Trust and Reliance in Human-AI Collaborative Text Summarization

RUIJIA CHENG*, University of Washington, USA
ALISON SMITH-RENNER, Dataminr, USA
KE ZHANG, Dataminr, USA
JOEL R. TETREAULT, Dataminr, USA
ALEJANDRO JAIMES, Dataminr, USA

As automatic text summarization systems becoming increasingly important in people's lives, it is crucial to understand people's needs in trust and reliance when they are *interacting* with or *assisted* by AI. Working towards this goal, we present this exploratory study: we first designed prototypes to represent five different types of interaction in AI-assisted text summarization; we then interviewed 16 users, aided by the prototypes, to understand their expectations, experience, and needs regarding reliance and trust with AI in text summarization. We discussed the initial design considerations based on our findings.

CCS Concepts: • **Human-centered computing → User studies**.

Additional Key Words and Phrases: human-AI collaboration; text summarization; AI-assisted text generation; user study; trust and reliance in AI

ACM Reference Format:
Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R. Tetreault, and Alejandro Jaimes. 2022. Trust and Reliance in Human-AI Collaborative Text Summarization. In *Workshop on Trust and Reliance in AI-Human Teams, at CHI2022, April 30, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

In this era of rapid information consumption, access to high-quality summaries, such as online news highlights and research paper abstracts, is increasingly important. However, summarization is difficult for humans, demanding high cognitive load and expertise [16]. Algorithmic approaches can automate summarization by generating many summaries quickly, while still require large collections of high-quality human-written summaries for training. At the same time, while it is difficult and slow for humans to write summaries, human-generated summaries often outperform machine-generated ones [12, 17]. Given the complementary skills of human and machine, could summarization benefit from human-AI collaboration?

Traditionally in text summarization, AI systems leverage human input in the data preparation [22] or final evaluation [19] stages. Novel systems have emerged in recent years, exploring new interactions between human and AI in text summarization. In our prior work [9], we identified five distinct types of these human-AI interactions from literature on text summarization and more general text generation tasks (illustrated in Figure 1):

*The research work presented in this paper was conducted while the author was interning at Dataminr.

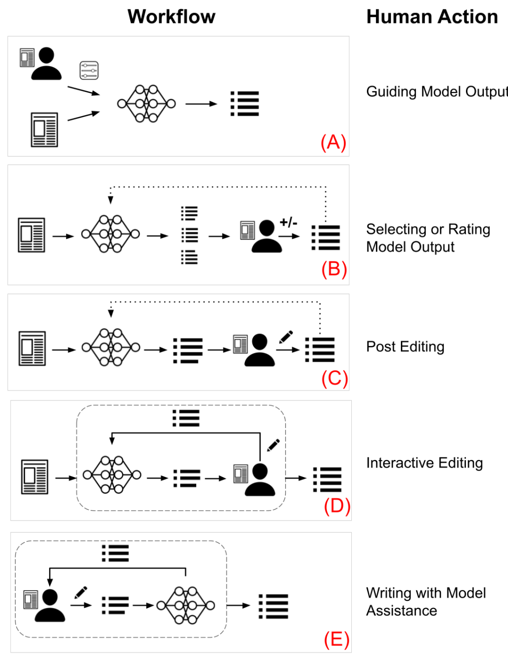

Fig. 1. Five human-AI interactions in text generation from Study 1, illustrated as summarization tasks.

- **Guiding Model Output**: humans provide preferences to the model (e.g., style, length) and the model produces summaries based on those inputs (e.g., the systems of Clark et al. [10], Passali et al. [25]).
- **Selecting or Rating Model Output**: humans select or rate model-generated summary candidates, either to choose the final output or for online training of the model (e.g., Bohn and Ling [4], Stiennon et al. [27]).
- **Post Editing**: humans edit model output summaries, which can be used as the final output or future training data (e.g., Moramarco et al. [23], Peris and Casacuberta [26]).
- **Interactive Editing**: model iterates on human-edited text to update and generate more text for continued human editing, iteratively and in real-time (e.g, González-Rubio et al. [14], Weng et al. [28]).
- **Writing with Model Assistance**: humans write summaries while the model provides suggestions along the way (e.g., Calderwood et al. [8], Padmakumar and He [24]).

Given these emerging human-AI interactions for text summarization, it is important for us to understand how users experience each interaction and what different needs they may have. Prior research in HCI communities has paid increasing attention to the issues of trust and reliance in human-AI collaboration (e.g., [2, 5, 20]), discussing important questions such as how to support "appropriate reliance" [21] and how to design for more efficient collaboration and trustworthy experience with AI [1]. Inspired by prior work, we explore and compare between the five different types of human-AI interactions: (1) to what extend users **rely** on AI in text summarization as opposed to controlling the process themselves and (2) to what extend users **trust** with AI in text summarization. To this end, we first developed prototype interfaces to represent the five types of interactions. We then conducted interviews with 16 users using the prototypes and identified varied user needs regarding reliance and trust. These interviews inform design considerations for human-AI collaborative text summarization systems.

2 USER STUDY ON HUMAN-AI TEXT SUMMARIZATION

We present a user study evaluating interactions in human-AI text summarization through interviews aided by prototype interfaces. Our goal is not to prescribe which interface is the "best" but to achieve a qualitative understanding of user experience and needs regarding trust and reliance with each interaction to inform future research and design.

2.1 Prototype Interfaces.

We first developed prototype interfaces to represent the five interactions. While some prototypes for these interactions exist in the literature for broader text generation tasks, many include additional features and visual design that may affect users' perceptions, therefore, we develop our own set of consistent, simple prototypes for exploring text summarization specifically. Each interactive prototype, implemented in Figma[1] or Google Docs[2], allowed participants to read an online news document and generate summaries with the support of a hypothetical AI model. Figures 2-6 show the screenshots of five prototype interfaces:

(1) **Guiding Model Output**: participants could change the desired summary length and style (formal or informal) using sliders and highlight parts of the original text that should be in the summary. We asked participants what additional guidance they wanted to offer to the model.
(2) **Selecting or Rating Model Output**: participants chose from three AI-generated summaries.
(3) **Post-editing**: participants saw a text box with an AI-generated summary and talked through how they would edit it. [3]
(4) **Interactive Editing**: given an AI-generated summary (text box), participants chose possible edits to the first sentence (dropdown menu) and then requested the model to update the summary based on those edits. We asked participants to imagine an alternate interface where they could edit anywhere in the summary.
(5) **Writing with Model Assistance**: following a "wizard-of-oz" prototyping method [18], a researcher acted as an AI bot in a Google Doc. As the participants typed their summaries, the "bot" provided suggested next sentences and added comments.

All interfaces except Writing with Model Assistance contain the same original text (a news article from the articles used in the warm-up activity) that needs to be summarized on the left. The representations of different types of human-AI interactions are on the right side of the interface. All the "AI-generated" summaries, outputs or suggestions were pre-defined and written by the research team. While the human-written summaries used in this study might not necessarily imitate the quality and style of real AI-generated summaries, the goal was to provide participants with the idea of interaction. Participants were not asked questions around the content of the summaries. We used the prototypes in our user interviews to elicit needs, expectations, and experience around trust and reliance in human-AI text summarization.

2.2 Participants.

We recruited 16 participants (10 females, 6 males, all based in the U.S.) from Upwork[4] with experience in writing and editing, varied professional backgrounds, and varied familiarity with the domains of Reddit posts, online news, and

[1]https://www.figma.com/
[2]https://www.google.com/docs/about/
[3]We implemented the the post-editing interaction in Figma to stay consistent with the rest of the prototypes. Although users cannot actually edit the summary, since all of our participants have some level of experience with text editing, it was easy for them to imagine the interaction of post-editing with this prototype.
[4]https://www.upwork.com/

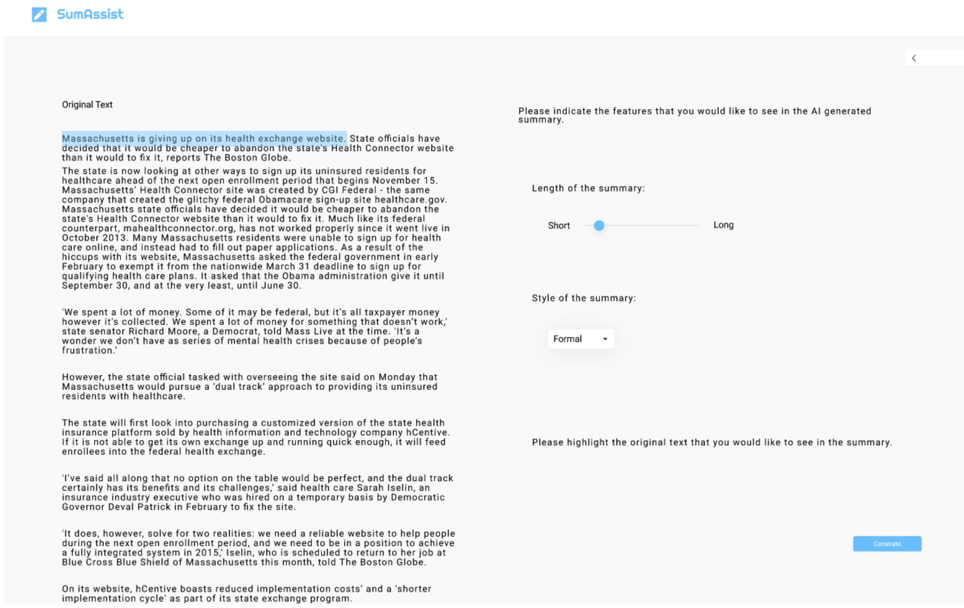

Fig. 2. The interface for **Guiding Model Output**. Users can change the desired summary length and style (formal or informal) using sliders and highlight parts of the original text that they want to include in the summary. Users can press the "Generate" button to get the "AI-generated" summary based on their inputs.

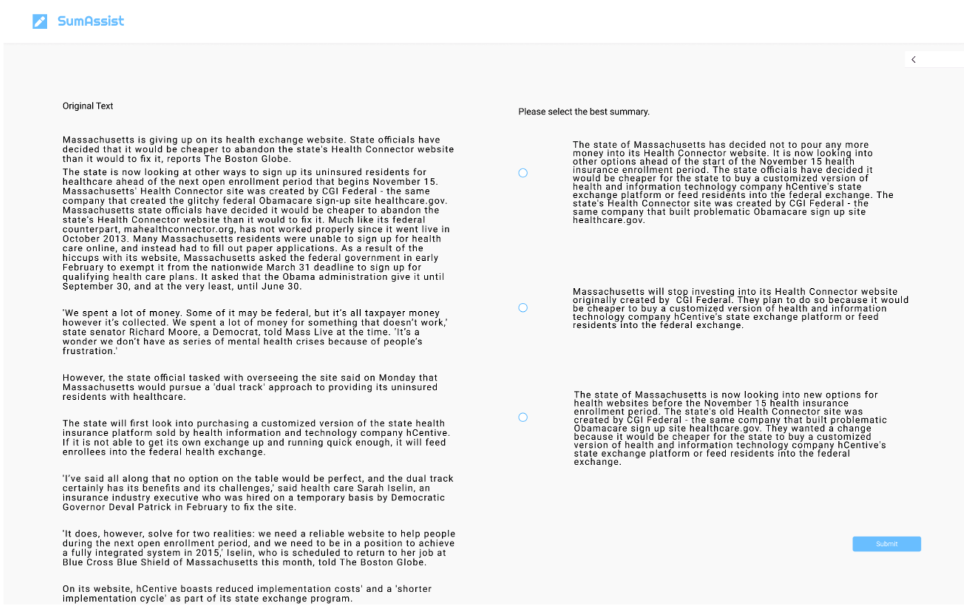

Fig. 3. The interface for **Selecting or Rating Model Output**. Users can chose the final product from three "AI-generated" candidate summaries.

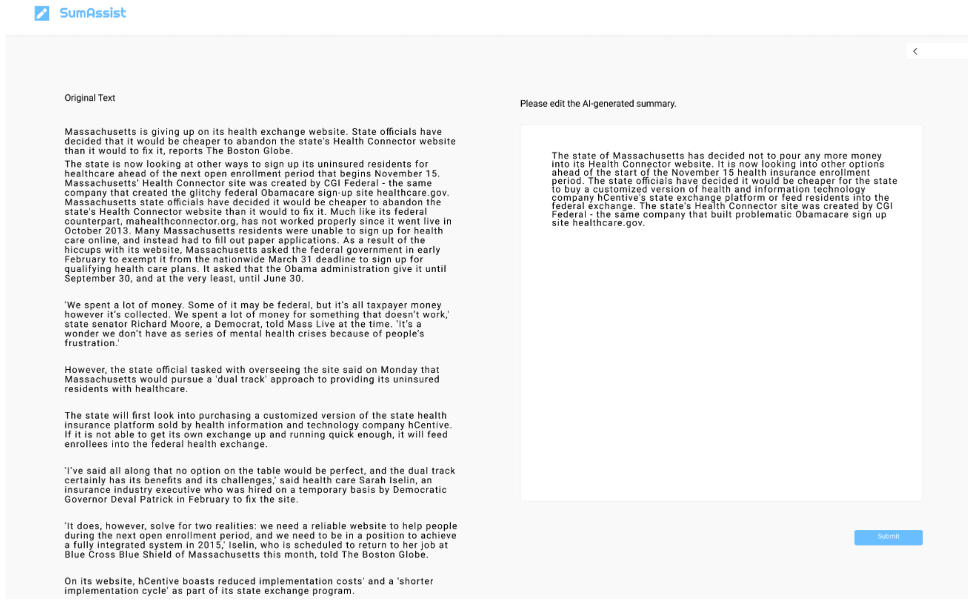

Fig. 4. The interface for **Post-editing**. Users see an "AI-generated" summary in the text box that they can hypothetically edit.

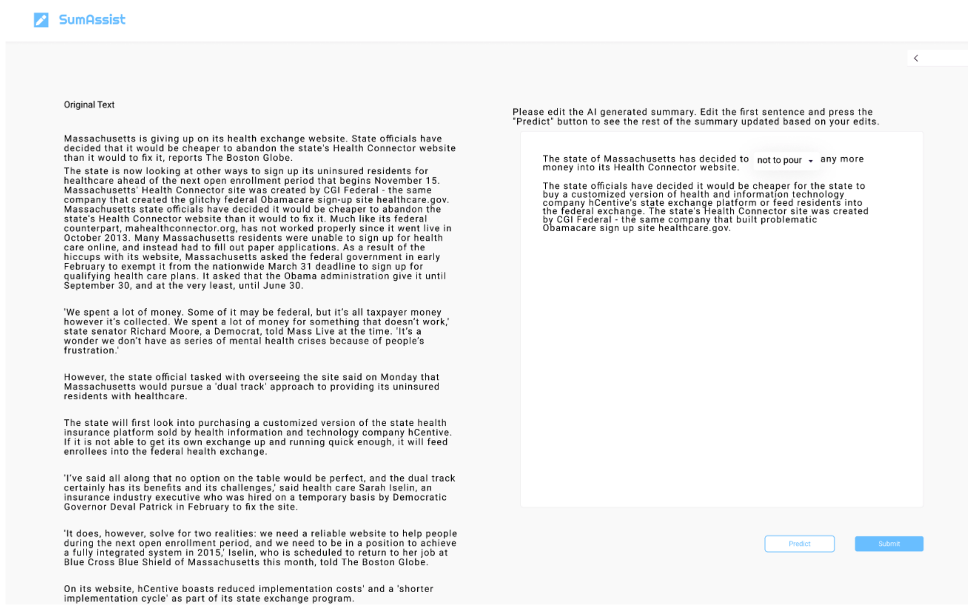

Fig. 5. The interface for **Interactive Editing**. Users see an "AI-generated" summary in the text box. They can use the drop-down menu to change certain words in the first sentence. They can then press "Predict" to request the model to update the rest of the summary based on those edits.

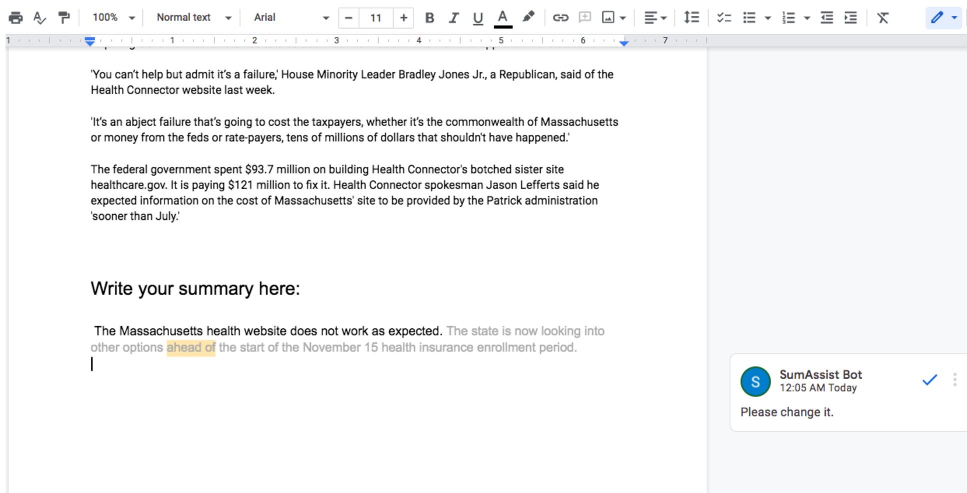

Fig. 6. The interface for **Writing with Model Assistance**. In a Google Doc, users can see the original article on the top and they can write their summary under the section "Write your summary here:". First, the user types a sentence for their summary, then a Bot (played by a researcher who log in with the "SumAssist Bot account") will insert the next sentence in gray fonts. The Bot will also insert comments on words in the user written sentence and suggest them to make changes.

U.S. government bills. The study took 2.5 hours, and we paid each participant $60.[5] The demographic details of our interview study participants can be found in Table 1.

2.3 Procedure.

Each participant first did a 60-minute offline warm-up activity less than 48 hours before the interview, where they summarized six articles (two Reddit posts on scams or finance,[6] two news articles from CNN/Daily Mail,[7] and two U.S. government bills[8]). This activity aimed to expose participants to summarization with articles written in different styles and with varied domain contexts.

Then, during the 1-on-1 semi-structured recorded video interviews (90 minutes), participants first reflected on their experience in the warm-up summarization tasks and then interacted with all five prototypes in random order as users of human-AI summarization systems. They were shown a news article from the warm-up task and also asked to imagine using the prototypes for the other documents from the warm-up. They interacted with the interfaces and received pre-determined outputs that mimicked AI assistance.

Participants were then prompted to talk through experience with each prototype. We collected and transcribed 22.6 hours of interview recordings, which were analyzed using thematic analysis [15]. We performed two rounds of open coding and developed themes reported in the following sections. We refer to participants as P1-16 with gender non-specific pronouns (i.e., they, them).

[5] Adequate payment in the United States.
[6] extracted from r/scam and r/wallstreetbets
[7] https://paperswithcode.com/sota/text-summarization-on-cnn-daily-mail-2
[8] https://www.tensorflow.org/datasets/catalog/billsum

ID	Gender	Age	Occupation	Education	Experience: summarization	Experience: editing	Familiarity: government bills	Familiarity: online news	Familiarity: Reddit posts
P1	M	50-59	Newspaper writer	Bachelor	6	6	6	6	5
P2	M	20-29	Student	Bachelor	6	6	6	6	6
P3	M	30-39	Student	Bachelor	5	7	5	7	7
P4	F	60-69	Freelance editor	Bachelor	5	7	5	7	5
P5	F	30-39	Freelance editor	Doctorate	7	7	5	7	5
P6	F	30-39	Project manager	Master	6	6	5	7	7
P7	M	30-39	Freelancer editor	Bachelor	5	4	3	6	5
P8	F	30-39	Freelance writer	Master	7	7	1	6	1
P9	F	30-39	Marketing consultant	Master	5	7	3	7	7
P10	F	20-29	Student	Bachelor	7	5	2	6	6
P11	F	50-59	Publicist	Bachelor	7	7	4	7	7
P12	M	60-69	Artist	Bachelor	6	6	1	6	1
P13	F	30-39	Freelance writer	Bachelor	6	7	2	7	5
P14	M	40-49	Engineer	Bachelor	5	7	1	7	7
P15	F	20-29	Student	High school	7	7	6	7	7
P16	F	30-39	Student	Bachelor	7	7	7	7	7

Table 1. Demographic information of interview participants. All the information are self-reported by the participants. All the participants were based in the United States. Column "Experience: summarization" reports their answers to the question "rate the following statement: 'I am experienced in text summarization' on a scale of 1 to 7, with 1 being least experienced and 7 being most experienced." The other columns on experience or familiarity reports their answers to questions in the same format.

2.4 Findings

2.4.1 General Expectations & Needs.

Different desire for control. While expecting AI-assistance to improve summarization, participants expressed different opinions on how much they might rely on AI or how much control over the summarization process that they wanted. Most participants agreed they at least wanted the ability to proofread the AI-generated summaries, or to *"have the final say"* on whether it was good as a final product (P3). Some said this *responsibility* was a habit of professionalism; others were cautious of the work done by AI and wanted to ensure quality: *"it was drilled into my head that these devices are tools and they can fail...we're always responsible for overseeing what the computer does"* (P7). Beyond simple editing, participants had a varied desire for control. Some felt summarization was *"not necessarily a creative enterprise"* and, therefore, were willing to *"relinquish a little bit of control to AI"* for efficiency (P3), while others wanted to participate in the entire generation process. These participants preferred to compose their own summary using AI strictly as an aid, e.g., *"it would just simply be used as a tool for me, not as something to replace my work"* (P12). Many felt uncomfortable using AI-generated summaries directly or after only proofreading edits due to the sense of *"plagiarism"*, and as a result, wanted the ability to rewrite summaries into their own words. Desired control could also vary by situation. For example, P7 wanted more control when summarizing for the bills because that was a more *"serious and important task,"* while P8 would be more lenient when summarizing the Reddit posts: *"even [the summary] doesn't capture everything, it is good as long as the summary kind of outlines the the key points of the article."*

Assumptions on how AI works and whether to trust AI.. Participants have their own theories on how AI works and assumption on whether AI would work well in certain situations. Such pre-existing opinions would impact how much they trust the AI to perform the tasks. Some participants thought AI would work best for formal text such as the Bills. Their rationales was that while the Bills were written in a format that was hard for human to parse, the formalized structure might actually be easy to program and for a computer to analyze. In P9's words, *"a bill would be easier to summarize only because it looked like it had a good format. It had the problem it was addressing, these are the people who will be impacted, this is how the program will work, this is who will fund it... So it would be easy for a coder to develop a software that automatically pulls those key points from the bill."* On the other hand, they stated that they would trust the AI less when it was summarizing informal and opinionated text, such as the Reddit posts. For example, P3 worried that AI would not be able to preserve the humanized aspects of the posts: *"it's a lot of personal linguistic tics, stylization, bias. (Those are) stuff that's trickier for an algorithm to sort of clock on its own, I guess."* P7 shared that AI might not be trustworthy for summarizing personalized social media writing because it would not take the background information about the person into account: *"you don't even know who that person is, you don't know their age, their language competency and stuff. So I would think that the accuracy would be less trustworthy for that."* (P7)

Need to understand AI to reduce over-reliance and boost trust. Participants were concerned that they might rely on AI too much and lacked confidence to correct it even when it was wrong. For example, P7 felt AI-generated summaries were an *"authority that has given you this thing"*, saying that *"for most people, if presented with something, they're going to go with it."* As a result, users could lose confidence when they disagree with the AI. P8 shared their hesitation to dramatically edit AI-generated summaries: *"it's almost feeling like you're pivoting against the AI...should I question what the AI thinks is important?"* This apprehension might increase when participants are summarizing for unfamiliar or difficult documents. Specifically, some anticipated a lack of trust when summarizing challenging articles because they could not reasonably assess the AI's output: *"I probably wouldn't use it for a lengthier subject that I wasn't familiar with...just because I wouldn't know if the AI was writing something I wanted to write"* (P6).

To foster trust, many wanted information about *how* the AI generated the summaries or suggestions—why the certain information is included and whether there were any hidden presuppositions by the model. For example, P8 said, *"knowing, in a very basic sense, how the AI is generating these summaries, [will] give me a good idea of essentially how much I can trust it."* As the prototypes did not include explanation features, participants noted that they did not trust the AI since they did not understand the mechanism, as P12 put: *"there's too many variables that you don't know. Too many unknowns for me."*

2.4.2 Interaction-Specific Experience & Needs. We report participants needs and expectations on control and trust for each of the five interfaces.

Guiding Model Output. Most appreciated the control over the summary generation by adjusting parameters. For example, P8 liked the text highlighting feature, as *"it gives you the amount of control in terms of being able to choose the parts that you think are important."* Many envisioned using the interface to customize the summary for their target audiences. For instance, P6 imagined using it to tailor summary styles for different colleagues: *"with my staff, I would use the short style...for my boss, I might use a longer formal summary to look a little more professional."* On the other hand, some were concerned about the lack of editing control: *"it doesn't have as much control as it seems. When you get to this [final] stage, you're stuck with it"* (P7).

Participants felt they had a reasonable understanding of the AI mechanism in this prototype and thus could trust it. Since they could change parameters (e.g., length, style) to experiment with different aspects of the summary, they better understood the process: *"we could [trust it more] maybe because I can play around with it. The long and short allows me to kind of have control"* (P12).

Selecting or Rating on Model Output. Despite the efficiency advantage, many complained about the lack of control, specifically the inability to influence or edit the AI-generated summaries. For example, some felt that choosing the best might not ensure quality: *"what if three of these are presented, and none of them are really good enough. Then it's just a matter of picking the least bad one"* (P7). Further, since comparing and selecting were simpler tasks than writing or editing, participants paid less attention and thought less critically: *"evaluating already written summaries and trying to decide which one is the best is different from just writing your own summary...I am not like super mentally invested in it, if I were writing my own, I'd very careful with word choices"* (P5).

Many struggled in selection as they did not know how the candidates were generated: *"how do you determine, from an AI standpoint, what information to keep and what information to get rid of? How do you determine the priority as what stays and what goes? Is it biased in any way?"* (P14).

Post-editing. Participants were satisfied with the editing control granted by this interface. For instance, P10 shared why they liked it more than the prior: *"You can edit it to however you'd like. I think the freedom to edit appeals to me a little bit more."*

Similar to other prototypes, participants hoped to see more information on how the AI generated the summary. For example, P16 said it would be helpful to visually see which parts of the article were emphasized in the AI-generated summary, so that they could decide what to focus on. Similarly, P7 imagined a quantifiable way to indicate how much of the content in the summary was matched with the original article. They hoped for *"some advanced algorithms checking to make sure that it did it right"* to decide how much to trust it.

Interactive Editing. Many valued that in addition to editing, they could also experiment with different versions due to the dynamic updates. *"it still has the human touch and it's not completely computer generated. So there's still some sort of organic thought process behind it, rather than just mechanical."* (P8) However, some viewed it pointless to iterate with the AI and would rather complete editing in a single turn: *"it's giving me a choice that I don't necessarily want...I want it to be as close to a final draft as possible, because then my editorial choices are final and have the feeling of finality"* (P3). Participants also worried about unpredictable AI actions that might impact their edits: *"I don't have any idea what the second paragraph is going to be until I make a choice with the wording of the first paragraph"* (P11); *"it is kinda stressful because if you use just one different word, it's going to change the entire thing"* (P15).

To ease this uncertainty, many wanted to understand how the AI updates based on their edits. P5 shared that they tended to discuss with coworkers on how certain choices were made—*"every word is intentional."* And, they hoped to have similar interaction with the AI, *"to know the reasoning behind the changes, the kind of logic flow,"* so that they could make better decisions on what to edit.

Writing with Model Assistance. Despite of the high control over the final output and whether to take AI's suggestions, participants wanted to control *when* they received assistance during summarization. Many viewed the auto-completion and suggestion intrusive and distracting, especially when they were not ready to receive help: *"it's harder to write* *when you have constant changes being thrown your way"* (P9). Comparing it with the Interactive Editing interface, P7 found the latter allowed more control over when AI helps: *"since you're pressing a button, you still feel like you have some control. And you have control the timing too, which is important, because, what if you want to think about your first sentence?"*

Similar to other interfaces, participants also wished to know why the AI made certain suggestions, so that they could decide whether and how to follow: *"I am a why person and I like to understand what I am doing. So if you're telling me to change something, you need to give me the reason"* (P6). In addition, some thought auto-completion might amplify human mistakes as it was learning from their writing: *"when I wrote my first sentence, I wasn't confident... And then for the bot to come in with that... it's not going to be a good summary, because I didn't know what I was writing."*

3 DISCUSSION & DESIGN IMPLICATIONS: FOSTER APPROPRIATE TRUST OF AI.

Our findings echo literature that humans can both over- and under-rely on AI systems [6, 7]. For example, consistent with Bhat et al. [3], some participants' viewed AI-generated text as an authority and anticipated that they would be conservative on making edits. Others were uncertain on the reliability of the AI-generated text or suggestions, especially when working with important text. Humans need *appropriate* trust of AI for collaborative text summarization (or other, more general writing tasks), so they can rely on "good" AI suggestions and ignore unneeded or "bad" suggestions. In the following, we outline some initial discussion toward designing for appropriate trust.

Support human validation of AI outputs. Systems should support users to understand *how* the model generates text, so that they can decide whether to rely on it or not. One technique is to allow humans to participate in the model decision process, such as the **Guiding Model Output** interaction which allowed users to specify preferences and experiment with different outputs. Systems can also offer explanations to model mechanism, perhaps through visual representations [13, 29]. Finally, systems should provide *contextual support* so that users who are unfamiliar with the domain and background can be empowered to evaluate the AI-generated summaries. For example, interview participants, regardless of interaction case, had issues working with AI when summarizing Government Bills, as they were unfamiliar with the format and jargon. To this end, systems should equip users with sufficient context, so that they can effectively judge the quality of AI suggestions and make decisions accordingly. These contextual supports could include embedded dictionaries, resource search, or Q&A support.

Allow humans to have the final say. In general, humans like to have the *"final say"* on AI-generated text. Even when participants' role was choosing model output, they still wanted the option to edit to ensure quality. As such, future human-AI text generation systems should provide editing options for the final output. Beyond that, when humans write with AI assistance, they desire control over the *timing* of when AI steps in. Unsolicited auto-completion and suggestions disrupted participants' writing experience. For example, users might want writing suggestions only on-demand, instead of models provided them automatically.

Customize the extent of AI interventions. While humans in general like editing support such as dictionaries or substitution suggestions, participants were more skeptical about dynamic updates in the **Interactive Editing** case, as AI may make major changes when they intended to make only minor edits. Therefore, they desired to adjust the *extent* of AI-predicted updates based on their intention. Echoing literature on predictable AI systems [11], similar systems should consider user intentions and empower users to preview AI actions. Systems could also model human editing intention, perhaps via action history like number and location of edits, and adapt AI actions accordingly to better serve user goals.

4 CONCLUSION

We developed prototype interfaces for five different human-AI interactions in text summarization and interviewed 16 users, uncovering varied user needs regarding reliance and trust when collaborating with AI in text summarization tasks. Considering both general and interaction-specific user experience and needs, we outlined considerations for researchers, developers, and designers when human-AI collaboration in text summarization systems, such as supporting human validation and control of AI outputs.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. 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 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing?. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. 116–121.

[4] Tanner Bohn and Charles X Ling. 2021. Hone as You Read: A Practical Type of Interactive Summarization. *arXiv preprint arXiv:2105.02923* (2021).

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[8] Alex Calderwood, Vivian Qiu, K. Gero, and Lydia B. Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study. In *HAI-GEN+user2agent@IUI*.

[9] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2022. Mapping the Design Space of Human-AI Interaction in Text Summarization. In *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL '22)*. Pending publication.

[10] 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*. 329–340.

[11] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2021. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Frontiers in Robotics and AI* 8 (2021), 194. https://doi.org/10.3389/frobt.2021.642201

[12] Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887* (2016).

[13] Sebastian Gehrmann, Hendrik Strobelt, Robert Krüger, Hanspeter Pfister, and Alexander M Rush. 2019. Visual interaction with deep learning models through collaborative semantic inference. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 884–894.

[14] Jesús González-Rubio, Daniel Ortiz-Martínez, Francisco Casacuberta, and José-Miguel Benedí. 2016. Beyond prefix-based interactive translation prediction. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 198–207.

[15] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. Sage Publications.

[16] Suzanne Hidi and Valerie Anderson. 1986. Producing Written Summaries: Task Demands, Cognitive Operations, and Implications for Instruction. *Review of Educational Research* 56, 4 (1986), 473–493. http://www.jstor.org/stable/1170342

[17] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. *arXiv preprint arXiv:2104.14839* (2021).

[18] John F Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 193–196.

[19] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561* (2021).

[20] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. 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/10.1145/3290605.3300717

[21] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.$50_3$0392 PMID: 15151155.

[22] Elena Lloret, Laura Plaza, and Ahmet Aker. 2013. Analyzing the capabilities of crowdsourcing services for text summarization. *Language resources and evaluation* 47, 2 (2013), 337–369.

[23] Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating Consultation Notes with Post-Editing. *arXiv preprint arXiv:2104.04402* (2021).

[24] Vishakh Padmakumar and He He. 2021. Machine-in-the-Loop Rewriting for Creative Image Captioning. *arXiv preprint arXiv:2111.04193* (2021).

[25] Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards Human-Centered Summarization: A Case Study on Financial News. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 21–27. https://aclanthology.org/2021.hcinlp-1.4

[26] Alvaro Peris and Francisco Casacuberta. 2019. Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language* 58 (2019), 98–126.

[27] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325* (2020).

[28] Rongxiang Weng, Hao Zhou, Shujian Huang, Lei Li, Yifan Xia, and Jiajun Chen. 2019. Correct-and-memorize: Learning to translate from interactive revisions. *arXiv preprint arXiv:1907.03468* (2019).

[29] Yi Zhang, Dingding Wang, and Tao Li. 2011. iDVS: an interactive multi-document visual summarization system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 569–584.