Beyond Summarization: Designing AI Support for Real-World Expository Writing Tasks

Zejiang Shen[†], Tal August[□], Pao Siangliulue[□], Kyle Lo[□], Jonathan Bragg[□]
Jeff Hammerbacher[□], Doug Downey[□,◇], Joseph Chee Chang[□], David Sontag[†]

[†]Massachusetts Institute of Technology, [□]Allen Institute for AI, [◇]Northwestern University

{zjshen,dsontag}@mit.edu,{tala,paos,kylel,jbragg,jeffhammerbacher,dougd,josephc}@allenai.org

ABSTRACT

Large language models have introduced exciting new opportunities and challenges in designing and developing new AI-assisted writing support tools. Recent work has shown that leveraging this new technology can transform writing in many scenarios such as ideation during creative writing, editing support, and summarization. However, AI-supported *expository writing*—including real-world tasks like scholars writing literature reviews or doctors writing progress notes—is relatively understudied.

In this position paper, we argue that developing AI supports for expository writing has unique and exciting research challenges and can lead to high real-world impacts. We characterize expository writing as evidence-based and knowledge-generating: it contains summaries of external documents as well as new information or knowledge. It can be seen as the product of authors' sensemaking process over a set of source documents, and the interplay between reading, reflection, and writing opens up new opportunities for designing AI support. We sketch three components for AI support design and discuss considerations for future research.

KEYWORDS

AI-Assisted Writing, Summarization, Expert Writing, Augmented Writing, Expository Writing.

1 INTRODUCTION

The advent of large language models (LLMs) [4, 27, 32] has brought about a dramatic change in the design space of AI-assisted writing.[1] The language understanding capabilities and high-quality text generation of LLMs promise to semi-automate cognitively-demanding writing tasks, i.e., help produce outlines or even generate long and grammatically correct paragraphs based on a short natural language input prompt.[2] As a result, there is growing interest from both the research and commercial communities in exploring new designs for intelligent writing support systems, including supporting creative story writing [7, 21, 31], blog posts or email composition,[3] personal knowledge management,[4] and so on.

While prior work has explored many exciting applications for LLMs, we argue that **expository writing** is a task that is understudied in existing work on AI augmented writing. We define expository writing pieces as articles that summarize facts and produce new knowledge or information. For example, this could be researchers collecting and reading multiple papers to write a survey paper [23], or doctors reviewing clinical notes to devise a treatment plan [18]. In these cases, authors not only summarize the source documents but also add information or bring new insights that do not exist in the source, e.g., organizing the relevant papers or synthesizing the patient's symptoms and test results to create a differential diagnosis and possible treatment options.

Compared to other types of writing that do not involve external documents, e.g., creative story or argumentative writing [21, 33], expository writing requires authors to comprehend the source, generate new insights, and faithfully reference and represent extracted information in the final article. The interaction between reading and writing brings new challenges for developing AI support, and the design is largely unexplored for incorporating the latest generations of language models. While summarization is often a necessary component, expository writing is also distinguished from document summarization tasks in terms of its goal to bring about new information that does not exist in the source.

Expository writing can be seen as a **sensemaking** process [29], and different types of sub-tasks are involved: typically, authors start with iteratively exploring and reading multiple relevant documents to *identify and extract key evidence*, then they *organize the evidence into useful schema* and further *synthesize into coherent writing* to communicate new knowledge or information [24]. Therefore, the role of AI may vary during the process of writing, and we argue it is important to design different types and levels of AI support to maximally help the authors while minimally influencing or shaping their opinions. In one approach, in the early stages of writing, the writers would initiate and drive the work and AI should only provide limited supporting functions. As ideas manifest and authors have a better sense of the writing content, AI could assume more responsibilities with authors "supervising" the model's work. For example, when starting writing a survey paper, the authors come up with the query to find relevant papers first, and AI helps execute the search and discover related documents; after authors read the retrieved papers and come up with ideas for the writing, AI can help generate the writing text based on authors' ideas, and the writers only need to proofread model output. AI discovery helps the user learn better, and editing support makes the writing more efficient and polished; most importantly, writers are in full control of the thinking and the ideas included in the produced articles.

Expository writing occurs frequently in many real-world tasks, and we argue that *real-world expository writing* is a high value open problem for AI augmented writing support, with many challenges and opportunities in the space. Many realistic expository writing

[1]In the remainder of this work, unless otherwise specified, AI-assisted writing refers to the use of LLMs to support writing.
[2]https://platform.openai.com/examples
[3]For example, commercial tools like copy.ai, https://copy.ai and Lex, https://lex.page/.
[4]Including apps like mem, https://get.mem.ai and Notion-AI, https://www.notion.so/.

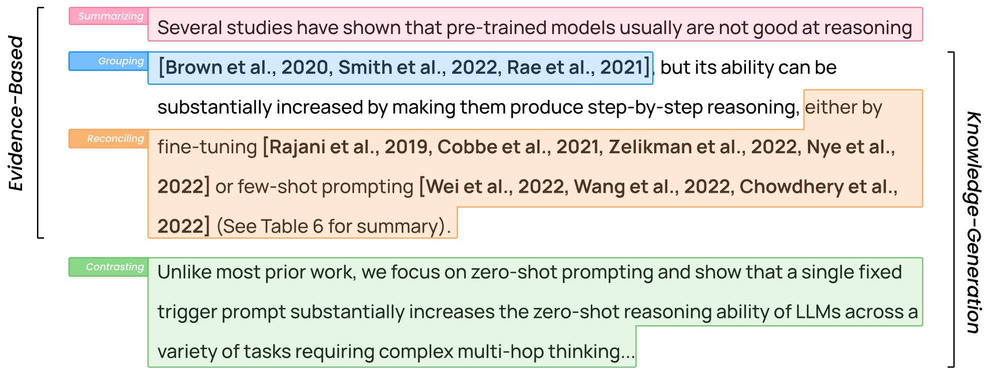

Figure 1: Dissecting an expository writing piece. In this example, the writing is a paragraph of a paper's related work section [20]. While it contains summaries of relevant previous work, it also elects to group several papers together as they describe similar concepts, or reconcile different findings under a shared framework. Finally, it contrasts the current paper with the previous work. The writing is **based on evidence** from existing work, but also **presents new knowledge** about the relationship between existing work as well as the novelty of the current paper, which are the final product of the authors' **sensemaking** process during the writing.

tasks require domain experts, e.g., describing key events in lawsuits [30], explaining scientific concepts or ideas [19], and briefing on patients' conditions and treatments [1]. Successful augmented writing systems for these tasks stand to both reduce the expensive expert hours required to perform the writing, and to improve the quality of the output (e.g. Bell et al. [3] reports that 1 in 5 patients find a mistake in their clinical notes written by doctors or nurses). Often, supporting such real-world tasks also allows us to draw upon existing rich repositories of example data and established evaluation protocols, which can further the development of future AI augmented writing systems. In the following sections, we formally define the expository task and sketch components of the design that assist the writing process rather than just the final article.

2 CHARACTERIZING EXPOSITORY WRITING

In contrast to other writing tasks, expository writing has two unique characteristics: it is *evidence-driven* and *knowledge-generating* as illustrated in Figure 1. Formally, given a set of *source documents* and a collection of *writing objectives*, expository writing aims to compose a *piece* that synthesizes the information from the source and produces *new information* in accordance with the objectives. There are different ways for authors to synthesize the source, including but not limited to the *selection*, *grouping*, *contrasting*, or *reconciliation* of multiple pieces of similar or different (even conflicting) statements, and the authors' synthesis brings information not present in the source documents. For example, a doctor might *select and group* a set of relevant observations from clinical notes, and *reconcile* with a possible condition to achieve the goal of *devising* a treatment plan. The writing objectives guide both the reading and synthesis process, and they can also be updated given new insights generated during the course of writing. Unlike a summarization task, here, it is not necessary to involve all source documents in the final writing: in fact, the choice to not include a document also constitutes new information in the writing product (i.e., by indicating the document's relative importance to the goal).

3 DESIGNING AI SUPPORTS FOR EXPOSITORY WRITING

To optimally involve LLMs to help expository writing, we argue that there are three components: 1) supporting the reading and evidence-seeking for correct and efficient understanding of the source, 2) assisting information synthesis and the production of new knowledge and ideas, and 3) facilitating text composition to communicate relevant evidence and insights. Expository writing pieces aim at bringing about new knowledge or perspectives, but it takes significant effort for authors to comprehend the source and convert the thinking into writing. By reducing the cognitive load and the interaction costs for information extraction and sensemaking during the reading, it can help authors focus on the reasoning and idea generation [16, 23]. On the other hand, while LLMs have demonstrated strong performance in document understanding and text generation, they currently suffer from hallucination [15] and are limited in reasoning [4, 32] and long-form generation capabilities. As such, it is sub-optimal to use LLMs to generate the whole writing piece altogether and post-edit [6]. A modular design that provides varying AI support at different stages of writing can be most helpful, and we detail the components as follows.

Augmenting Reading and Collecting Evidence. Reading relevant documents to gather evidence is a crucial early step for expository writing. While there have been significant research efforts on document discovery and comprehension (e.g., for scientific documents [8, 9, 14, 17, 25]), AI tools for supporting *reading for writing* is relatively sparse [12, 16, 26]. Experts typically expend significant manual effort reading through many source documents to identify key relevant information to help them synthesize new knowledge.

This is referred as "establishing the working memory" [2, 13], a key step during the cognitive process of writing. One existing approach to facilitate reading is using language models to automatically generate summaries for long documents [4, 34], but they can also be prone to hallucination [15]. More importantly, Ziegler et al. [35] shows that recent instruction-tuned models are merely "smart copiers" when performing summarization, and overly trusting the model outputs risks biasing authors' views and may lead to authors collecting inaccurate evidence.

As such, authors' reading and understanding the source still plays a central role, and designing for *reading for expository writing* should focus on the *augmentation* of reading and extraction of key information from source documents [16], but at the same time providing support for verifying the extracted evidence. For example, the system can automatically generate a list of the key facts relevant to an author's goal from a set of source documents, and guide the authors to examine and check the facts' correctness by linking each one back to relevant parts of the original source documents.

Supporting Information Synthesis. As an important middle step between reading and writing, the writer typically takes a long time to inspect and reason over the evidence collected and produce new ideas to be written, e.g., finding the similarities and differences between papers and reconciling them under a shared framework as illustrated in Figure 1. Synthesizing a large of collection of information gathered across source documents can be challenging, and different interfaces and techniques have been developed to support this step [5, 12, 26]. Language models can be involved to improve the efficiency and accuracy of these approaches by, e.g., automatic grouping based on the textual representation [28] or enabling semantic search. Beyond better organizing evidence to reflect an author's mental model, a system could also leverage LLMs for inspiration during synthesis [11]. One example is proposing connections between evidences and suggesting relevant/different topics to discover, though it is important to also provide mechanisms so that authors can verify and ground model-generated ideas with relevant evidence. The involvement of AI models is not intended to be a replacement for authors' thinking and contemplation but enhancement for producing new ideas, and discovering otherwise unseen connections or findings.

Facilitating Text Composition. The translation of ideas into language has been considered to be equally (if not more) cognitively demanding and challenging as the other components: as Flower and Hayes [10] puts, it needs "juggle all the special demands of written English" and "the process can interfere with ... planning". The latest language models have demonstrated strong capabilities in generating fluent and high-quality text, and some studies suggest the generations are even on par with freelance writers for certain summarization tasks [34]. It is promising to incorporate language models to support text composition, but several design choices need to be considered to reduce the risk of generation errors and biasing authors' opinions [22]. For example, it may be important for language models in this case to be evidence-aware and explicit—the generation should only be invoked when the authors call it, and relevant evidence and a short prompt about the writing intent needs to be provided. An instance of model generation could be relatively short, e.g., one or two sentences, with the authors calling it in multiple turns. Compared to generating long completions at one time, e.g., a whole paragraph, we speculate that this could decrease the possibility of erroneous and biased generation, and decrease the cost of verification by the authors.

4 CONCLUSION

Expository writing is genre of evidence-driven and knowledge-generating writing that takes place in many real-world settings. In this paper, we argue that the unique characteristics in expository writing open up new opportunities for designing AI support. We sketch the components of the design and highlight considerations and challenges for future implementation.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and suggestions. We appreciate the helpful discussion and advice from Oren Etzioni, Arvind Satyanarayan, Dennis Wei, Prasanna Sattigeri, Subhro Das, Barbara Lam, Lauren Yu, and Ruochen Zhang.

REFERENCES

[1] Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4794–4811. https://doi.org/10.18653/v1/2021.naacl-main.382

[2] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559.

[3] Sigall K Bell, Tom Delbanco, Joann G Elmore, Patricia S Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G Leveille, Thomas H Payne, Rebecca A Stametz, Jan Walker, et al. 2020. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA network open* 3, 6 (2020), e205867–e205867.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[5] Arie Cattan, Sophie Johnson, Daniel Weld, Ido Dagan, Iz Beltagy, Doug Downey, and Tom Hope. 2021. Scico: Hierarchical cross-document coreference for scientific concepts. *arXiv preprint arXiv:2104.08809* (2021).

[6] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R Tetreault, and Alejandro Jaimes. 2022. Mapping the design space of human-ai interaction in text summarization. *arXiv preprint arXiv:2206.14863* (2022).

[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*. 1–19.

[8] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Annual Meeting of the Association for Computational Linguistics*.

[9] Joseph Chee Chang et al. 2022. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. *ArXiv* abs/2022.99999 (2022).

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

[11] K. Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for Science Writing using Language Models. *Designing Interactive Systems Conference* (2021).

[12] Han L. Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with Text Across Documents. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[13] John R Hayes. 2000. A new framework for understanding cognition and affect in writing. *Perspectives on writing: Research, theory, and practice* (2000), 6–44.

[14] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2020. Augmenting Scientific Papers with Just-in-Time,

Position-Sensitive Definitions of Terms and Symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2020).

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *Comput. Surveys* (2022).

[16] Hyeonsu B Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. *arXiv preprint arXiv:2208.03455* (2022).

[17] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Joseph Chee Chang. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. *ArXiv* abs/2302.06780 (2023).

[18] Amy JH Kind and Maureen A Smith. 2008. Documentation of mandated discharge summary components in transitions from acute to subacute care. *Advances in patient safety: new directions and alternative approaches (Vol. 2: culture and redesign)* (2008).

[19] Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search. *ArXiv* abs/2203.08436 (2022).

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916* (2022).

[21] 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*. 1–19.

[22] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958* (2022).

[23] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. *arXiv preprint arXiv:2302.06754* (2023).

[24] Peter Pirolli. 2007. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis.

[25] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S. Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. *27th International Conference on Intelligent User Interfaces* (2022).

[26] Napol Rachatasumrit, Gonzalo A. Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. *26th International Conference on Intelligent User Interfaces* (2021).

[27] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *ArXiv* abs/2112.11446 (2021).

[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[29] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[30] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. *arXiv preprint arXiv:2206.10883* (2022).

[31] 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 Transactions on Computer-Human Interaction* (2022).

[32] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Anand Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *ArXiv* abs/2201.11990 (2022).

[33] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774* (2022).

[34] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *CoRR* (2023). arXiv:2301.13848 [cs.CL] http://arxiv.org/abs/2301.13848v1

[35] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).