Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries

Hai Dang
hai.dang@uni-bayreuth.de
Department of Computer Science, University of Bayreuth
Bayreuth, Germany

Karim Benharrak
Karim.Benharrak@uni-bayreuth.de
Department of Computer Science, University of Bayreuth
Bayreuth, Germany

Florian Lehmann
florian.lehmann@uni-bayreuth.de
Department of Computer Science, University of Bayreuth
Bayreuth, Germany

Daniel Buschek
daniel.buschek@uni-bayreuth.de
Department of Computer Science, University of Bayreuth
Bayreuth, Germany

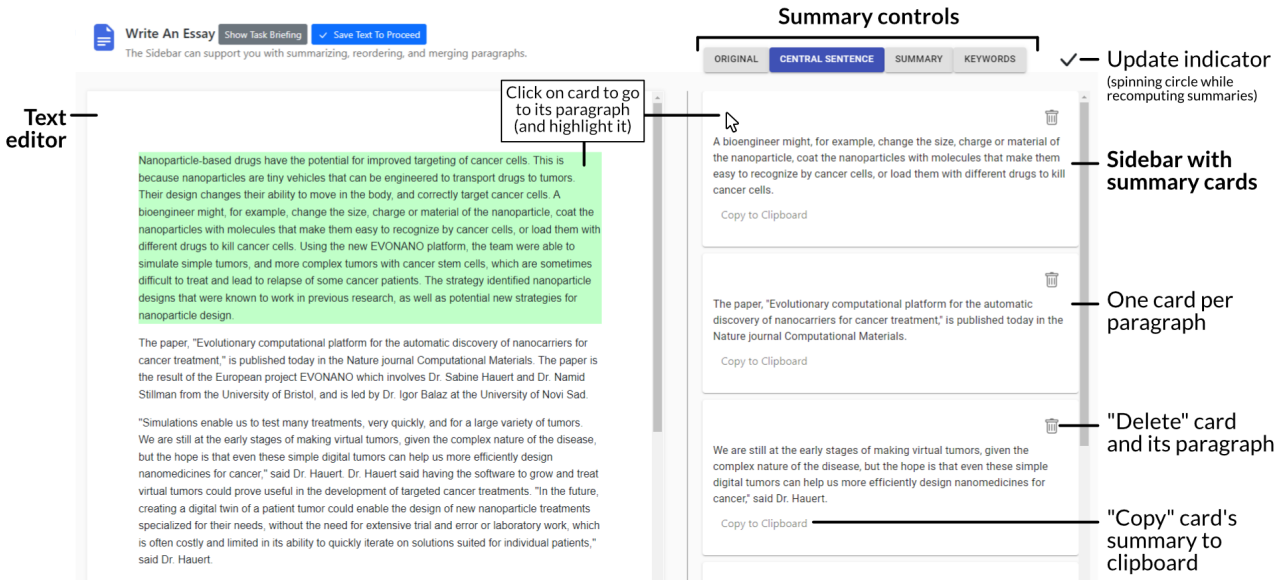

Figure 1: Our UI (in the final version for study 2) with text editor (left) and a sidebar (right), which shows one card per paragraph. Typing is possible only in the editor. The cards' content is controlled with the top buttons (original, central sentence, summary, keywords). These summaries are updated while typing. Editor and sidebar can be scrolled independently. Clicking a card scrolls to its paragraph in the editor and highlights it (in green for 1 second).

ABSTRACT

We propose a text editor to help users plan, structure and reflect on their writing process. It provides continuously updated paragraph-wise summaries as margin annotations, using automatic text summarization. Summary levels range from full text, to selected (central) sentences, down to a collection of keywords. To understand how users interact with this system during writing, we conducted two user studies (N=4 and N=8) in which people wrote analytic essays about a given topic and article. As a key finding, the summaries gave users an external perspective on their writing and helped them to revise the content and scope of their drafted paragraphs. People further used the tool to quickly gain an overview of the text and developed strategies to integrate insights from the automated summaries. More broadly, this work explores and highlights the value of designing AI tools for writers, with Natural Language Processing (NLP) capabilities that go beyond direct text generation and correction.

UIST '22, October 29-November 2, 2022, Bend, OR, USA

ACM ISBN 978-1-4503-9320-1/22/10…$15.00
https://doi.org/10.1145/3526113.3545672

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Text input**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Text documents, text summarization, semantic zoom, reverse outlining, Natural Language Processing

ACM Reference Format:

Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3526113.3545672

1 INTRODUCTION

Writing is important and ubiquitous for many people. Numerous tools have been developed to aid writers in their creative process by automatically performing spell and grammar checking or suggesting continuation phrases to increase human performance and productivity. Yet, good writing goes beyond correct spelling and grammar – it needs to both convey the writer's intention and address the reader's needs. Writing is a complex cognitive activity, as pointed out by Flower and Hayes [12]. In particular, writing interweaves various sub-activities. Revision is one of them, which is important but often challenging for several reasons: (1) *Practically*, investing into preparations for writing or repeated revisions often stands at a tension to finishing the work. (2) *Organizationally*, as the body of text grows, it is increasingly difficult to oversee the draft and maintain a big picture during writing. (3) *Skill-related*, it is tempting, in particular for amateurs, to focus on local revisions (e.g. word-level) instead of scope and structure [19].

Despite these challenges, providing assistance for the specific activity of text revision has remained under-explored in HCI research, as pointed out in the community [4, 50, 51]. Recent work here focuses on the aforementioned local revisions (e.g. word replacement [16, 26], error correction [1, 11, 58]), increasingly applying Artificial Intelligence (AI) methods from Natural Language Processing (NLP). However, such co-creative use of AI has not been examined yet for higher-level revisions. These aspects motivate our research question for this paper: *How might we enable writers to benefit from computational (NLP/AI) capabilities beyond text generation and correction, in particular for revision?*

We address this question by exploring how we might support reflection and (structural) revision via *automatic summarization*. This approach was inspired by taking an existing cognitive strategy for revision as a starting point, namely *reverse outlining* [19]: Writers manually create reverse outlines *after* a part of the draft has been written, summarizing it, to reflect and identify opportunities for improvements. This guided our design and prototype: We automate this summarization to continuously update paragraph-wise summaries shown to writers next to their text.

We tested this in two in-depth user studies with a total of 12 participants writing analytic essays about a given article. The summaries facilitated reflection and overview: They gave users an external perspective on their writing (e.g. comparing own expectations against what's in the summary). People used this to reflect on content and scope of paragraphs (e.g. identifying redundancy), leading to revisions. They further used the tool to quickly gain an overview of the given article, and developed strategies to integrate insights from the automated summaries into their text.

In summary, we contribute a text editor prototype with automatic text summaries and its evaluation in two user studies. More broadly, our work contributes to the literature on interactive NLP, text interaction and co-creative AI: It explores and highlights the value of assisting writers with NLP capabilities beyond direct text generation and correction.

2 BACKGROUND AND RELATED WORK

This work draws upon the areas of writing reseach, AI and NLP, and reading and writing support tools.

2.1 What is a Summary and What is its Goal?

Here we clarify our understanding of the term "summary" in this paper. Overall, we align with the definition by Radev et al. [35], who define a summary "as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that." They further state that the "main goal of a summary is to present the main ideas in a document in less space", which fits to our application as well, although we target the paragraph-level instead of the whole document. They also highlight a distinction in the literature between *indicative* summaries (what is the text about, without specific content) and *informative* ones (short version of the text). Our concept and system use the informative type.

Moreover, Ceylan and Mihalcea [5] distinguish between *objective* and *interpretive* summaries to inform their work on a book summarisation system: The former type objectively summarises the plot, while the latter also integrates the subjective view of the person writing the summary. This classification is interesting to consider for our use case: Our system creates *objective* (automatic) summaries of the user's text in order to prompt and facilitate the user's *own interpretation* (i.e. self-reflection).

Finally, *extracts* are short versions reusing parts of the full text verbatim, while *abstracts* are short recreations [35]. Our concept and system explore both types (see Section 3).

2.2 Automatic Text Summarization

There are two main NLP approaches for automatically creating summaries of a given text [20]: *Extractive* methods aim to select the most important/relevant parts of the text. This selection directly yields the summary. In contrast, *abstractive* methods aim to write a new piece of text to serve as a summary, similar to what most humans would do when tasked to summarize a text. There are also methods that mix aspects of both approaches, such as pointer concepts [41], which allow a model to copy some parts verbatim while (re)writing others. In our prototype, we employ a *TextRank* [30] approach for extractive summaries and the *T5* model [36] for abstractive summaries (details in Section 4).

2.3 Reading and Writing Support Systems

A number of interactive systems support reading and writing with summarization: For example, Leiva [23] used extractive summarization to make websites more responsive to the device size – not only by adapting the visual layout but also the length of its text

content. The system by Wang et al. [52] summarizes mobile UIs to create succinct descriptions of screen content, and ter Hoeve et al. [47] proposed a conversational UI (chatbot, voice assistant) that reports information from a document when asked about it. Li et al. [25] combined speech recognition with summarization to show text snippets that help users navigate long audio content.

Related, but not using summarization, are the many tools that support creating text; this is only a small overview: Already in 1982, Macdonald et al. [27] built a "writer's workbench", which could comment on stylistic features. Recent work typically generate text suggestions (e.g. for the next sentence [38] or paragraph [54] or in a sidebar [42]). More broadly, this co-creation through interleaved human writing and AI-generated text is a common approach (e.g. [4, 6, 22, 45, 57]). Further work explored controllable generation of plots with AI [7, 46], or addressed writing poetry [14], metaphors [13], slogans [8], fictional characters [40], and emails [4, 18]. Strobl et al. [43] provide a recent survey of 44 academic writing tools. Further recent work supports revision on the level of word replacement [16, 26] and (typing) error correction [1, 11, 58], partly using NLP models. Finally, Arnold et al. [2] recently highlighted opportunities for supporting writing without generating text, which motivates our design direction here.

In summary, the literature has mainly used (1) automatic summaries as reading support, and (2) text generation as writing support. We explore the remaining combination, namely *using (AI) summarization to support writing* as it happens. Conceptually, we build on an existing writing strategy, described next (Section 2.4).

2.4 Text Summarization in Writing Research, Practice and Instruction

Summarizing text is an integral step in the strategy of *reverse outlining* [19] (or *backward-outlining* [39] or *post-outlining* [32]), which has been called "the Swiss army knife of revising" [49]: Instead of a typical outline that is created before drafting a text, a reverse outline is created after (a version of) the draft has been written. Concretely, the writer summarizes each paragraph. If done on paper, the result can be cut into one note per paragraph for easy rearrangement [28, 29]. These summaries support self-reflection about the text's structure (e.g. *Is the text's claim supported by each part?*, *Are the aspects covered in a suitable order?*, *Do paragraphs contain a single thought each?*) and subsequent actions based on the gained insights (e.g. reordering, splitting, merging or deleting paragraphs). This strategy is a part of many university writing guides and courses (e.g. [33, 34, 39, 49]), and other teaching [19, 24, 48], and valued by professional writers (e.g. [15]).

More broadly, reverse outlining is an instance of (self-)annotation. Annotations are "concise descriptions" of a work and can be descriptive or evaluative [56]: Our use of AI summaries is descriptive since it does not include statements on what is good or bad about the summarized text. As suggested in studies with students [10], the act of self monitoring afforded in this way can improve control over one's own writing. In this light, our work explores AI-supported self monitoring during writing.

3 CONCEPT DEVELOPMENT

Here we report on our concept development, starting with the core conceptual inspiration, before covering UI aspects and interaction.

3.1 Core Concept: AI Text Summarization as (Self-)Annotation

As described in Section 2.4, reverse outlining is an effective strategy for reflecting on and revising a text structurally – but without interactive support yet. In our concept development, we took this as a starting point to explore how we might support it technically in an interactive system. Concretely, reverse outlining offers two key conceptual aspects that we pick up on in our design: *Annotating* what has been written, and the fact that this is *one's own text* (and not that of another person). We comment on both in relation to our concept and design direction in more detail here:

LeVan and King [24], Yaylı [56] emphasize the value of self-assessment with self-annotation in structuring and understanding text during writing. They mention "sideline commentary for their writing" as a strategy for students to externalize their thoughts and gain an overview of their work in progress. As a key part of our conceptual exploration, we decided to automate the summarization step of such self-annotation with AI to support self-assessment. While the act of manual summarization can be seen as a part of the reflection process, we assume that it is also an initial hurdle to enter a reflection and revision step (e.g. investing time and effort into summarizing what you have already written vs writing more). By making summaries automatically available to writers our concept thus is intended to invite and support reflection – and potentially revision, if identified as needed.

Automatic summaries are not a "zero cost" feature for writers because they need to read them, to benefit from them. In reverse outlining, this in itself is considered useful: Tully [48] noted that seeing reverse outlines gives a "fresh eye" on the text as it motivates taking writing "breaks" and "time away from a draft weakens the memory's tie to the narrative" – thus stimulating (self-)reflection. Since writers in our concept do not write summaries themselves, it may be even easier for them to achieve this "fresh" view through the summaries. Indeed, we found this in our study (Section 7.3).

3.2 UI: Summaries as Margin Annotations

The writing interface mimics known writing software and is split into two main views. One view represents a classic word processing editor. The other view represents the margin, the usual place for annotations. To stay in line with common terminology in user interface design, we termed this second view the "sidebar". Writing guides suggest using the document margins or separate notes (or post-its) for additional text annotations, for example to develop a reverse outline (e.g. [21, 32]), or as "sideline commentary" [24]. Concretely, we represent each paragraph as a card, based on the suggestions of Messuri [28] to cut summaries on paper into multiple snippets. The details of this from an interaction point of view are covered below (Section 3.3). The text editor and the sidebar are inherently linked by their content. Each text paragraph in the editor produces a summarised annotation card in the sidebar. To support this, we highlight the paragraph when clicking on its card (cf. Fig. 1).

3.3 Interaction: Summaries as Cards

Within the sidebar, the summaries are presented as "cards". This is motivated as a metaphor considering the use of pen and paper in (reverse) outlining and other revision activities (e.g. margin notes, post-it stickers). There, notes may be written on cards (or post-it notes) to afford a set of relevant interactions, which we map to interactions with our digital cards as described next. In each part, we indicate the mapping of *physical (paper) action → UI interaction.*

3.3.1 Reordering: Move paper notes → Drag & drop cards. Related work has commented on the benefit of reordering paragraphs in the writing process [28, 29, 39, 49]. The underlying strategic motivation and intention is that the ability to step back from the text to reorder paragraphs supports and enhances structural revision. With a (reverse) outline on a physical piece of paper this means cutting the outline paper into smaller snippets (if not already written on cards) so that these can be moved around and arranged in a new order. In our UI, we mapped this to cards (i.e. summaries are "precut" per paragraph) and the interaction of dragging and dropping these cards vertically in the sidebar.

3.3.2 Removing: Throw paper away → Delete card. Removing content is one possible action that writers might identify when reflecting on a (reverse) outline of their text [15, 32, 39, 49]. For example, this might be triggered by the insight that some parts of the text are redundant. With paper notes, writers can remove said note while in our UI we provide a "delete" button for each card (see Fig. 1).

3.3.3 Splitting: Cut/rewrite paper notes → Add line break. Another related action is splitting, which writers might identify as a way of improving their text based on insights they gained from a (reverse) outline [28, 29, 33, 49]. For example, they might see that one paragraph mixes two topics. On paper, writers may cut a note (or rewrite it on two new pieces of paper). In our UI, writers can simply add a new line break in the main text to split a paragraph into two, which automatically updates the cards in the sidebar accordingly.

3.3.4 Merging: Glue/rewrite paper notes → Delete line break or drag & drop cards onto each other. Inversely to to splitting, (reverse) outlines may also reveal to writers that two text parts should be merged into one [33]. On paper, writers might glue notes together or rewrite them onto a single new card. We support this in two ways: For a simple concatenation, users can move the involved paragraphs next to each other (if they are not already subsequent) and delete the line break between them, which also updates the cards accordingly. Alternatively, if the merge is semantically more involved, users can drag & drop one card *onto* another one, which will trigger a dedicated merge view with an automatically created merge suggestion (see Fig. 3 and Section 4 for details). Users can accept or cancel this suggestion. If they accept it they can of course also further edit the result in the main text view.

3.3.5 Revision: Rewrite/reprint main text → Copy card content. Finally, writers might sometimes identify content in their annotations as suitable pieces for a revision of the main text, for example, to achieve a more succinct version. On paper, writers could rewrite/reprint the main document or manually cross out or write over it. In our UI, we provide a button to replace a paragraph directly with its summary (in version 1 of our prototype), which we later turned into the more flexible concept of copying the content of a card to the clipboard (in prototype version 2). We report more details on this conceptual change in our results and discussion.

UI Element	Version 1 (study 1)	Version 2 (study 2)
Cards-text link	Click on card navigates to paragraph	Added text highlighting when card is clicked (see Fig. 1)
Copy summary (compare in Fig. 2)	"Replace original text with this summary"	"Copy to clipboard"
Merge	No visual highlight of merged content	Added visual highlights of retained and cut parts of the merged paragraphs (see Fig. 3)
Summary levels (compare in Fig. 2)	*Level 0:* original paragraph in full; *Levels 1 - 4:* extractive summary with (4, 3, 2, 1) sentence(s); *Level 5:* abstractive summary	*Original:* full text; *Central sentences:* extractive with 1 sentence; *Summary:* abstractive; *Keywords:* extract up to 5 keywords

Table 1: Prototype changes implemented in Version 2 after the feedback from study 1 (Version 1).

3.4 Annotation Content and Controls

Another key aspect of our concept concerns the content of the annotations/cards. Many choices are possible here, considering different kinds of summaries (Section 2.1) and degrees of granularity. We thus offer the user control over multiple levels. Providing more than a single option in this way is further motivated by varying comments on length and type of summaries for reverse outlining in writing guides (e.g. one sentence, main idea) [15, 21, 48]. In our protoype, we explored two sets of levels: 1) Abstract numbers ranging from one to five representing *increasing summary zoom levels* and 2) descriptive summary levels (see Fig. 2).

4 IMPLEMENTATION

Our prototype has a server client architecture, with the front end UI, a Python server to host the text summarization methods, and a webserver and database to serve the web app and questionnaires and store data collected in the studies. We improved the prototype after the first study (described in detail in Section 7.1). Table 1 summarizes the differences between the two versions.

4.1 Frontend / UI

We implemented the UI (Fig. 1) as a web app with React.js. It realizes the concept described in Section 3: It is split into a main text editor and a sidebar with summary cards. Each paragraph is represented by one card. The cards can be reordered, deleted or merged.

4.1.1 Card Interactions. We further implemented all card interactions as described in Section 3. Card dragging was implemented using react-beautiful-dnd. In version 2, clicking a card highlighted its pargraph in the text in color for one second (green box in Fig. 1).

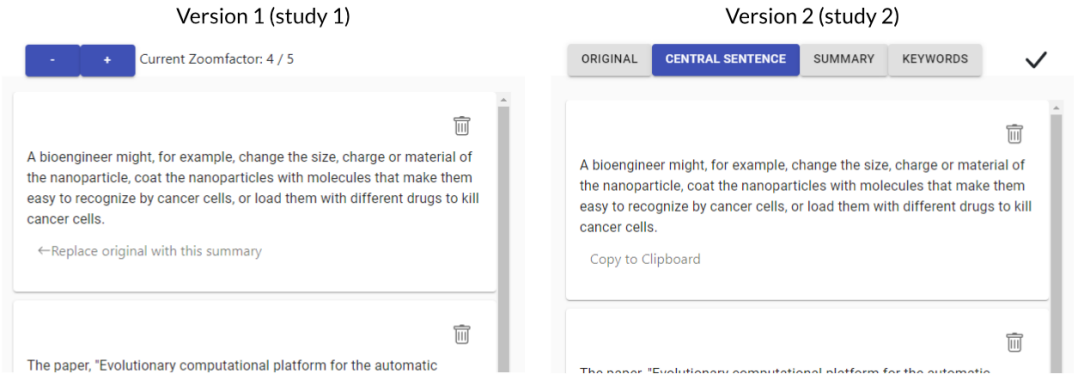

Figure 2: Our design iteration of the summary controls, from version 1 (used in study 1) to version 2 (used in study 2): We redesigned the summary levels (at the top of the UI) and replaced the "Replace in Text" action with a more flexible "Copy to Clipboard" action (at the bottom of each card). We also added an indicator in the top right that shows whether summaries are currently being updated in the background (spinning circle or checkmark).

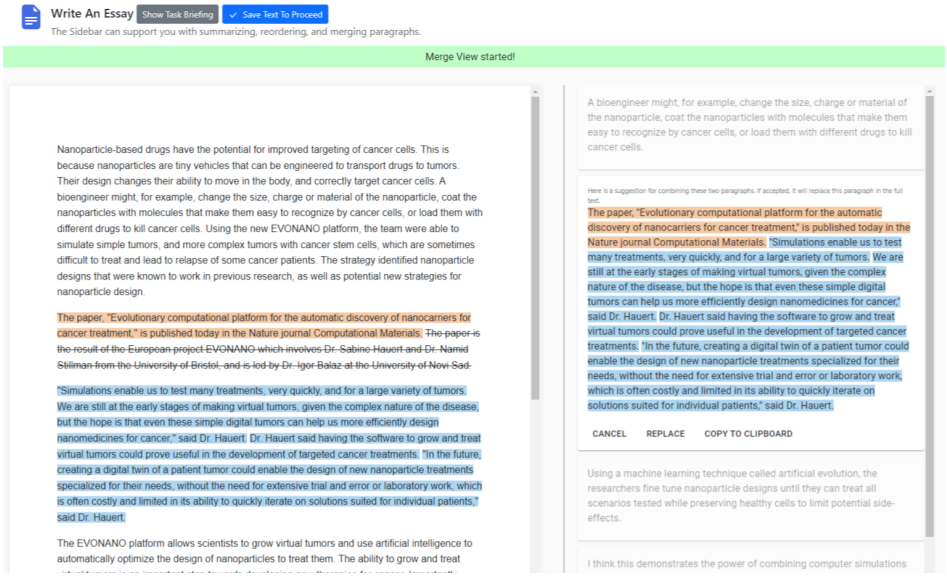

Figure 3: Merge view (in prototype v2 / study 2), triggered by dragging one card onto another. The sidebar shows a merge suggestion in a merged card. A green message (top), removed summary controls and grayed out other cards indicate that no further interaction is possible until the merge is accepted or canceled. The suggestion is computed by ranking sentences in each paragraph (Section 4). The page view highlights retained and cut parts of the merged paragraphs (color-coded, plus strikethrough for cuts). Version 1 was the same but showed no colors and strikethrough.

4.1.2 Merge. For the merge functionality, the user drags and drops a card onto another card. The system then applies extractive summarization to rank the sentences in each paragraph. The merge suggestion is composed by concatenating the five top ranked sentences across both paragraphs. The original order of these sentences in the text is preserved. In version 2, the page view highlighted retained and cut parts (see Fig. 3). A green message, removed summary controls and grayed out cards indicate that no further interaction is possible until the user has clicked "cancel", "replace" (which applies the merge in the main text directly) or "copy to clipboard".

4.1.3 Summary Levels. The toolbar at the top provides summary options, which we redesigned after study 1 (Fig. 1 shows the improved version; for a comparison see Fig. 2). Table 1 describes the options in each version. Section 4.2 describes the NLP methods.

4.2 Language Models and NLP Methods

Here we describe the used language models and NLP methods. We made pragmatic model choices for our interactive system since our focus was the writing insights. Concretely, we chose a tradeoff

between quality and computation speed, as assessed by trying out related models of different sizes.

4.2.1 Extractive Summarization. The extractive method uses *GloVe* embeddings[1] and *TextRank* [30] to find the top k sentences. In study 1, the extractive method was used for summarisation Levels 1 to 4, beginning with k=4 sentences on Level 1 with k decreasing by 1 per level until k=1 on Level 4. In study 2, extractive summarisation was used for the Central Sentence level with k=1.

4.2.2 Abstractive Summarization. Our abstractive method used the *T5* transformer model [36] available via *huggingface*[2], with the language modeling head. We mostly relied on defaults but determined these "generate" parameters by early explorations with our prototype: *num_beams*=4 , *no_repeat_ngram_size*=2, *early_stopping* was enabled and *max_length* was set to 70% of the source token count. In study 1, abstractive summarization was used in Level 5. In study 2, the level named *Summary* used it.

4.2.3 Keyword Extraction. We implemented the keyword extraction using the open source library *wordwise*[3] (which uses a RoBERTa model[4]). Keyword extraction was not implemented in study 1 but was used for the *Keyword* level in study 2.

4.3 Backend / Server

The prototype was hosted on a university server with 32GB RAM and a GPU with 12GB memory. We implemented a lightweight Python Flask API to process the text. Each method's backend call receives a JSON Array containing all of the required paragraphs and returns the desired results as a JSON object with each index representing one paragraph. We created a cache in order to improve loading and processing times. It stores the paragraphs as well as the return values of each summary technique that was previously used on that paragraph. It also records which paragraphs have been modified as a result of the author's revisions to the original text. Thus, the system only needs to newly compute the summaries of changed paragraphs. Overall, the resulting summary updates were almost instantaneous for extractive methods and keywords, and took up to 2 seconds (when adding long text at once i.e. via copy/paste) for the abstractive method.

5 METHOD

Our evaluation methods consist of a writing task with think-aloud, followed by a semi-structured interview, all during online sessions (video calls), plus a final questionnaire. For better clarity, we describe these methods upfront here. The study procedure involving these methods is then presented in detail in Section 6.

5.1 Think-Aloud During Interaction

Participants interacted with the prototype in a writing task, for which we asked them to articulate their thoughts. Occasionally, we also asked questions to better understand their thinking (e.g. when they seemed to be looking for something) or to remind them of thinking aloud. We recorded these video calls including audio and participants' screens. On top of that, two researchers involved in the video call took unstructured notes on comments from the participants as well as critical observations. Both researchers later shared, compared and merged their notes.

5.2 Semi-Structured Interviews

A semi-structured interview followed the writing task to dig deeper and allow people to share thoughts in reflection on their experiences. We asked questions about notable moments we observed during the task, plus detailed questions about which aspects of the prototype people particularly liked or disliked or would prefer to change.

5.3 Coding of Think-Aloud & Interviews

We used the researcher notes from the online sessions and transcripts of people's comments in an approach adopting Grounded Theory [9, 31]: In an *open coding* round, two researchers individually assigned inductive codes to each note. In an *axial coding* round, these two researchers compared their codes and clustered them into higher thematic groups. For example, emerging clusters grouped several codes that related to *improvement of the prototype* or codes that related to the *interaction strategy* of the participants. Afterwards, the researchers jointly agreed on a "cluster label" that best describes its corresponding codes. These resulting aspects also serve us as a structure for the results in this paper (Section 7). Finally, in a *selective coding* round, three researchers went through the full transcripts of all interviews to find further evidence (and potentially counterexamples) for these aspects as developed in the previous step. Since we had created the transcripts automatically, we double-checked all relevant parts again in the original videos.

5.4 Questionnaire

In an online questionnaire after the video call, people rated each feature on a five-point Likert scale (plus a *did not use* option). Two open questions asked about (1) other writing use cases where a system such as ours might be useful, and (2) what participants would like to add or change in the current prototype.

6 USER STUDIES

We conducted two user studies via video calls, with a design iteration of our prototype in between them. The procedure was identical in both studies. To avoid redundancy, we report on it here once.

6.1 Participants

In total, N=12 people (5 female, 7 male) participated in the two studies (4 in the first study, 8 in the second one; no one participated in both studies). Their age ranged from 22-36 years. We recruited them from networks across a few universities and personal contacts. Their backgrounds included professionals, undergraduate students, and (HCI) researchers to whom we reached out via email. While this is a convenience sample and we reflect on limitations of the study in our discussion, our sample covers relevant users for writing tools with an interesting range of (professional) writing experiences and regularity. Concretely, about half of participants indicated to write (substantially) daily, the others less than once a week. Participants were not native English speakers yet all used English regularly

[1] https://nlp.stanford.edu/projects/glove
[2] https://huggingface.co/t5-base – These T5 transformer models are trained for summarization and T5-base refers to a medium model size.
[3] https://github.com/jaketae/wordwise
[4] https://huggingface.co/docs/transformers/model_doc/roberta

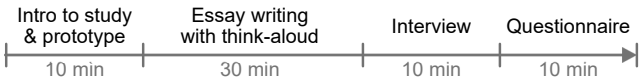

Figure 4: Overview of our study procedure (see Section 6.2).

professionally and/or personally and were informed beforehand that they would be asked to write an essay in English. We further informally confirmed their high proficiency based on the task observations and resulting texts. People were compensated with a 10 Euro gift card for an online shop.

6.2 Procedure

On average, a study session took 60 minutes. It was structured as shown in Fig. 4. We explain the involved steps in more detail next.

6.2.1 Study Intro (10 minutes). In line with our institutional regulations and informed consent procedures, an intro page in our web app explained the study, informed about data collection and privacy regulations, and further general study information. One of the researchers then demonstrated the features of the prototype as an introduction via screen sharing. The text used in this demonstration was taken from an arbitrary daily article on Wikipedia. As part of this demonstration, the researcher showed the different summarization levels, the merge functionality, and the linking between cards in the sidebar and paragraphs in the page view. People could ask questions, for example to clarify how the interactions worked.

6.2.2 Essay Writing (30 minutes). In the main part, people chose one of two opinionated articles and wrote an analytical essay on how the respective author builds their argument. The task prompt and articles were taken from the analytical writing section of practice test prompts for the Standardized American Tests (SATs)[5]. We chose this task because it allows users to work with an existing text and thereby increases the opportunity to experience the prototype without an extensive "creative" drafting period. At the same time, this task requires people to analyze and write about the text, not to simply copy it, thus ultimately leading them to write an essay. We deemed this a useful trade-off between starting from scratch and starting with a given text in the prototype (i.e. pure editing task). We reflect on this choice in our discussion. People opened the article in a separate browser tab so that they could switch between the prototype and the article. We encouraged thinking aloud (see Section 5.3). With people's consent, we recorded audio and screen during the essay writing and the following interview.

6.2.3 Concluding Interview and Questionnaire (10+10 minutes). We conducted the semi-structured interview after the writing task (see Section 5.2). Finally, after the video call, people filled in the questionnaire (see Section 5.4).

7 RESULTS

We structure this report into design insights and identified interaction, writing and reflection strategies. The mean duration of interaction and interviews was 40 minutes, matching our planned time for these parts. The mean final text length was 416 words. All summary settings were used and both extractive and abstractive methods were used considerably in the revised design (see Fig. 5).

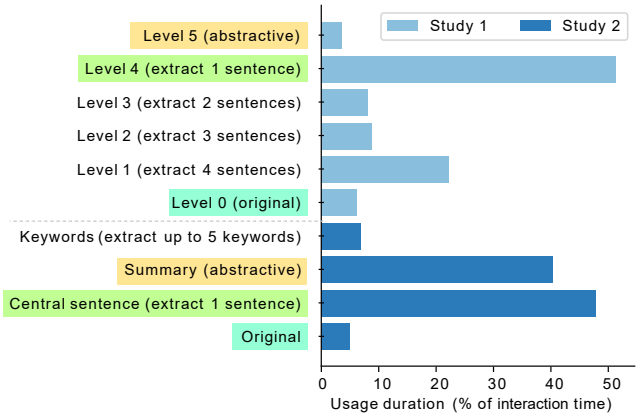

Figure 5: Time spent in each summary setting. Colored labels show which settings from prototype version 1 (study 1) correspond to which settings in version 2 (study 2). Uncolored settings only appeared in one design/study.

7.1 Design Insights and Improvements

We report findings in study 1, that motivated prototype changes before study 2. We report these together with a description of the changes and observed consequences in study 2.

7.1.1 Unclear Semantic Zoom and Summarization Levels (Study 1) were Redesigned (Study 2). The initial zoom metaphor was a source of confusion. First, the minus/plus symbols of the zoom tool where interpreted as text length, which is the inverse mental model: As text length, "minus" should decrease summary length but it was designed as "zooming out" semantically (i.e. seeing more of the original text, thus actually increasing length).

Second, people had trouble understanding how the levels differ: *"I felt level one to three did not show much difference [...], this may also be because I included little information"* (P_1, study 1)[6]. Indeed, for short paragraphs, the first three zoom levels may not result in text changes because extracting a varying number of sentences does not make a difference if a paragraph has fewer ones anyway: *"I didn't realize that there were two more levels [...] since nothing changed, I felt that the few paragraphs I have created may not have been sufficient to trigger something here"* (P_4, study 1).

These issues motivated us to redesign this tool: With version 1, we explored a rather fine-grained set of summarization options. However, as described above, this was not particularly helpful for people. We thus decided to only keep four levels in version 2, also motivated by what was used the most (Fig. 5): *original* text paragraph, single *central sentence*, short *abstractive summery*, and extracted *keywords*. We also labeled these levels with text captions on four separate buttons, which avoids the plus/minus. These changes

[5]https://blog.prepscholar.com/sat-essay-prompts-the-complete-list (Topic 1: Benefits of early exposure to technology, Topic 2: Preservation of natural darkness)

[6]Participants' quotes were translated into English by the authors.

were clearly successful: We found no indication of misunderstandings in both observed usage behavior and think-aloud comments.

7.1.2 Unclear Correspondence of Text and Cards (Study 1) and Improvements (Study 2). Participants in the first study often commented on "losing orientation" when interacting with the cards: For example, some people had issues deciding which card matched which text paragraph at a glance.

We partially solved the issue by highlighted the corresponding text paragraph when a user clicked a card. All participants in study 2 appreciated this; for example: *"It's a good feature to be able to highlight this, then it is easier to find [corresponding text]"* (P_{11}, study 2). However, this highlighting on click was only from cards to text. As some people's behavior in study 2 showed, it should be bidirectional (i.e. clicking on paragraph to highlight its card). For example, as a workaround, one person (P_6, study 2) quickly clicked through the cards until their paragraph of interest was highlighted.

A related issue in version 1 was the button in each card to replace content in the text (see version 1 in Fig. 2, bottom of card). The resulting change lacked visual feedback: *"I thought that something was happening [...] but suddenly it's gone [the text paragraph] and something has changed"* (P_3, study 1). Considering further observations, we solved this problem not with more feedback here but more generally by changing this feature to *Copy to Clipboard* (see version 2 in Fig. 2), as described next (Section 7.1.3).

7.1.3 Unflexible Direct Changes (Study 1) were Replaced with a Copy to Clipboard Feature (Study 2). In the first version and study, the *Replace in Text* button in each card replaced the corresponding paragraph with the card's summary directly. This had three downsides: (1) The replacement was inflexibly tied to the summarized paragraph – it was not easily possible to use the summary elsewhere in the text. (2) As mentioned above, the lack of visual feedback was confusing. (3) Finally, this button encouraged people to view the summaries as text *suggestions* rather than as annotations.

We thus change this to a *Copy to Clipboard* button. This requires users to perform an extra step (paste) to include the summary into their writing. This offers more freedom to decide where in their text they want to insert it. At the same time, as we indeed confirmed in study 2, this change discouraged people to view the summaries as text suggestions that are presented to be "accepted".

7.1.4 Difficulty of Assessing Merge Changes (Study 1) was Addressed with Better Merge Preview (Study 2). People in study 1 initially found it difficult to spot the differences of the summarized text and the original paragraph. Especially when merging two paragraphs they wanted to see more clearly which parts of the text were kept in the merge suggestion and which were removed. The second iteration of our prototype thus highlighted this with colors and font styling (see Fig. 3). This was a successful change as people in study 2 commented positively on this visual feedback.

7.2 Interaction and Writing Strategies

Here we report on key interaction patterns and writing strategies.

7.2.1 Using Automated Summaries to Understand Longer Text. One distinct strategy was to gain a quick overview of longer text, for example, by copying parts of the source article into our editor and reading the summaries: *"Let's see what the editor says about the individual paragraphs. [I'm going to] look at the text summaries, or more concretely, the central sentences in order to get a better overview"* (P_3, study 2). This was perceived as fast: *"It is definitely faster because one does not have to reread everything"* (P_8, study 2).

Related, summaries simplified the text for reading and selection for building an argument: *"I wanted to look at a shortened version, to see whether the summary [of a part of the source article] is irrelevant for my argument"* (P_6, study 2). As can be expected, this strategy was employed at the beginning, prior to writing. It started with trying out summary levels before settling for one for reading. The strategy ended after reading the summaries and was followed by two distinct transitions to writing: (1) One group integrated the summaries directly into their text as a starting point. (2) Others deleted the article form the editor again to start with a blank page. We describe the larger strategies emerging from this next (Section 7.2.2).

7.2.2 Developing Text Top-down vs. Bottom-up. Two larger distinct writing strategies emerged in our study: First, in what we call *bottom-up* text development, participants started with an empty editor and then built up their text through typing, interleaved with checking the reference article for the task. Occasionally, they copied specific snippets from this source article into the editor to read the summaries and/or use the text as a basis for further writing.

In contrast, in what we call *top-down* text development, participants copied the entire source article into the editor and used the summaries as text building blocks. In this strategy, people kept the source's argumentative structure and mainly focused on drafting sentences to connect the summary-based paragraphs.

Across both these strategies, people edited summaries (or directly used source text) to avoid plagiarism and to match their own writing style. We describe this behavior in more detail next (Section 7.2.3).

7.2.3 Adapting Text as Part of Integrative Leaps and "Reverse Leaps". When using text from summaries to revise or draft text, the most dominant strategy was modifying the summary text to match one's own writing: *"This is not copy-pasted one-to-one. I have partly adapted it [the copied summary] and reformulated it"* (P_6, study 2). These edits can be seen as *integrative leaps*, as recently introduced by Singh et al. [42]: It describes people's efforts to integrate (AI) suggested material into their writing. We also made the *opposite* observation where people adapted their own writing to influence the summaries: For example, one person combined multiple text paragraphs to explore changes in the summaries. Similarly, one person said: *"I'll remove all empty spaces between the paragraphs to see whether the message I want to deliver comes through"* (P_3, study 1). This also demonstrates reflection on paragraph structure and scope, which we elaborate more on in Section 7.3.2.

7.2.4 Summaries as Textual Building Blocks for Overview Texts. People further considered the text from automated summaries to write "overview sections". For instance, one person framed this as building blocks to formulate their text's conclusion: *"I want to add final summary sentences at the end, [...] like a conclusion. I quickly skimmed through existing text summaries to recall what the main messages were"* (P_5, study 2). While another participant mentioned a similar strategy to build an abstract: *"I'd take the entire text and try to generate an abstract out of it"* (P_{11}, study 2).

7.3 Reflection Strategies

Here we "zoom in" on people's reflective use of our system.

7.3.1 Considering Summaries as Another, External Perspective. A key aspect that characterizes the reflection processes with the summaries is that they were considered as an external view on the writing; such as: *"And that's how I worked with [the summary], I'd first write down my text and would then compare: What does the bot suggest? And then I've integrated [the summary] in my text"* (P_{11}, study 2). And: *"It's good to read [the paragraph], in other wording, and [...] slightly summarized"* (P_{12}, study 2).

Related, also hinted at in the first quote above, people *compared* the AI summary with their own "mental summary" for the same paragraph. Noticing differences prompted people to more closely analyze why this was the case. In doing so, they reflected on their written work. For example: *"[The key message] is essentially what the tool also proposes as a summary. [...] It's interesting that what the tool proposes is longer than my own summary"* (P_9, study 1). This reflection often also led to edits.

7.3.2 Using Summaries to Check and Revise Paragraph Scope and Structure. Another way of using the summaries for reflection was to check the scope of individual paragraphs, for example to determine if one indeed included its intended content or message. For example: *"I would take a look at both [the paragraph and its summary]; both things are important. I would like to see in the summary whether all my arguments are also reflected [...]. I'd expect the summary to help me reflect where I set my focus in the section [...]"* (P_5, study 2). If the summary did not include the intended aspect, people applied two revision strategies: (1) Either the main message is elaborated on in the paragraph, or (2) the paragraph is split.

Related, restructuring actions were also valued: *"The most helpful [features] were definitely the reorder and merging of paragraphs. When I noticed that the summaries at a specific zoom level were too short, I could just merge them; that simplified the work and made work much faster. That was the most practical aspect. Other than that, the summary and simplification of large paragraphs"* (P_1, study 1).

7.4 Feedback in the Final Questionnaire

The final questionnaire asked people to think about where (else) they would use this system. The answers echoed the topics of overview and reflection identified during the interaction and interviews: *"[...] it could be helpful for organizing my chapters and for getting a quick overview of different parts"* (P_2, study 1). And another person thought it would be useful *"[w]hen writing texts where structure and conciseness matters (blog entries, news articles, summary/conclusion [...])"* (P_5, study 2).

The questionnaire also asked about feature ideas and changes. We addressed the replies from study 1 in our design iteration (e.g. highlighting text changes, cf. Section 7.1.4). Replies from study 2 revolved mainly around further improving the visual link of paragraphs and summaries (e.g. one person suggested to highlight the paragraph already on hover, not click).

Fig. 6 shows the results of the included Likert items on individual features. Overall, not everyone found every feature equally helpful (or used it) yet all features were helpful for a considerable proportion

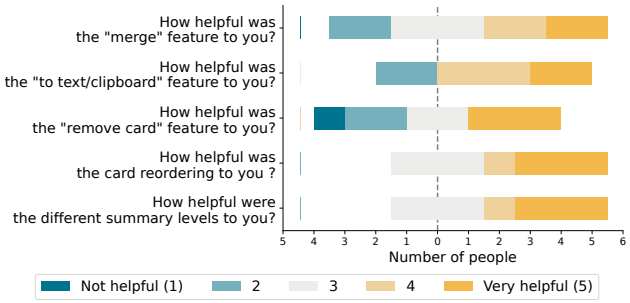

Figure 6: Likert ratings of specific features asked in the final questionnaire. The number of ratings varies because a "did not use" option was available for each question.

of people. This fits to the observed diversity of people's writing approaches. "Delete" was the least helpful feature.

7.5 Comparisons against Writing without Automatic Summary

Even without a comparative baseline a few participants explicitly commented on their writing experience with and without automatic summaries. Their comments hint at the use of different kinds of manual summarization in their usual work flows, and how these benefit from our tool. For example, a common reading technique involves highlighting important parts of the text, which corresponds to our "central sentence" level: *"I would have just started writing [...] and summarized the relevant paragraphs from the source in the process. I would have highlighted important parts in the text and only after finishing the entire text I would come back to think about the structure of the text."* (P_8, study 2).

In another example, a participant explained that *"[w]ithout the margin summaries, I would probably first draft the entire text finally to recognize that my submission was too long [...] only after [that] I would start summarizing and shortening the text"* (P_4, study 1).

Finally, automatic summaries shorten reading time when writing with source material, as in the study: *"[...] it really formulates concretely what I need. In the end it is exactly what I would have usually done laboriously by hand. I would read for couple hours. Here I can just copy the text."* (P_9, study 1).

8 DISCUSSION

8.1 How Continuous Automatic Text Summaries Support Writers

Writing is an iterative process interleaving several sub-processes that relate to planning, drafting and evaluating text [12]. Our continuous automatic text summaries consider this in that they are updated as users write their text and thereby succinctly reflect the state of the draft, as an offer for evaluative processes. This concept was accepted by participants and utilized in diverse ways: Crucially, the summary annotations were used and valued by people as a support for reflection and overview. Concretely, people reflected on the content, scope and structure of their paragraphs, by integrating insights from the corresponding summaries. This also matches the

intention of the reverse outlining strategy, which we had used as a starting point for our design. Moreover, the summaries helped writers to gain a quick overview of both the given text at the start as well as their own essays at the end. Finally, as a key finding on the perception and human-AI relationship here, people viewed these summaries as another, external view on the text. It is worth noting that this perception was not limited to summaries of text copied from the given articles but also was the case for summaries of the users' own writing. This "external AI view" can be helpful, especially when otherwise writing alone.

8.2 Self-Annotation and "AI Annotation"

We critically reflect on the shift from self-annotation to "AI annotation" in this work. First, manual summarizing can be helpful as part of (meta)cognitive strategies for writing, text understanding and learning to write [53]. Automatic summaries might take this away. However, participants indeed found the summaries helpful, as described above. Interestingly, a typical behavior involved *comparing* a mental summary to the generated summary. This indicates that AI annotation influences cognitive processes in self-annotation, instead of replacing them. Crucially, as pointed out previously, AI annotations were to some extent considered as an external view. Future studies could further relate this to perceived agency and authorship, as examined for writing with text generators (e.g. [4, 54]). Notably, low perceived authorship *for annotations* might have positive utility for writers (i.e. as "fresh eyes" [48], as suggested in our study) – in contrast to the case when generating or editing text, where AI might be seen as taking away agency.

A second point is expressiveness: Self-annotation offers more flexibility than our UI, both spatially and symbolically, (e.g. freehand drawings, also beyond the margins). The spatial aspect appeared in our study in comments on locating annotations in the main text (cf. Section 7.1.2) yet people did not seem to miss or expect visual annotations as part of our concept. Our takeaway here is that the *correspondence* of AI annotations and user content needs to be clarified explicitly (e.g. our click-to-highlight design change) because people lack the "cognitive context" that would come implicitly via manual annotation. At the same time, users' expected *modality* of NLP-powered annotations in our design was clearly textual.

Combining these two points, our findings guide the community towards exploring AI annotations as *complementing* self-annotation with two concrete directions: (1) supporting a mix of manual and automated annotations (e.g. reusing UI concepts from human-human collaboration, such as differently colored annotations for writer and AI), and (2) annotations beyond text (e.g. by combining our work with the stylus diagramming interactions by Subramonyam et al. [44] or even text-based sketch generation systems [17]).

8.3 Limitations & Reflections on Methodology

We do not claim that the generated summaries represent the latest state-of-the art in NLP. Although people sometimes had reservations about summarization quality (e.g. fearing at the start that it might lack important context), they productively worked with the summaries and also integrated them into their own text. From our observations we hypothesize that even with "perfect" quality users would experience AI summaries as external because they are not written by users themselves, need to be read, and trigger comparison to own expectations. Related, in future work, it might be interesting to examine the impact of how well AI summaries match a user's writing style.

It is challenging to let people experience not only free writing but also revision within the limited duration of an observed study session. Here, we reflect on lessons learned: First, our observations and people's feedback on the study suggest that writing experience may influence how comfortable people feel with being observed. We believe our task was useful here because it gave people something concrete to get started with, while familiarizing themselves with the study situation, without having to write immediately (i.e. reading the provided article first).

Second, however, the task of writing about a single article does not perfectly capture people's typical writing tasks. We see it as a trade-off: It allowed people in the study to start from a given material while requiring them to write a new text (about the article). Starting without any input would likely require a much longer study to experience some aspects of the writing process.

Third, we decided against a baseline setup where users write text without summarization support. We did this to maximize the writing time spent using our system. However, even without a comparative baseline some of the participants compared our system to their usual writing process (see Section 7.5).

Based on these experiences, we conclude that the writing task to analyze given argumentative texts was a suitable choice. We decided for a controlled study to gain insights into participant's thinking while using the tool, but we encourage future work to also consider in-the-wild studies to assess other writing setups. Related, we expect our system to work better for writing longer texts. Users' feedback mentioned long articles or theses as examples. Besides studying longer use, it could also be interesting to compare use for argumentative, informative, and creative texts.

8.4 AI Role Perception Through Design

Our first design had a *Replace in Text* button in each card that replaced the corresponding paragraph with the summary. As observations and think-aloud showed, this partly made people perceive summaries as *text suggestions* that should be included in the text, rather than annotations. While the use of summaries in one's own text is not inherently wrong, this clearly indicated a miscommunication of the intended role of the AI in our design. We thus changed *Replace in Text* to *Copy to Clipboard* in study 2. This suggests a more "passive" role of the summary, which could be put into the text with a dedicated further action (paste). Functionally, it is also more flexible, since users can copy it anywhere they like. Observations and think-aloud in study 2 indicate that this indeed shifted perception away from "text suggestions". This contributes a concrete example of how an arguably small design choice might evoke prior mental models around AI features (e.g. of known text suggestion features) that shape the perceived role of the AI.

In a broader view, we can position our work in the recent framework for modeling interaction in co-creative AI systems by Rezwana and Maher [37]: Their 2022 survey of 92 co-creative AI systems revealed three predominant interaction models: In two, the AI generates content, either turn-based or in parallel to the user. In the

third, the AI (also) evaluates the user's creation, in a turn based manner. With this paper, we contribute to the literature an exploration of another interaction model: The AI generates parallel content that empowers users in their self-evaluation. With this new direction, we also respond to the survey's conclusion "that the space of possibilities is underutilized" [37]. Considering their insights, future explorations of said space could address implicit human-AI communication (e.g. gaze-informed summaries).

8.5 Designing Human-AI Co-creative Systems from Existing Cognitive Strategies

In their work on envisioning less "obvious" NLP applications, Yang et al. [55] highlight that "[a]uthors are inherently better than algorithms at comprehending their unfinished writing". Consequently, they propose to reframe the author-AI relationship in terms of other NLP problems (conversation, retrieval/search, question answering). In this way, they used NLP concepts to inform interaction design, whereas we used human writing concepts (i.e. reverse outlining) to do so. We consider both approaches as complementary directions to explore and "[...] expand this narrow intersection between what is [of] value to users and what can be built" [55].

Dissecting this, our approach starts with a strategy in the target domain (reverse outlining in writing) and asks how it might inspire writing support with NLP. Concretely, the reverse outlining strategy guided and thus greatly facilitated our decision-making for initial design choices (e.g. layout, main interactions) – even if it was not our goal to "'force" writers to apply exactly this strategy when using the resulting system. Crucially, it also usefully constrained the role of the AI: Summarize only, no interpretation, no "idea" generation. However, getting the main features and their UI right required more than this starting guide, as the first study and design iteration showed (e.g. redesigned zoom levels, change from "replace text" to "copy to clipboard"). Overall, we recommend this approach based on our experiences here. Looking ahead, existing human strategies (and their respective treatments in writing research) might serve the text interaction community in a role akin to generative theory (cf. [3]). Our approach provides a template: We can examine other human writing strategies to inspire new NLP tools for writers.

9 CONCLUSION

We have investigated how writers can be supported in reflecting and revising their text. We have found that automatic AI-generated summaries help writers by providing a continuously updated overview and an external perspective during the writing process. Participants used these insights to check and revise the scope and structure of their text draft. This work demonstrates writing support beyond the current focus on text generation and correction. More generally, we envision future AI-powered writing tools to offer a mix of direct text edits and indirect reflection support for writers. We encourage future work to explore how other adaptive margin annotations may help writers in the text drafting process and also how non-textual annotations can be integrated. To facilitate such research in co-creative writing tools, we release the prototype and further material on the project website:

https://osf.io/v6zfn

ACKNOWLEDGMENTS

We thank Christina Schneegass and Lukas Mecke for their feedback on the manuscript. This project is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

REFERENCES

[1] Ahmed Sabbir Arif, Sunjun Kim, Wolfgang Stuerzlinger, Geehyuk Lee, and Ali Mazalek. 2016. Evaluation of a Smart-Restorable Backspace Technique to Facilitate Text Entry Error Correction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5151–5162. https://doi.org/10.1145/2858036.2858407

[2] Kenneth C Arnold, April M Volzer, and Noah G Madrid. 2021. Generative Models can Help Writers without Writing for Them. *2nd Workshop on Human-AI Co-Creation with Generative Models - HAI-GEN 2021* (2021), 8.

[3] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. 2021. Generative Theories of Interaction. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 45 (nov 2021), 54 pages. https://doi.org/10.1145/3468505

[4] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. 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 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[5] Hakan Ceylan and Rada Mihalcea. 2009. The Decomposition of Human-Written Book Summaries. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (Ed.). Vol. 5449. Springer Berlin Heidelberg, Berlin, Heidelberg, 582–593. https://doi.org/10.1007/978-3-642-00382-0_47 Series Title: Lecture Notes in Computer Science.

[6] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2287–2295. https://doi.org/10.1145/3292500.3330723

[7] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Vol. CHI'22. Association for Computing Machinery, New Orleans, LA, USA, 19.

[8] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[9] Juliet M Corbin. 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Sage.

[10] Andy Cresswell. 2000. Self-monitoring in student writing: developing learner responsibility. *ELT Journal* 54, 3 (July 2000), 235–244. https://doi.org/10.1093/elt/54.3.235

[11] Wenzhe Cui, Suwen Zhu, Mingrui Ray Zhang, H. Andrew Schwartz, Jacob O. Wobbrock, and Xiaojun Bi. 2020. *JustCorrect: Intelligent Post Hoc Text Correction Techniques on Smartphones*. Association for Computing Machinery, New York, NY, USA, 487–499. https://doi.org/10.1145/3379337.3415857

[12] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.

[13] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300526

[14] Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an Interactive Poetry Generation System. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 43–48. https://www.aclweb.org/anthology/P17-4008

[15] Aaron Hamburger. 2013. Outlining in Reverse. https://opinionator.blogs.nytimes.com/2013/01/21/outlining-in-reverse/ Cad: 1 Section: Opinion.

[16] Han L. Han, Miguel A. Renom, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2020. *Textlets: Supporting Constraints and Consistency in Text Documents*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376804

[17] Forrest Huang, Eldon Schoop, David Ha, and John Canny. 2020. Scones: Towards Conversational Authoring of Sketches. In *Proceedings of the 25th International*

Conference on Intelligent User Interfaces (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 313–323. https://doi.org/10.1145/3377325.3377485

[18] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964. https://doi.org/10.1145/2939672.2939801

[19] Cynthia L. King. 2012. Reverse Outlining: A Method for Effective Revision of Document Structure. *IEEE Transactions on Professional Communication* 55, 3 (Sept. 2012), 254–261. https://doi.org/10.1109/TPC.2012.2207838 Conference Name: IEEE Transactions on Professional Communication.

[20] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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, 540–551. https://doi.org/10.18653/v1/D19-1051

[21] Purdue Writing Lab. 2021. Reverse Outlining // Purdue Writing Lab. https://owl.purdue.edu/owl/general_writing/the_writing_process/reverse_outlining.html

[22] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22)*. Association for Computing Machinery, New Orleans, LA, USA. https://doi.org/10.1145/3491102.3502030

[23] Luis A. Leiva. 2018. Responsive text summarization. *Inform. Process. Lett.* 130 (2018), 52–57. https://doi.org/10.1016/j.ipl.2017.10.007

[24] Karen Sheriff LeVan and Marissa E King. 2017. Self-Annotation as a Course Practice. *Teaching English in the Two Year College* 44, 3 (2017), 289.

[25] Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical Summarization for Longform Spoken Dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 582–597. https://doi.org/10.1145/3472749.3474771

[26] Yang Li, Sayan Sarcar, Sunjun Kim, and Xiangshi Ren. 2020. *Swap: A Replacement-Based Text Revision Technique for Mobile Devices*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376217

[27] Nina Macdonald, Lawrence Frase, P Gingrich, and Stacey Keenan. 1982. The writer's workbench: Computer aids for text analysis. *IEEE Transactions on Communications* 30, 1 (1982), 105–110.

[28] Kristin Messuri. 2016. Revision Strategies. *The Southwest Respiratory and Critical Care Chronicles* 4, 14 (April 2016), 46–48. https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/263 Number: 14.

[29] Kristin Messuri. 2016. Writing Effective Paragraphs. *The Southwest Respiratory and Critical Care Chronicles* 4, 15 (July 2016), 86–88. https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/290 Number: 15.

[30] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[31] Michael J. Muller and Sandra Kogan. 2012. Grounded Theory Method in Human-Computer Interaction and Computer-Supported Cooperative Work. In *The Human–Computer Interaction Handbook* (3 ed.). CRC Press. Num Pages: 21.

[32] Graduate Writing Center of the Center for Excellence in Writing. 2007. Strategies for Drafting & Revising Academic Writing. https://www.tnstate.edu/write/documents/DraftingRevisingEves2007.pdf

[33] University of Wisconsin-Madison. 2021. Reverse Outlines: A Writer's Technique for Examining Organization. https://writing.wisc.edu/wp-content/uploads/sites/535/2018/07/reverseoutlines_uwmadison_writingcenter_aug2012.pdf

[34] Sam Park. 2008. Reverse Outlining Worksheet | Student Learning Center. https://slc.berkeley.edu/writing-worksheets-and-other-writing-resources/reverse-outlining-worksheet

[35] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the Special Issue on Summarization. *Computational Linguistics* 28, 4 (Dec. 2002), 399–408. https://doi.org/10.1162/089120102762671927

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]* (July 2020). http://arxiv.org/abs/1910.10683 arXiv: 1910.10683.

[37] Jeba Rezwana and Mary Lou Maher. 2022. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* (feb 2022). https://doi.org/10.1145/3519026 Just Accepted.

[38] Melissa Roemmele and Andrew S Gordon. 2015. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer, 81–92.

[39] Laura Saltz. 1998. Harvard College Writing Center - Revising the Draft. https://writingcenter.fas.harvard.edu/pages/revising-draft

[40] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages. https://doi.org/10.1145/3450741.3465253

[41] Abigail See, Peter Liu, and Christopher Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Association for Computational Linguistics*. https://arxiv.org/abs/1704.04368

[42] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* (jan 2022). https://doi.org/10.1145/3511599 Just Accepted.

[43] Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education* 131 (2019), 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

[44] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. 2020. *TexSketch: Active Diagramming through Pen-and-Ink Annotations*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376155

[45] Reid Swanson and Andrew S Gordon. 2008. Say anything: A massively collaborative open domain story writing companion. In *Joint International Conference on Interactive Digital Storytelling*. Springer, 32–40.

[46] Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2018. Controllable Neural Story Plot Generation via Reinforcement Learning. arXiv:1809.10736 [cs.CL]

[47] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with Documents: An Exploration of Document-Centered Assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3343413.3377971

[48] L. Danielle Tully. 2019. Reverse Outlines: Fueling Revision & Preparing for Writing Conferences. *The Second Draft* 32, 2 (2019), 6. https://ssrn.com/abstract=3465807

[49] Duke University. 2021. Revising Process | Thompson Writing Program. https://twp.duke.edu/sites/twp.duke.edu/files/file-attachments/reverse-outline.original.pdf

[50] Keith Vertanen, Mark Dunlop, James Clawson, Per Ola Kristensson, and Ahmed Sabbir Arif. 2016. Inviscid Text Entry and Beyond. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3469–3476. https://doi.org/10.1145/2851581.2856472

[51] Keith Vertanen, Kyle Montague, Mark Dunlop, Ahmed Sabbir Arif, Xiaojun Bi, and Shiri Azenkot. 2017. Ubiquitous Text Interaction. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 566–573. https://doi.org/10.1145/3027063.3027066

[52] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2words: Automatic mobile UI summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 498–510.

[53] Harriet Salatas Waters and Wolfgang Schneider. 2009. *Metacognition, Strategy Use, and Instruction*. Guilford Press.

[54] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, and Toby Jia-Jun Li. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *IUI 2022 Workshop on Human-AI Co-Creation with Generative Models (HAI-GEN 2022)*. 10.

[55] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300415

[56] Demet Yaylı. 2012. Tracing the benefits of self annotation in genre-based writing. *The Journal of Language Learning and Teaching* 2, 1 (2012), 45–58.

[57] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[58] Mingrui Ray Zhang, He Wen, and Jacob O. Wobbrock. 2019. Type, Then Correct: Intelligent Text Correction Techniques for Mobile Text Entry Using Neural Networks. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 843–855. https://doi.org/10.1145/3332165.3347924

Average Rouge Score	Central sentences	Abstractive summaries
Rouge-1 R	0.5503	0.2344
Rouge-1 P	0.5316	0.2588
Rouge-1 F	0.5355	0.2302
Rouge-2 R	0.4886	0.0705
Rouge-2 P	0.4930	0.0895
Rouge-2 F	0.4897	0.0723
Rouge-l R	0.5422	0.2127
Rouge-l P	0.5266	0.2349
Rouge-l F	0.5295	0.2087

Table 2: Results from the ROUGE analysis of the T5 model against human selected central sentences and human written (abstractive) summaries.

A ADDITIONAL MODEL EVALUATION

To assess the quality of the summaries displayed in the study, we report on an additional evaluation. Ideally, we would evaluate summary quality after each interaction by comparing the automatically created summary in each card to a human-created summary for that paragraph. This is not practical because the text and thus summaries change all the time during interaction.

Instead, we thus decided to assess summarization quality on the *given articles*. To do so, we recruited three people who wrote abstractive summaries for each paragraph of the articles. They also selected what they thought is the most central sentence per paragraph.

We used this data to compute ROUGE scores (Table 2). The mean ROUGE-L F-score is ca. 0.2 against human summaries, and ca. 0.5 against human-selected sentences. This is generally comparable to the ROUGE scores reported on the benchmark dataset in the T5 paper [36]. This indicates that the T5 model performed similarly to its published benchmark when used on the domains that people wrote about in our study.

However, note that the dataset and reference generation method are different. Thus, we mainly report these values with the goal of providing a point of comparison for future work that employs AI summaries in systems and use-cases similar to ours here.

Finally, we also compared the sentences selected by our central sentence method against the sentences selected by the three human annotators. They matched (i.e. same sentence chosen by system and annotators) in 48 % of the cases.