a significant difference (F(1, 126) = 4.17, p < .05). The difference was more pronounced in Task 2, where the correct answer was located within one of the highlights (F(1, 54) = 4.84, p < .05): readers took an average of 93.8s with Scim (σ = 81.6s) versus 127.3s with the standard reader (σ = 77.8s). In Task 3, where the correct answer was not located within one of the highlights, there was no significant difference in time (F(1, 54) = 0.58, p = .45), with participants taking 94.8s with Scim (σ = 68.7s) versus 108.0s with the standard reader (σ = 74.7s).

There was no significant difference (F(1, 126) = 0.22, p = .64) in participants' accuracy with Scim (μ = 0.80, σ = 0.40) compared to a standard document reader (μ = 0.76, σ = 0.43). There was also no significant difference (F(1, 119) = 0.01, p = .92) in readers' perceived difficulty in answering questions with Scim (μ = 2.32, σ = 0.89, with 5.00 indicating strong difficulty) compared to a standard document reader (μ = 2.31, σ = 1.02). Altogether, the results show that Scim reduced the time it took for readers to seek information in papers, with no difference in accuracy or effort.

7 STUDY 2: LONGITUDINAL DIARY STUDY

Participants in our usability study noted that it would take some time to acclimate to a novel reading interface like Scim before they felt comfortable using it. To better understand realistic long-term use, we therefore also conducted a two-week long diary study. This study let readers use Scim for papers of their choice from a list relevant to their discipline, leading to alignment of their motivation with typical motivations for skimming. Participants could choose when they read, and for how long, as long as they skimmed using Scim at least once a day.

7.1 Study Design

We designed the diary study to provide insight into the following research questions:

RQ1. *What value can intelligent highlight-based skimming aids provide to researchers?*
RQ2. *How do researchers make use of skimming aids as they read?*
RQ3. *In what scenarios do researchers find skimming aids useful?*
RQ4. *What are the limitations of highlight-based skimming aids?*
RQ5. *What features should future intelligent skimming tools have?*

7.1.1 Participants. We recruited participants through university-affiliated mailing lists, Slack channels, and public posts from the authors' Twitter accounts. Participants were required to have prior experience reading or writing research papers. Preference was given to those with experience reading papers in the field of natural language processing (NLP), because the collection of papers we preprocessed for this study came from a recent NLP conference. A total of 12 participants were recruited for the study (6 male, 6 female). Two were master's students, and ten were PhD students. PhD students spanned a range of experience, with 1 first-year student, 3 second-year students, 2 third-year students, 3 fourth-year students, and 1 fifth-year student. No participants had participated in any of the prior lab studies. Participants were compensated $100 USD at the end of the study.

7.1.2 Reading Materials. Though Scim's pipeline was able to process arbitrary papers within a few seconds, we wished to reduce the time it took for participants to load papers during the diary study. As a result, we preprocessed a set of papers we felt would be exciting for participants to read, specifically the proceedings of NAACL 2022, one of the most recent and widely-read NLP conferences. We selected these papers since Scim had been fine-tuned primarily on NLP papers, and we expected it would perform appropriately for this collection. We also anticipated the NLP community would provide a broad audience from which we could recruit participants for a diary study. We also allowed participants to request other papers outside of this collection to read with Scim throughout the study; in total, an additional 10 NLP papers were preprocessed.

7.1.3 Procedure. The diary study consisted of three stages: a welcome session, a two-week long observational period, and an exit interview. During the welcome session, participants completed a tutorial of how to use Scim. They were given a few minutes to try out the interface, and to ask questions. Each participant was also shown the online diary (hosted in a Google Doc), and briefed on the protocol for recording their skimming experiences.

Then, during the observational period, participants were asked to spend 5–10 minutes each day, for 10 days, skimming at least one paper and completing a structured reflection in the diary.[2] On the first day of the observational period, participants skimmed papers using a standard (non-Scim) document reader for the first day to provide a point of comparison with Scim. During the subsequent nine days, they skimmed papers with Scim. Following each skimming session, they completed a diary entry, consisting of the following questions:

(1) Which papers did you skim today, and how long did you spend skimming each one?
(2) What highlights (if any) drew your attention to something you might have missed without the highlights?
(3) Did highlights help you skim this paper? Explain.
(4) List one or more ways the system could have helped you better skim this paper.

---

[2]Nearly all participants succeeded in completing 10 days of diary entries. Only 1 of 12 failed to complete all required entries; they completed only 7 of 10. When a participant fell behind in their diary entries, we sent them light email reminders.

Table 1. A summary of usage of Scim's features during the diary study. Notably, most readers used most features at least once. Use of the highlight controls varied widely, with some readers using them heavily (P6, P10, P12), and others less often (P1–3, P6, P7–9). All readers used the highlight browser on multiple occasions. Faceted highlights are omitted because we could not collect log data as to when readers looked at highlights.

Feature	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Highlight Browser	5	10	8	10	9	4	12	3	5	20	6	19
Global Highlight Controls	1	0	3	0	2	16	3	3	0	22	9	4
Local Highlight Controls	3	2	1	8	0	12	0	0	0	16	3	34
Context Linking	0	1	0	3	30	2	0	0	0	6	0	8

After the observational period, we conducted exit interviews with participants. They were asked to reflect on their experience using Scim in detail, including how it supported their skimming and opportunities for improvement.

7.1.4 Analysis. We conducted a thematic analysis on the qualitative data — diary entries and transcripts from exit interviews — following the approach described by Creswell and Poth [19]. One author identified significant excerpts from the diary entries and transcripts, and iteratively developed and refined a set of themes represented in the data. A second author validated the analysis by reviewing the themes, checking their alignment with the excerpts, and proposing revisions. A total of 177 responses to diary prompts were analyzed (participants left responses to some questions blank). We also instrumented and analyzed behavioral logs detailing interactions with Scim. In reporting results, we refer to participants with the pseudonyms P1–P12. The utterances presented below were edited to elide identifying information while preserving their meaning.

7.2 Results

Next, we present the findings of our diary study as they pertain to each of our research questions.

7.2.1 The value of Scim as a skimming aid **(RQ1)**.

For many readers, Scim helped with skimming by allowing them to focus their attention and attain a high-level understanding of the paper (P5, P6, P9, P10). Furthermore, Scim helped readers identify key concepts and review the main ideas of papers. P5 described Scim as guiding her to the important contributions of the papers she skimmed, and the highlights as offering a "gist of the paper beyond what was in the abstract."

And though highlights helped readers review the paper as a whole, they could also help them orient to specific aspects of a paper they wanted to understand. For instance, P1 and P2 both noted that the highlights helped them to understand the results of the paper more quickly, which are often quite dense and text-heavy. Scim's highlights also helped readers attend to interesting details in sections of papers they might have otherwise skipped over (P1, P4, P11). This was described as "slowing down" and skimming with greater care:

> This was a paper that is very light on methods and most content is about results, which I tend to skim over. So the highlights helped me slow down and slightly more carefully read a few places. (P4)

For some readers, skimming without Scim required two passes, first skimming a paper to identify relevant passages, and then re-reading passages of interest in greater detail (P5, P8). Scim could alleviate the need for multiple passes:

> *With highlights, I usually spend more time reading and understanding the highlighted content and skimming the other content. Without the highlight[s], I need to scan the entire content first, identify the critical points and then understand it. The highlights save me time in skimming the whole paper.* (P8)

*7.2.2   How researchers made use of SCIM **(RQ2)**.*
Usage of SCIM entailed usage of its constituent features of highlights, the highlight browser, controls, and context linking. All readers made use of most features at least once (Table 1). We surmise that the most frequently used feature was the highlights, for several reasons. First, the feature was always turned on. Second, highlights figured prominently in our conversations with readers, as evidenced by the rest of this section. And third, most readers reported that the highlights helped them find useful information during their daily readings (see Section 7.2.3).

While the predominant method of interaction with SCIM was likely to view highlights within the paper, a second commonly used feature was the highlight browser: all readers opened SCIM's side bar more than once, with the average reader opening it 9.3 times. Readers described the highlight browser as supporting navigation and providing a rapid understanding of paper contents (P7, P9, P10). It was also described as an "extractive summary" (P2). One reader thought the highlight browser provided a "better way to skim" in comparison to highlights, which at the time of their diary entry, they believed made the paper "difficult to read" (P7).

Nearly all readers used both global and local controls to configure the number of highlights. Global controls were typically used a small handful of times to achieve an acceptable density of highlights (which was then persisted into subsequent skimming sessions). Only a few readers adjusted the highlights via the global controls across multiple papers skimmed. When asked, readers typically reported that the default density of highlights was appropriate (P2, P6, P7). That said, most did adjust the number of highlights with paper-level controls at least once. Exit interviews confirmed that readers tended to tune the level of highlights to the preferred level on the first day of use. One participant asked for highlight controls with coarser options, for instance enabling them to toggle between one mode showing only the most important highlights, and another with many highlights for a deeper skim.

Readers seemed to use SCIM to augment, rather than replace, their existing skimming strategies. Readers reported directing their attention both to the highlights and to conventional paper landmarks like section headers and visual content. For example, P9 described their process as navigating through the main sections of a paper as they might in a typical skim, and then using the highlights to identify important information within those sections. P4 similarly described skimming using the combination of section headers and highlights.

For some readers, it took some time to become accustomed to using SCIM (P8, P10). One issue seemed to be developing trust in what was highlighted (P8). In their exit interview, P10 described their how their trust and interactions with SCIM evolved over the course of the study:

> *I feel like I just got more used to the highlights. … When I would see an objective highlight, I would trust it. I found the results highlights to be very helpful, so I would immediately focus on those. I would open the side panel right away instead of waiting during the end of the paper. I just got used to the tool, and I learned how to use it fast, depending on the paper and what I wanted to get from the paper.* (P10)

*7.2.3   Circumstances in which SCIM was useful **(RQ3)**.*
Overall, the intelligent highlights appeared to be useful during a majority of skimming sessions. In response to the diary question, "Did highlights help you skim this paper?" 74 of 105 (70.4%)

responses answered in the affirmative. There were a handful of circumstances in which readers reported SCIM as particularly useful.

One circumstance where SCIM was useful was in reading dense passages of text. The highlights made long passages that were absent of "visual support" such as figures more approachable (P3, P5). SCIM helped one reader skim a detailed experimental section and identify several important details which, due to the density of text, they "might have skipped if not for the highlights" (P5). Readers reported SCIM as helpful not just for dense passages, but also for papers that were text-heavy as a whole, such as survey papers (P5, P11).

Intelligent highlights were seen as useful to readers who sought information from papers on a topic they did not typically read about (P8, P10), assisting them in identifying and focusing on important paper content:

> For me it was also generally useful for reading papers that were a little out of my comfort zone. … In that case the highlighting helped me focus on, read, and conceptualize better certain parts of the methodology in order to better understand the conclusions. (P10)

The highlights also provided a summary of the paper in their own right. One reader described a situation where they were "not particularly interested in this paper." For them, the highlights served as "a summary" that they could read in lieu of looking closely at the paper (P4). This suggests an interesting possibility for intelligent highlights to help not just highly-motivated skimmers, but also those skimming papers in lower-motivation contexts.

7.2.4 Limitations of SCIM's model of intelligent highlighting **(RQ4)**.

Readers identified several ways that intelligent highlights might be extended to be made more useful. One concern was that highlighted passages sometimes lacked sufficient context to be understood alone (P1–3, P7–8, P11).

> When reading the highlights, the context is often missing. Sometimes it is just in the lines before and after, but sometimes we need to find it which then makes reading difficult as there is now more back and forth instead of a linear reading. (P7)

SCIM was designed with the hope that readers would look for such "context" in the surrounding text by simply moving their focus from highlighted to unhighlighted text. In practice, it could be disruptive for readers to seek out this context. Necessary context could appear just before or after the highlight in the paragraph, and in some cases even in other sections. For one reader, skimming highlights that lacked context therefore became a process that resembled "more back and forth instead of a linear reading" (P7).

Some readers desired tighter integration between SCIM's highlights and existing visual cues within a paper. While SCIM did not occlude or hide text that the author had emphasized (e.g., bolded text, section headers, or bulleted lists), this emphasized text was often not highlighted. As a result, readers discovered inconsistencies between the visual cues introduced by authors and the highlights suggested by SCIM (P2, P4, P5), such as bolded result statements or contributions in a list, which were not consistently highlighted.

Sometimes, text was highlighted in other unexpected and undesired ways. For instance, SCIM sometimes highlighted only one contribution from a list of bulleted contributions, when readers believed it should have highlighted all of them (P1, P7). SCIM was also unpredictable when highlighting passages that contained dense math notation (P1, P6, P11), and readers wished for highlights to apply to visual content like tables and figures (P2, P5, P7, P12).

7.2.5 Envisioning future intelligent skimming tools **(RQ5)**.

SCIM represents just one way in which intelligent assistance could support the paper skimming process, and readers described alternative ways that future tools could help them skim. For some

readers, Scim's highlights provided too much detail, particularly if they desired only a high-level understanding of the material (P6, P8). Readers suggested that an abstractive summarization of paper content (e.g., "with a bit of info pulled from tables/graphs/figures/examples" (P12)), could lessen the effort required to understand dense sections of papers (P1–2, P7–P8, P12). Recent large language models have achieved impressive advances in summarizing scientific texts, and future tools could leverage these models to augment the reading experience with abstractive summaries.

Readers also believed they could be aided with better tools for navigation. One reader desired the ability to use a paper's abstract or introduction as an index into related highlights in the rest of the paper (P2). Another reader wished to see the paper summarized in question-and-answer format, realizing they often sought answers to questions while they skimmed, such as "What are the research questions? What are the novelties/contributions of this study? What data/model/evaluation methods do they use? What are the main results? What are the limitations?" (P8). While Scim addresses these information needs through faceted highlights, future tools could support more conversational interactions between readers and the papers they skim.

8 DEPLOYING SCIM AT PRODUCTION SCALE

To evaluate the impact of automatically-generated highlights at a larger scale, we adapted features from Scim into Semantic Reader, a publicly-available augmented reading interface for scientific papers.[3] Through their usage of the online reading interface, we envisioned scholars could interact with the features naturally throughout their skimming process. As part of the integration, we iteratively refined Scim's features by incorporating feedback from participants in our prior two studies, and explored opportunities to make the design more accessible for individuals with visual disabilities. Finally, in scaling Scim's automatically generated highlights to several hundred thousand papers, we added several architectural improvements to improve the robustness and efficiency of the paper processing pipeline.

8.1 Refining Customizability and Improving Accessibility

Findings from the two prior studies showed that colored highlights are an effective visual cue for communicating distinct facets of information in the text. However, one limitation of such a design is that it remains inaccessible to individuals who use assistive technology such as screen readers, those with difficulties seeing low-contrast visuals, or users with certain types of color blindness. Seen through the lens of inclusive design, we believed this one-size-fits-all approach of interface augmentations created an inequitable skimming experience. To accommodate such users who may engage with the text using a screen reader, we ensured snippets in the side bar are fully screen reader accessible. Thus, these users may read with the same attention guidance augmentations that users without assistive technology would encounter. While we find that this solution is a step toward a more accessible scientific paper reading experience, we acknowledge that reading side bar snippets is not the same experience as viewing highlighted text as visual cues within the paper itself. Future work could enable screen readers to verbally announce the beginnings and ends of highlighting in-line, which could help preserve the context around each highlight.

Furthermore, we iterated on Scim's design to include three options that allow users to customize the presentation of skimming highlights based on their individual needs.

Margin flags. We added optional labels adjacent to each highlight in the paper's margin which could indicate a highlight's facet at a glance without relying on color. Users who may have difficulty remembering the color associations or who cannot distinguish between the colors could benefit from these margin flags, which convey facet information via text rather than color.
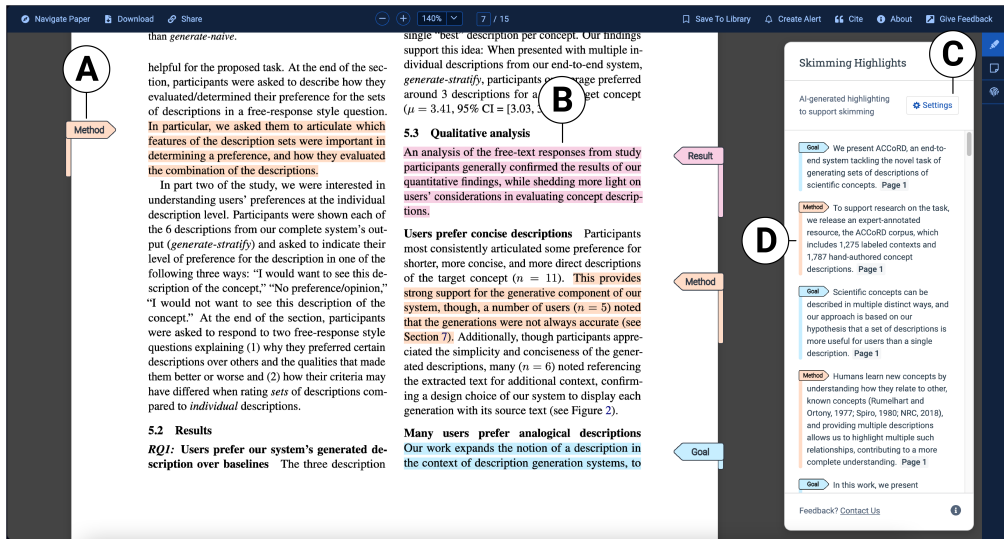

Fig. 6. We integrated SCIM's interface augmentations into a publicly-available reading interface for scientific papers, Semantic Reader, refined into the following set of features: margin flags which can be turned on or off (A); multi-faceted highlights (Goal, Method, and Results), the density and visual characteristics of which can be customized (B); user-controlled settings for further customizability of the interface (C); screen-reader accessible snippets containing the extractive text from each highlight (D).

Customizable highlight contrast and opacity. We provided controls that allow users to customize the opacity and contrast of the highlights. Some users may find the bright highlights distracting, while others may be unable to distinguish low contrast highlights against a paper's white background. These contrast controls are designed to accommodate both needs.

Coarse-grained highlight density controls. We simplified SCIM's original paper- and paragraph-level controls to reduce the effort required for users to determine the appropriate amount of highlights in a paper. Our diary study suggested that the predetermined number of highlights in SCIM was a reasonable amount to support skimming, and that few users cared to adjust the highlights at a finer-grained level while reading. As a result, we provided two global adjustments for highlight density, allowing users to either select less or more highlights than the default.

8.2 Improving the Robustness of the Paper Processing Pipeline

In addition to highlights which were irrelevant, lacked sufficient context, or were classified with an incorrect facet, participants in the diary study also sometimes noted inaccuracies in the highlights that appeared to stem from errors in the document processing component. We found participants were particularly sensitive to these errors, which were easily distinguishable at a glance, and the prevalence of these errors became more noticeable as we scaled the number of papers for which highlights were available. For instance, some highlights were incorrectly split across multiple columns or pages in a paper, or on undesirable paper content such as footnotes, page numbers, and references. The presence of these errors may distract scholars as they skim, and even worse, potentially undermine their confidence in quality of the highlights overall.

8.2.1 Analysis of errors in highlights from document processing. To assess the frequency and types of potential errors that may appear in the displayed highlights, we conducted a two-phase evaluation.

First, three members of the research team annotated and coded errors within a random, traffic-biased set of 30 papers (uniformly sampled from papers available on the Semantic Reader with more than two views over last 90 days). Annotators were asked to identify and categorize errors as they encountered them.

Annotators found a diversity of document processing issues that diminished the quality of the extracted highlights. These errors included:

(1) **Non-body text errors** — The model responsible for extracting the main body text from paper PDFs occasionally included other text on the page as part of the main body, such as the paper's preamble (e.g., title, author, affiliations), page numbers, headers, footnotes, and references. It also sometimes identified text from within figures and tables; we did not intend to highlight these visual or tabular components within papers in Scim (though future aids may consider the utility of highlighting such content). Including non-main body text within an extracted sentence could degrade the accuracy of the subsequent highlight classification model, and also draw attention to irrelevant content within a page, visually distracting readers.

(2) **Sentence splitting errors** — Sentences were sometimes incorrectly split on non-sentence boundary punctuation (e.g., splitting on the period in "In Section 4.1, we first introduce...") or on column and page boundaries. These resulted in highlights that seemed to begin or end abruptly, rather than capturing a complete sentence.

(3) **Layout errors** — If the token bounding boxes were not accurately extracted, the highlights could appear shifted on the page. This resulted in some highlights that appeared over empty regions of a page, across sentence boundaries, or overlap figures.

In our error analysis, we noticed that papers with an arXiv source had fewer errors and generally higher quality data, since these papers were often had cleaner PDFs compiled from LaTex source documents. We also found that the paper processing pipeline tended to perform better on computer science papers, possibly due to biases in the training data. As a result, we conducted a second annotation on 97 computer science papers with an arXiv source. For this set, we observed 60.8% of papers had at least one non-body text error, 54.4% of papers had at least one sentence splitting error, and 5.2% of papers had at least one layout error. No errors were detected in 35.1% of papers.

Overall, this error analysis revealed the existence of a substantial quantity and diversity in document processing errors that could propagate into the reading interface, introducing undesirable distractions for users while skimming. It further suggests how building effective and usable intelligent reading interfaces requires not only furthering the capabilities of models for language understanding, but also addressing the seemingly small and nuanced errors that may emerge when processing long and complex PDF papers, impacting the augmented reading experience.

8.2.2 Additional structural and linguistic predictors. To address these issues, we enhanced the original paper processing pipeline (Section 5) with additional heuristic and learned predictors.

(1) **Margin predictor** — This uses a clustering algorithm to detect margins of a page containing the body text of a paper. Identifying these regions helps to remove extraneous symbols within the margins such as page numbers or watermark text (e.g., identifiers on arXiv preprents).

(2) **Word predictor** — This fixes words that get split between rows (e.g., with hyphenation).

(3) **Block predictor** — This introduces heuristics to improve VILA block predictions (originally described in Section 5.1) for non-body text components of papers. Accurately identifying these blocks helps in ensuring certain types of content are disregarded by the highlight selection model, such as text within references, figures, tables, equations, section headers, preamble (e.g., title, authors), headers, footers, and footnotes.
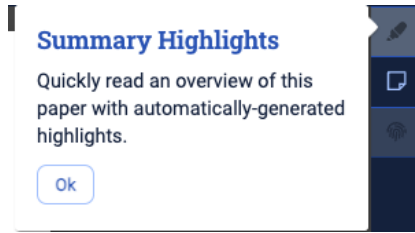
Fig. 7. When users in the deployment study first encountered a paper with skimming features available, the capabilities were introduced with a popup window in the upper right, directing the user to controls for opening the sidebar with highlights and customizable settings.

(4) **Sentence predictor** — This merges sentences that span across columns and pages, and filters out overly long sentences (>160 tokens).

(5) **Sentence perplexity predictor** — This estimates the perplexity of sentences using a unigram language model. Specifically, sentences with an average log word probability of less than -20 are removed.[4] This is to ensure sentences that are mostly numbers (e.g., sentences parsed from tables) get filtered out.

(6) **Sentence header predictor** — This matches each sentence with its appropriate (possibly nested) section headers. It improves upon the original header predictor in Scim (which was originally used in a heuristic ensuring consistency between facet labels and sections; see Section 5.3), and is used to render the section header for each snippet in the side bar.

In an updated pipeline, these predictors are combined with the original faceted highlight detection and classification language models (described in Section 5.2). The specific language models used for classification could eventually be substituted with more performant models, e.g., generative large language models, which even with minimal prompt engineering and simple few-shot prompting strategies have exhibited impressive capabilities on many language understanding tasks [8, 9].

8.3 Study 3: Deployment Study

Between August 20, 2023 and October 6, 2023, we ran a randomized controlled trial on 350,000 recent computer science, arXiv papers on Semantic Scholar's Semantic Reader platform [48]. Fifty percent of users experienced a baseline version of the Semantic Reader and the other 50% were given the reader with production skimming features as shown in Figure 6. On their first visit to a skimming-enabled paper, the feature was explained with a one-time, dismissable popup window (Figure 7), directing the user to controls to open the highlight panel and adjust controls.

Over the trial period, 20,200 unique users read 42,700 documents with skimming available. 19% of these users actively used the skimming capabilities, and an average of 12.7 times per user — with a mixture of clicking on paper highlights or flags, hovering or clicking on side bar snippets, and adjusting settings. 197 users (5% of users engaged with skimming) adjusted settings, such as the number, density or type of highlights displayed. 49 users (1.3%) turned the feature off. No statistically significant difference was seen between the control and treatment groups with respect to key metrics: papers viewed and scroll distance per user, nor in Semantic Scholar site-wide metrics such as the number of returning sessions per user or the number of discovery clicks per user. As a result, the test was concluded and skimming was enabled on 521,000 papers (as of October 22, 2023) for all users of Semantic Reader.

---

[4]Log word probability is calculated using word frequencies from the Web 1T Ngram corpus [7]. See [71] for similar details.

Enabling Scim at production-scale across a large corpus of papers served to highlight a number of challenges in designing effective intelligent user interfaces, particularly in the context of scholarly support tools. Qualitative feedback collected throughout the deployment study suggests that a large opportunity remains in improving the models used in identification of relevant content to highlight. The current models we developed remain imperfect and also appeared to perform worse on out-of-distribution papers, such as those outside of computer science. Beyond improving the underlying automated techniques, a key tension in the design of Scim was in augmenting papers in a way that accelerated rather than distracted from readers' own skimming processes. As a result, significant effort was dedicated to iteratively tailoring the distribution, density, and visual characteristics of the highlights to facilitate effective skimming support. The nontrivial number of data processing improvements and guardrails that were incorporated into the pipeline before deployment also indicates the importance of mechanisms for identifying and resolving errors throughout complex, multi-component systems, as errors that propagate from any component, even beyond the natural language processing models themselves, can negatively influence users' perceived quality of these intelligent tools. Our design process also revealed the importance of customizability within these interfaces, for instance in allowing users to tailor the information presentation according to their specific needs, research backgrounds, and familiarity with a particular paper. Future systems can also continue to ensure equitable access by involving inclusive design principles as a core consideration from the onset, ensuring that these tools are universally usable.

9 DISCUSSION

In this paper we sought to understand how intelligent user interfaces could support readers in skimming scientific papers. We designed and evaluated Scim, a tool that augments the skimming experience with automatic faceted highlighting of paper content. A lab study showed Scim reduced the amount of time to complete short information seeking tasks in scientific papers, with no significant difference in readers' self-reported task difficulty. In a subsequent diary study, we observed how researchers might use Scim in more realistic settings. Readers believed Scim helped them develop a high-level understanding of papers and determine which passages to skim or skip. Scim was seen as useful for skimming dense texts and papers from unfamiliar domains. Next, we consider these results, amidst limitations of our research approach, and suggest implications for the agenda of developing intelligent skimming aids.

9.1 Skimming versus Scanning

In the lab study, the information seeking tasks we used were intended to measure participants' speed and accuracy while skimming. However, we noticed that participants often exhibited behavior that more closely resembled *scanning*. Unlike skimming which involves a rapid high-level comprehension of text, scanning is a subtly different reading process concerned with locating specific pieces of information within a body of text. In the lab study, some participants utilized conventional scanning strategies such as "Control+F", using keywords in the question or answer choices as prompts. While this strategy was typically unsuccessful since the questions were designed with text which nullified this keyword-based scanning strategy, this behavior suggests participants did not necessarily attempt to skim the paper to gain an understanding of the paper to answer the questions, but instead scanned the text for keywords to locate the exact answer to the question. For readers like these, our results may be less indicative that Scim helps with the skimming process but rather the scanning process.

9.2 Supporting Experts and Novices

Scim was designed to help experienced skimmers get more out of skimming. It was not, however, designed to help inexperienced skimmers develop proficiency with skimming. Skimming assistants for teaching skimming may or may not have a lot in common with Scim. As AI models become increasingly adept at identifying salient paper content, those models may be useful not just to identify important content, but to coach skimmers to find this content as well. The development of such AIs and accompanying interfaces poses the interesting design problem of ensuring that readers have a consistently productive experience learning the essentials of skimming, regardless of their prior background or the documents they skim.

9.3 Risks to Attention

One risk of introducing technologies like Scim is that they may have unintended consequences for a reader's attention. Skimming requires significant attention to understand the idiosyncrasies and nuances within papers. Tools that augment the paper with assistive affordances may inadvertently deplete a reader's limited attention if they impose additional cognitive burden, as might be the case if it highlights content in a way readers do not expect. The tools could also lead a reader to pay less attention to the paper as they skim, once they are no longer required to drive the skimming process themselves. Furthermore, should such skimming assistants be readily available, readers may not be incentivized to deeply read papers, but rather enticed to skim by the presence of highlights.

Skimming aids like Scim therefore need to be designed in a way that respects a reader's attention and the value of deeply reading. They should be accurate and reliable. They should be deployed alongside studies that understand their effect on readers' engagement with texts. Furthermore, they should be developed and deployed in tandem with tools that support and encourage researchers to deeply read, and in general encourage good reading practices within the research community. For instance, affordances for skimming might be made available only when searching quickly through multiple papers or while reading on the go, but then are limited in scenarios befitting a deeper read.

9.4 Limitations of Highlights

Without sophisticated controls and affordances enabling more goal-driven or personalized skimming, highlights only present a single pathway through a paper. Highlighting is a cueing mechanism that directs reader attention and assists in the foraging of information, but it does not address other aspects of the information gathering and sensemaking processes scholars undergo while skimming. As readers suggested in the diary study, there are numerous ways in which skimming aids could provide more holistic support. For instance, augmented intelligence could offer additional context for highlighted passages, integration with existing visual cues, highlighting of visual content, complementary usage of abstractive summarization, or enhanced navigation support.

9.5 Future Work

Through our design and evaluation of Scim, we identified several opportunities to further explore the potential of augmented intelligent interfaces for skimming scientific papers.

9.5.1 Improving Highlight Quality. The effectiveness of Scim, like that of many other AI-infused user interfaces, is limited by the accuracy of its underlying AI models. Future research could improve the utility of highlights by building upon our models with other computational approaches. One promising direction could involve long-form summarization models sensitive to our highlight-relevant design guidelines, and another could leverage the discriminative capabilities of general-purpose large language models. Features including a paper's hierarchical structure, author-cued content, or visual content may also be leveraged to improve highlighting accuracy. Furthermore, our

findings suggest improvements to PDF processing are necessary to improve the user experience for augmented reading interfaces like Scim. Minor errors in Scim's PDF processing resulted in content like footnotes, section headers, tables, or figures being concatenated with paper sentences, which led both to poor classification of those passages and highlights that included disparate content. Resolving such issues in PDF processing could significantly improve classification accuracy and the cleanliness of the presentation of highlights within a reading interface.

9.5.2 Social Annotations. Our studies suggested that readers may be hesitant to adopt an augmented reading interface like Scim due to distrust in the AI's ability to provide the relevant highlights. Some readers mentioned that they might trust highlights created by other people (e.g., fellow researchers) more than those generated by AIs. Could social annotations be used to produce better highlights? Social highlights have been extensively explored in other settings, including studies on the effect of social annotation on attention within public multimedia content [13], news reading [39], and education [24, 87, 90]. Modern online publishing platforms such as Medium also show "popular highlights," suggesting the potential for social highlights in reading tools for scientific literature as well. Such affordances could port nicely into a system like Scim, and the coordination of social highlights with AI-generated highlights can be an interesting area for future research. More broadly, the development of community-driven features, such as enabling readers to share and discover highlighted content curated by their peers within the same field, suggests how collective intelligence can be leveraged to personalize and enrich the skimming experience.

9.5.3 Personalization of Skimming Aids. As readers continue to interact with augmented reading interfaces, we envision an opportunity for AI-infused systems to learn from repeated reader interactions, providing personalized and proactive reading support to help mitigate undesirable cognitive overhead introduced by these systems. They could also be tailored to readers' individual reading behaviors by considering their experience reading papers within a particular field, their typical information needs, or their goals for reading a particular paper. Providing this interactivity will empower users to fine-tune the systems to their own preferences, enhancing the relevance of the content surfaced over time. Furthermore, while the current design of Scim highlights content according to four common facets, future skimming aids could allow users to dynamically define and refine these facets beyond this initial set. And unlike the current paper-agnostic facets, additional benefit may lie in augmenting the reading interface with more informative or fine-grained facets, dependent on the specific content in a paper users are actively reading.

10 CONCLUSION

In this work, we presented the design, development, and evaluation of Scim, an intelligent augmented reading interface that supports scholars in skimming scientific papers with faceted, evenly-distributed, minimally intrusive, configurable highlights. A lab usability study found participants located information in papers more quickly with Scim than with a standard document reader. In a two-week-long diary study, participants remarked ways in which Scim supported a rapid, high-level skimming of papers. Based on feedback from these two studies, we iteratively refined Scim's design and architecture, before conducting a production-scale deployment study by integrating Scim's features into the Semantic Reader, a public augmented paper reading interface. Scim was found to be particularly useful for dense passages of text and for papers from unfamiliar domains. Altogether, these studies suggest the potential for intelligent tools to support researchers in skimming scientific literature, and highlight opportunities for further model improvement and personalization.

ACKNOWLEDGMENTS

REFERENCES

[1] Mitchell Abrams and Matthias Scheutz. 2022. Social Norms Guide Reference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 1–11.

[2] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Association for Computational Linguistics, New Orleans, LA, USA, 84–91.

[3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 74 (Sept. 2023), 38 pages.

[4] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 661–671.

[5] Jöran Beel and Bela Gipp. 2009. Google Scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*, Vol. 1. Rio de Janeiro, Brazil, 230–241.

[6] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes.* Morgan & Claypool Publishers, San Rafael, CA, USA.

[7] Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. https://catalog.ldc.upenn.edu/LDC2006T13

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[10] George Buchanan and Tom Owen. 2008. Improving Skim Reading for Document Triage. In *Proceedings of the Second International Symposium on Information Interaction in Context* (London, United Kingdom) *(IIiX '08)*. Association for Computing Machinery, New York, NY, USA, 83–88.

[11] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4766–4777.

[12] Patricia L. Carrell. 1985. Facilitating ESL Reading by Teaching Text Structure. *TESOL Quarterly* 19, 4 (1985), 727–752.

[13] Scott Carter, Elizabeth Churchill, Laurent Denoue, Jonathan Helfman, and Les Nelson. 2004. Digital Graffiti: Public Annotation of Multimedia Content. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1207–1210.

[14] Bay-Wei Chang, Jock D. Mackinlay, Polle T. Zellweger, and Takeo Igarashi. 1998. A Negotiation Architecture for Fluid Documents. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA). Association for Computing Machinery, New York, NY, USA, 123–132.

[15] Joseph Chee Chang, Amy X. Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S. Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 737, 15 pages.

[16] Ed H Chi, Michelle Gumbrecht, and Lichan Hong. 2007. Visual foraging of highlighted text: An eye-tracking study. In *Proceedings of the 12th International Conference on Human-Computer Interaction*. Springer, Beijing, China, 589–598.

[17] Ed H. Chi, Lichan Hong, Michelle Gumbrecht, and Stuart K. Card. 2005. ScentHighlights: highlighting conceptually-related sentences during reading. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, San Diego, CA, USA, 272.

[18] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained Language Models for Sequential Sentence Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Online, 3691–3697.

[19] John W Creswell and Cheryl N Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications, USA.

[20] Geoffrey Duggan and Stephen Payne. 2009. Text Skimming: The Process and Effectiveness of Foraging Through Text Under Time Pressure. *Journal of experimental psychology. Applied* 15 (09 2009), 228–242.

[21] Geoffrey B. Duggan and Stephen J. Payne. 2011. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*. Association for Computing Machinery, Vancouver, BC, Canada, 1141.

[22] Allen Institute for Artifical Intelligence. 2022. MMDA - Multimodal Document Analysis. https://github.com/allenai/mmda

[23] Robert L. Fowler and Anne S. Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3 (1974), 358–364.

[24] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1555–1564.

[25] Jamey Graham. 1999. The Reader's Helper: A Personalized Document Reading Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, PA, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 481–488.

[26] Tovi Grossman, Fanny Chevalier, and Rubaiat Habib Kazi. 2015. Your Paper is Dead!: Bringing Life to Research Articles with Animated Figures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, Seoul, Republic of Korea, 461–475.

[27] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 413, 18 pages.

[28] William C. Hill, James D. Hollan, Dave Wroblewski, and Tim McCandless. 1992. Edit Wear and Read Wear. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, CA, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 3–9.

[29] Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*. Association for Computational Linguistics, Hyderabad, India, 381–388.

[30] Roderick Hunt. 1995. The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review* 2 (1995), 105–112.

[31] Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 158–168.

[32] Edward W. Ishak and Steven K. Feiner. 2006. Content-aware scrolling. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, Montreux, Switzerland, 155.

[33] S. Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37 (2007), 83–84.

[34] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating Document Reading by Linking Text and Tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, Berlin, Germany, 423–434.

[35] Juho Kim, Amy X. Zhang, Jihee Kim, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Content-aware kinetic scrolling for supporting web page navigation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, Honolulu, HI, USA, 123–127.

[36] Tae Soo Kim, Matt Latzke, Jonathan Bragg, Amy X. Zhang, and Joseph Chee Chang. 2023. Papeos: Augmenting Research Papers with Talk Videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 19 pages.

[37] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization.

[38] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting References Between Text and Charts via Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Toronto, Ontario, Canada, 31–40.

[39] Chinmay Kulkarni and Ed Chi. 2013. All the News That's Fit to Read: A Study of Social Annotations for News Reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2407–2416.

[40] Natasha Lacroix. 1999. Macrostructure construction and organization in the processing of multiple text passages. *Instructional Science* 27, 3/4 (1999), 221–233.

[41] Byungjoo Lee, Olli Savisaari, and Antti Oulasvirta. 2016. Spotlights: Attention-Optimized Highlights for Skim Reading. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, San Jose, CA, USA, 5203–5214.

[42] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A Dataset for Multi-Span Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 1250–1260.

[43] 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 (April 2012), 991–1000.

[44] Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the Conceptualisation and Zoning of Scientific Papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Valletta, Malta, 2054–2061.

[45] Mary J. Lindstrom and Douglas M. Bates. 1990. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* 46, 3 (1990), 673–687.

[46] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation* 61, 6 (Jan. 2005), 700–712.

[47] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. arXiv:2303.14334 [cs.HC]

[48] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. arXiv:2303.14334 [cs.HC]

[49] Fernando Loizides and George Buchanan. 2009. An Empirical Study of User Navigation during Document Triage. In *Research and Advanced Technology for Digital Libraries*. Vol. 5714. Springer, Berlin, Heidelberg, 138–149.

[50] Tonja Machulla, Mauro Avila, Pawel Wozniak, Dillon Montag, and Albrecht Schmidt. 2018. Skim-Reading Strategies in Sighted and Visually-Impaired Individuals: A Comparative Study. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference* (Corfu, Greece) *(PETRA '18)*. Association for Computing Machinery, New York, NY, USA, 170–-177.

[51] Damien Masson, Sylvain Malacria, Edward Lank, and Géry Casiez. 2020. Chameleon: Bringing Interactivity to Static Digital Documents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Honolulu, HI, USA, 1–13.

[52] Michael E. Masson. 1982. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8, 5 (1982), 400–417.

[53] Michael E. J. Masson. 1983. Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition* 11, 3 (May 1983), 262–274.

[54] Martha J. Maxwell. 1972. Skimming and Scanning Improvement: The Needs, Assumptions and Knowledge Base. *Journal of Reading Behavior* 5, 1 (March 1972), 47–59.

[55] Microsoft. 2022. Scroll bar map mode and bar mode - Visual Studio (Windows). https://docs.microsoft.com/en-us/visualstudio/ide/how-to-track-your-code-by-customizing-the-scrollbar

[56] Mozilla. 2022. PDF.js. https://github.com/mozilla/pdf.js

[57] MUI. 2023. Material UI. https://material-ui.com/

[58] Ani Nenkova and Kathleen McKeown. 2012. A Survey of Text Summarization Techniques. In *Mining Text Data*, Charu C. Aggarwal and ChengXiang Zhai (Eds.). Springer US, Boston, MA, 43–76.

[59] Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. PAWLS: PDF Annotation With Labels and Structure. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 258–264.

[60] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719.

[61] David N. Rapp and Paul van den Broek. 2005. Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science* 14, 5 (2005), 276–279.

[62] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* 11, 3 (Nov. 2017), 269–282.

[63] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data Programming: Creating Large Training Sets, Quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3574–3582.

[64] Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. 2016. So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychological Science in the Public Interest* 17, 1 (May 2016), 4–34.

[65] William R. Reader and Stephen J. Payne. 2007. Allocating Time across Multiple Texts: Sampling and Satisficing. *Human–Computer Interaction* 22, 3 (Aug. 2007), 263–298.

[66] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992.

[67] John P. Rickards. 1980. Notetaking, Underlining, Inserted Questions, and Organizers in Text: Research Conclusions and Educational Implications. *Educational Technology* 20, 6 (1980), 5–11.

[68] Bill N. Schilit, Gene Golovchinsky, and Morgan N. Price. 1998. Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Los Angeles, CA, USA) *(CHI '98)*. ACM Press/Addison-Wesley Publishing Co., USA, 249–256.

[69] Athar Sefid and C. Lee Giles. 2022. SciBERTSUM: Extractive Summarization for Scientific Documents. In *Document Analysis Systems: 15th IAPR International Workshop* (La Rochelle, France). Springer-Verlag, Berlin, Heidelberg, 688–701.

[70] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. 2022. VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *Transactions of the Association for Computational Linguistics* 10 (2022), 376–392.

[71] Luca Soldaini and Kyle Lo. 2023. *peS2o (Pretraining Efficiently on S2ORC) Dataset*. Technical Report. Allen Institute for AI. ODC-By, https://github.com/allenai/pes2o.

[72] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-Training for Language Understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1414, 11 pages.

[73] Katherine Stasaski and Marti Hearst. 2022. Semantic Diversity in Dialogue with Natural Language Inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 85–98.

[74] Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 1075–1085.

[75] Craig S. Tashman and W. Keith Edwards. 2011. Active Reading and Its Discontents: The Situations, Problems and Ideas of Readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2927–-2936.

[76] Carol Tenopir, Donald King, Sheri Edwards, and Lei Wu. 2009. Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. *Aslib Proceedings* 61, 1 (2009), 5–32.

[77] Simone Teufel and Marc Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 28, 4 (Dec. 2002), 409–445.

[78] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 1493–1502.

[79] Hedwig von Restorff. 1933. Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung* 18, 1 (Dec. 1933), 299–342.

[80] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2140–2151.

[81] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 485, 13 pages.

[82] Alan J. Wecker, Joel Lanir, Osnat Mokryn, Einat Minkov, and Tsvi Kuflik. 2014. Semantize: visualizing the sentiment of individual document. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. Association for Computing Machinery, Como, Italy, 385–386.

[83] Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document Misinformation Detection based on Event Graph Reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 543–558.

[84] Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Oluwasanmi Koyejo. 2022. A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, WA, USA, 587–599.

[85] Qian Yang, Gerard de Melo, Yong Cheng, and Sen Wang. 2017. HiText: Text Reading with Dynamic Salience Marking. In *Proceedings of the 26th International Conference on World Wide Web Companion*. Association for Computing Machinery, Perth, Australia, 311–319.

[86] Ji Soo Yi. 2014. QnDReview: read 100 CHI papers in 7 hours. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, Toronto, Ontario, Canada, 805–814.

[87] Dongwook Yoon, Nicholas Chen, Bernie Randles, Amy Cheatle, Corinna E. Löckenhoff, Steven J. Jackson, Abigail Sellen, and François Guimbretière. 2016. RichReview++: Deployment of a Collaborative Multi-Modal Annotation System for Instructor Feedback and Peer Discussion. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, CA, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 195–205.

[88] Polle T. Zellweger, Bay-Wei Chang, and Jock D. Mackinlay. 1998. Fluid Links for Informed and Incremental Link Transitions. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space* (Pittsburgh, PA, USA) *(HYPERTEXT '98)*. Association for Computing Machinery, New York, NY, USA, 50–57.

[89] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing Quantity Hallucinations in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2237–2249.

[90] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful Classroom Deployment of a Social Document Annotation System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1883–1892.