# Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence

NIKHIL SINGH* and GUILLERMO BERNAL*, MIT Media Lab, USA
DARIA SAVCHENKO* and ELENA L. GLASSMAN, Harvard University, USA

While developing a story, novices and published writers alike have had to look outside themselves for inspiration. Language models have recently been able to generate text fluently, producing new stochastic narratives upon request. However, effectively integrating such capabilities with human cognitive faculties and creative processes remains challenging. We propose to investigate this integration with a multimodal writing support interface that offers writing suggestions textually, visually, and aurally. We conduct an extensive study that combines elicitation of prior expectations before writing, observation and semi-structured interviews during writing, and outcome evaluations after writing. Our results illustrate individual and situational variation in machine-in-the-loop writing approaches, suggestion acceptance, and ways the system is helpful. Centrally, we report how participants perform *integrative leaps*, by which they do cognitive work to integrate suggestions of varying semantic relevance into their developing stories. We interpret these findings, offering modeling and design recommendations for future creative writing support technologies.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Interaction techniques**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: creativity support, story, writing, multimodal, audio, images, audiovisual, interface, AI, human-AI interaction

## 1 INTRODUCTION

Augmented writing systems pervade human-computer interaction in everyday life, taking various forms to suit specific tasks. From spelling and grammar checkers to tappable word predictions and suggested email completion, these systems are typically designed to enhance human performance and productivity. Recent work in machine learning, intended to improve performance on these tasks and others (such as machine translation and text summarization), has given rise to formidable natural language understanding and generation models [16, 81]. These are often demonstrated by application to automated or semi-automated narrative generation tasks [1, 62, 76], an essentially creative domain. Given these advances, some recent work has begun to investigate the possibility for such models to be applied to enhancing human creativity, in a machine-in-the-loop setting [18, 22].

Much remains unexplored about how emerging methods in AI, machine learning, and natural language processing might influence creative writing, in part due to the ambiguity and variability of human writing processes. These processes go beyond the linear projection from idea to a full text; research shows how planning narratives, translating ideas into visible textual material, and reviewing are all happening and interacting throughout the process rather than simple sequential stages [36, 67]. However, this is a very familiar process for humans when communicating through writing; as every writer knows, having good ideas does not automatically

*These authors contributed equally to this research.

Authors' addresses: Nikhil Singh, nsingh1@mit.edu; Guillermo Bernal, gbernal@mit.edu, MIT Media Lab, Cambridge, MA, USA; Daria Savchenko, daria_savchenko@g.harvard.edu; Elena L. Glassman, glassman@seas.harvard.edu, Harvard University, Cambridge, MA, USA.


1073-0516/2022/1-ART1
https://doi.org/10.1145/3511599

produce a good text progression. The need for that "good idea" to be anchored and developed so that the reader can be invested takes a great deal of effort. In today's world, language generation models like GPT-2 [81], GPT-3 [15], and new ones coming down the line are typically silent on the inner processes of negotiation and decision that a human writer is working through. Additionally, possible forms contributions from these systems might take to influence writing are not limited to text; writers are able to engage multiple perceptual channels through their work: they may activate multisensory imagination through evocative imagery, invoking auditory and olfactory phenomena, and other forms of sensory description.

We propose to investigate how participants engage with a system that does the following: a multimodal writing support interface that bridges generated writing suggestions with multimedia retrieval to produce concept representations simultaneously in sight, sound, and language. We pair this interface with an extensive study that combines surveys, interaction, and semi-structured interviews during observed, think-aloud writing sessions.

Through this study, we examine and report in detail how participants receive, consider, and integrate suggestions from an intelligent tool into their writing. We explore prominent axes of individual and situational variation in these integrative behaviors, noting the different kinds of "leaps" participants make to understand suggestions and make the necessary compositional decisions and actions to incorporate new information contained in them, ranging from copying and pasting to re-writing core aspects of their entire story. We are specifically motivated by the following questions:

**RQ1** What kinds of assumptions, expectations, and understanding are brought into interacting with an AI creative writing system by different users?
**RQ2** How do different users process and integrate different kinds of writing suggestions, and how and why do they accomplish this?
**RQ3** How does this suggestion-informed interaction compare with unassisted (or potentially human-assisted) writing?
**RQ4** What does the combination and interaction of these three factors mean for intelligent writing support tools?

We design our study from a hybrid *Expectation-Process-Outcome* model (a visual depiction is shown in Fig. 1). We seek to capture prior *expectations*, which we do through what we call *explanatory models*, combining aspects of mental models and folk theories of technology. We study the *process* by closely observing participants as they write with the interface, asking questions, and encouraging them to describe and reflect on their thoughts and decisions. Finally, we include an evaluative survey through which participants report on their experience both independently and in comparison to a "blank page" style version of our interface. By combining these sources of information, we seek to document and communicate a range of behaviors, needs, creative processes, and results.

Our findings suggest that (1) different interaction approaches affect writer needs from system suggestions, (2) varied prior assumptions and explanatory models exist and may be both anchored to and adjusted during the interaction process, (3) suggestions support writing in more and less visible and direct ways, and (4) participants perform different kinds of *integrative leaps*, involving cognitive work to make suggestions useful to their writing. We interpret these findings and make commensurate design recommendations for future creative writing support tools.

## 2 RELATED WORK

There is a great deal of related work along multiple axes of this project. Here, we review significant precedents and influences on our work in six disciplinary areas. We additionally review relevant conceptual background areas that inform our empirical methodology and goals.

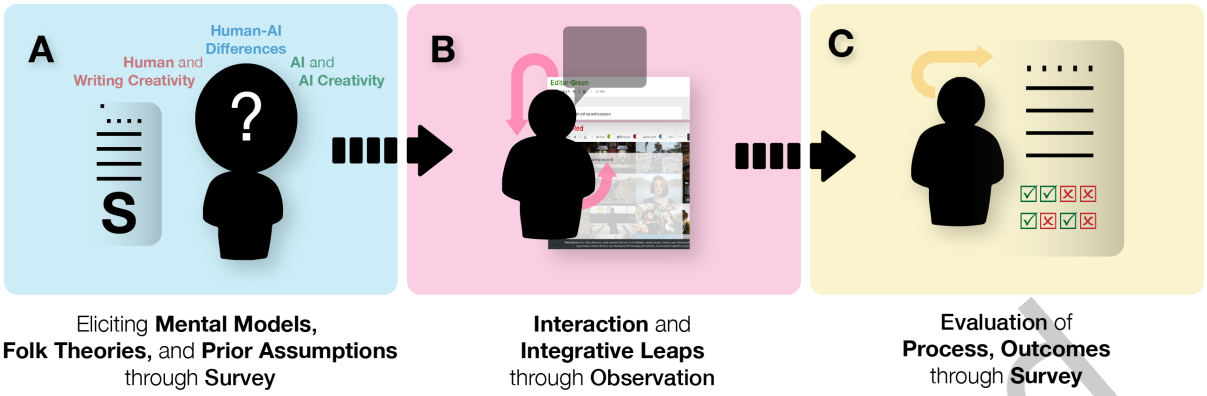


Fig. 1. **Our *Expectation-Process-Outcome* study model.** We seek to capture **(A)** each participant's "explanatory models" in areas relevant to our system, **(B)** the most salient features of their interaction and sense-making process in writing with it, and **(C)** their evaluation of the outcomes and experience.



## 2.1 Studying Writing

Flower and Hayes [36] describe what they term a cognitive process theory of writing. They model several components as part of this: the task environment includes text produced upto a given point, as well as the rhetorical problem at hand, and the writing process(es) involve planning (generating ideas, organizing them, and setting goals), translating (transforming ideas into visible text), and reviewing (evaluating and revising). At a theoretical level, these components are of interest to us because what they seek to model is how writers make decisions while writing and what factors affect this. We similarly seek to understand how writers make decisions and meaning through interaction with a supporting AI tool.

At a methodological level, they rely on *protocol analysis*, wherein participants perform an assigned rhetorical task as they think about their actions out loud and are recorded doing so. They note that this avoids the drawbacks of *introspective analysis*, in which participants report on their actions after-the-fact, observing that this tends to be colored by what they think they *should* have done. Participants are also instructed not to self-analyze under this method. While this provides a helpful starting point, our circumstance is different: participants are not following a task they know how to do and reporting on it. Rather, they are interacting with a new system and engaging in a new process (or a new version of a familiar process of writing), and as such we need more information from them to adequately understand aspects of their relationship with this system and adapted process. For this, we turn to an interpretive methodology, informed by *thick description* as we will describe later in this section.

More recent psychological approaches to studying creative writing have emphasized the role of retrieval, conceptual combination, and analogical mapping as some of the fundamental cognitive processes that explain creative cognition [35, 48, 95]. Other work on writing has emphasized social-interactive [69] and sociocultural models of writing, studied with a range of social scientific empirical methods that consider how writing activity is "*situated* in concrete actions that are simultaneously *improvised* locally and *mediated* by prefabricated, historically provided tools and practices" [78]. Robertson et al. characterized the conditions under which email-replies-suggestions generated by an AI system are perceived as problematic [84]. They highlight how social context, not just content, can influence how "brief suggestion-like email replies" that ignore social context have the potential to turn otherwise appropriate replies into inappropriate ones. Recognizing that these factors are also essential for understanding writing, especially as writing is *re-situated* and *re-mediated* with new technologies, our approach is informed by heterogeneous empirical studies of writing.

## 2.2 Writing Support

Systems that support writing tasks are ubiquitous, including word prediction, spelling and grammar checking, dictation (speech-to-text transcription), auto-completion of emails, and more. These systems have a substantial history, along with empirical investigations of their effects on writing, productivity, and behavior. For example, Smith and Goodwin [90] investigated lexicon-based single-key vs. double-key typing support for numeric keyboards in context-constrained settings, i.e., for jobs that required such text entry, indicating how computers may help resolve ambiguities arising from the former. Early work on spelling, punctuation, and grammar checking as well as additional textual analysis found that computer assistance could improve writing without substantially increasing writing time, but also found that it prompted users to think critically about their writing when they might not have prior to interaction with such systems [60]. Perhaps in contrast, Woodruff et al. [98] found that their high-level composition assistance tool was absorbed into a less critical sequence-of-suggestions text composition strategy, despite reports that it was helpful. In modern mobile devices, tools like tap-to-complete word prediction and correction [11] are commonplace, and have been shown to increase typing accuracy [38], though recent work has investigated the effects of such tools on other aspects of the writing experience [18]. Here we will discuss general purpose approaches that are designed to help develop ideas in writing, or otherwise influence or shape various writing tasks.

Moving beyond word-level predictions but retaining their contextual role, Arnold et al. provide methods to generate phrase-level continuations and demonstrate their impact by showing that phrase completions can be offered in a way that they are accepted by the user and interpreted as suggestions rather than predictions [2, 4]. More recent work by Arnold et al. examines the effects of predictive text on writing content, finding that efficiency enhancements apply not just to the process of writing but to the range of content emerging from this process as well [3]. They found that predictive text suggestionsâĂŤeven when presented as single wordsâĂŤare taken as suggestions of what to write. These suggestions often influence the length of the text generated by the user.

Some prior work has also provided multiple simultaneous suggestions to demonstrate different directions [4, 79]. Nicolau et al. identify *cardinality* as an important design factor for non-visual word completion systems [66]. More recent work by Buschek et al., conducted in parallel with ours, has examined the effects of parallel phrase suggestions on writers in an email-writing task [17]. They found that multiple parallel suggestions increased suggestion acceptance, especially for non-native English speakers. InkWell is a writer's assistant designed to help writers augment their creativity by generating various revisions of a given text, employing a synonym-based dictionary and a wide variety of soft constraints [41]. InkWell shows the importance of providing text variations to the user and how this can lead to better writing. Much of this work points to multiple suggestions being helpful, but these are often stochastically (or probabilistically) varying and cannot be reasoned about consistently or causally as complementary channels. Additionally, they do not capture the hierarchical structure of stories. Based on these findings, we decided to add two models to our system that can provide the user with suggestions corresponding to different hierarchical semantic targets in parallel.

For academic writing support, Liu and colleagues [59] introduced G-Asks, a system for improving students' writing skills, e.g., citing sources to support arguments and presenting evidence persuasively. Their system generates questions with a template-based approach by using Tregex [57], a robust algorithm used to replace keywords in a sentence and the Stanford Parser [26], a natural language parser program that works out the grammatical structure of sentences. As input, the system takes individual sentences and generates questions for the following citation categories: Opinion, Result, Aim of Study, System, Method, and Application. By doing this, the approach supports writing not only to produce content, but also to learn. Inspired by the implication of such a system we consider how suggestions might provoke cognitive processes.

Finally, several projects have considered the role of artificial agents or AI tools in creative story-writing support. Osone et al. found that more writers enjoyed working with a Japanese generative model than not [70]. Roemmele

and Gordon's Creative Help system makes suggestions that users can edit to incorporate them into stories [85]. As part of this work, they evaluate how much suggestions are edited as a proxy for suggestion helpfulness. In later work, they additionally study how randomness or unpredictability in suggestion can influence writers' attitudes, finding that increased randomness lowers ratings of factors like coherence and increases ones like perceived originality [86]. Both aspects of this work are relevant to our study, in which we examine how writers edit both suggestions and their stories to integrate suggestions, and additionally the relevance-variety trade-off in a more implicit sense, by observing and reporting the interaction. Gero and Chilton present Metaphoria, which generates metaphors to support creative writing [46]. Their work discusses both ownership and what they call "divergent outcomes" resulting from the suggestions, both of which our study addresses. Clark et al. propose a machine-in-the-loop creative writing system and study its application to stories and slogans [22]. The authors note several findings and make commensurate recommendations, some of which we build on. For example, they note the challenge of balancing between the easily-ignored "pull" interaction, and intrusive automatic suggestions. We build on this by combining direct invocation with a wait-threshold timer-based hint display. WordCraft [23] frames the collaboration between a human storyteller and AI within an open-ended dialog system with more explicit turns and turn types than what is implemented in this paper, but the effects of these design choices were not evaluated in a systematic way. Finally, like the creators of FairyTailor [10], we also introduce multimodality, which could have significant effects on findings, due to effects on the content (how are images or sounds translated to text?), the process (how is this information integrated into existing text?), and the overall experience. As such, this work additionally provides helpful background to contextualize our findings.

### 2.3 Language Models

Statistical language modeling has recently made substantial advances through the application of new techniques [93] to very large models [16] and corpora. The predominant paradigm in many natural language tasks has, as such, moved to transfer learning, in which general-purpose models are fine-tuned on downstream tasks. This mitigates issues of data scarcity, reduces training time, and improves performance. In our case, we are specifically interested in causal or autoregressive language models, which probabilistically predict successive tokens from prior ones. We fine-tune two pre-trained variants of the popular GPT-2 [81] language model in our prototype. While these models are no longer necessarily the state-of-the-art language modelers due to rapid developments in a fast-moving field, they are still competitive performers that are state-of-the-art in interactive systems, where other factors including ease of fine-tuning (flexibility), speed of response (interactivity), and open availability are important.

### 2.4 Multimodal Feedback

By presenting various communication channels, multimodal systems are considered to support human information processing by using a range of cognitive resources. This assumption is largely based on cognitive theories proposing multiple, modality-specific processing resources [5, 72]. One goal of a well-designed multimodal system is to integrate complementary input modes to create a synergistic blend, permitting each mode's strengths to overcome weaknesses in the others and support "mutual compensation" of feedback errors [71].

In addition to these cognitive benefits, multimodal feedback offers us a rich window into participants' reasoning and process of sense-making. While language processing alone demands high engagement to process and to make sense, we aim to study how a complementary blend of information representations can allow us to uncover varied aspects of participants' interaction with an intelligent system for creative enhancement. In this section, we look into our two non-textual modalities for feedback: still visual input (images) and auditory input (sound recordings). We review how each of these modalities has been used to support users on a given task, and consider how these approaches might indicate possible benefits for our task.

*2.4.1 Imagery.* iTell [56] supports retrospective storytelling with digital photos. It employs a design process based on providing support to help novice storytellers engage in the composition process like experts. To assist users with creating retrospective narratives about their personal experiences, iTell presents the users with four steps to complete: Brainstorm, Organize, Writing, and Add Personal Media. The user must finish each step before proceeding to the next step and cannot skip a step without completing it at least once. One interesting finding from the workshop conducted as part of their user study is the influence of the media modality on the novicesâĂŹ retrospective story development, how novices approach retrospective storytelling, and what is needed to make novices successful retrospective storytellers. In particular, the authors show benefits for novices to have access to mixed media in the story development process. One of the significant differences between iTell and our system is that iTell requires the user to gather any media material beforehand to retrieve and incorporate it during a writing session. Another significant difference is the lack of text suggestions to help the user in their writing.

Another example of a support tool for the development of new ideas is Design Daydreams (DD) [64]. DD is part of a suite of computational design tools that integrate ambiguity and juxtaposition into systems that users can use to discover new ideas. Using a low-tech augmented reality system to overlay digital images on top of objects visually, the Design Daydreams augmented "post-it note" fluidly extends the inspiration designers find online into the physically-interactive and collaborative brainstorming environment. Feedback suggested that the low fidelity of the tool provided a natural ambiguity that left room for interpretation as designers juxtaposed digital and physical concepts together to create new ideas. Like these projects, our visual feedback aims to discover mental constructs related to the story. It does this either indirectly, through the mood created by the image palette, or directly by layering diverse representations and allowing object or concept features to be distinguished and integrated into the developing story.

*2.4.2 Audio.* Specific attributes of the surrounding environment have been shown to support memory, foster creativity, enhance sensitivity to details, and balance cognitive load [21]. For instance, Mehta et al. found that moderate noise levels, like a coffee shop's ambient sound, facilitate abstract processing [63]. Zhao et al. built a multimodal mediated work environment, where they demonstrated effects on occupants' ability to focus and recover from stressful situations [100]. Sounds with attributable causes (i.e. where humans are able to aurally discern the source) have also been shown to impact memory [82], language learning [99], and, as a feedback modality, attention and information communication [42]. Motivated by this, we integrate an audio feedback system that retrieves sound by concept (rather than by content), to offer a semantically relevant aural dimension that may confer these benefits in the process of writing a story.

## 2.5 Interpretive Approaches

We approach our observation of participants' interaction through the lens of interpretation. Interpretation as a concept has been used in a number of papers in HCI [8, 55, 65, 89]. The interpretive perspective we maintain in this work is informed by anthropological approaches to make visible the alignments of factors of interaction that would otherwise go unnoticed due to common-sense understanding. Our theoretical approach is built on the dichotomy of social theory concepts of understanding as causal explanation (*erklÃďren*) versus understanding as interpretation (*verstehen*).

Following Max WeberâĂŹs distinction [96] between *explanation* that captures the causal sequence of actions and *understanding* that attends to the meaning of those actions, our research aims to analyze the interaction of the person with the AI system from the perspective of the latter (i.e. "meaning"). More specifically, the meaning of actions from the point of view of the participants, who organically construct meaning in the process of engaging with complex systems. As such, interpretation in this research is a form of understanding that makes it possible to discern the meaning production that occurs within the interaction between the human and the AI system. To that end, we aim to identify and observe how the interaction is influenced by the explanatory models of AI that

users have. We look at what type of conceptualization work is done on the part of the users in the process of engaging with AI, both in the world (prior assumptions and expectations) and locally in our study (impressions, integrative processes, and interactive reasoning), and how they rely on such conceptions to navigate the process and products of co-writing with AI.

"Understanding" implies the meaning of actions can be transferred through co-presence with a participant in one space, being able to build rapport, and engage with the participant so as to understand how people make sense of the world around them. For this research we aimed to complement the quantitative and qualitative data obtained from survey questions with qualitative data from semi-structured in-session interviews, observation of the participants, and what is called, in the social sciences, âĂIJthick descriptionâĂİ [43]. Thick description allows us to go beyond the observation of causal actions and acquire interpretation by the actors of not only their own actions but also of the context within which they operate. We detail our specific approach in a later section on study design.

## 2.6 Explanatory Models and Expectations of AI

We use the term "explanatory models" to refer to the super-set of two kinds of conceptual representations of computational systems, commonly referred to as "mental models" and "folk theories" respectively. Here we describe each and outline our rationale for combining them in our work.

*2.6.1 Mental and Conceptual Models.* Human-AI researchers often use the concept âĂIJmental model of AI," a term informed by psychology and cognitive science. In the context of human-machine collaboration, and even for human collaboration alone, a great deal of work has illuminated the importance of mental models in promoting team success [27]. In the case of AI, it has been shown that optimal inference does not necessarily yield optimal human-AI team performance. Bansal et al., for example, study mental models of AI performance in the context of human-AI teams [7]. They do note, however, the relevance of other types of mental models (such as those of how the system works) to collaborative settings.

Gero et al. study human-AI collaboration in a game setting, and their results suggest that understanding of the system alone insufficiently develops appropriate conceptual models [45]. The same authors distinguish between mental and conceptual models by indicating that the latter are held by those with extensive knowledge of the system, e.g., designers and experts. This follows Norman's formulation of these two terms, where he suggests that *conceptual* models are "invented by teachers, designers, scientists, and engineers," [68] noting that researchers then *conceptualize* the mental models through experiment and observation in order to produce systems and conceptual models that direct these mental models to be coherent and usable.

*2.6.2 Folk Theories.* While mental models offer insight into cognitive representations of a system's operation developed through experience, intuitive theories about the world structure learning, understanding, and cognition more broadly, in diverse ways despite a common psychological substrate [44]. Folk theories are a form of expectations that are based on some experience, but are not necessarily systematically checked [83]. Mental models are structured accounts of a system's mechanics and behavior, but folk theories and implicit beliefs arise from a great many sources of information and interaction, and are not constrained to nor will they necessarily contain an understanding of "the relationship between inputs and outputs" [40]. Folk theories may be especially salient in the domain of AI systems, given their dramatic and continued impact on culture and society. Few kinds of technical systems are as pervasive in the collective consciousness, due to rapid advances, news reports, economic incentives and concerns, and potentially profound implications for human identity and activity.

Folk theories have been captured in the study of cyber-social systems, often relating to algorithms employed in social media platforms. These theories may be elicited through *direct* investigation by researchers, often through interviews and associated methods, or *indirectly*, through inferential procedures applied to data "in-the-wild",

such as posts on a social media platform. As an example of the former, Eslami et al. elicit theories about the operation of Facebook's news feed algorithm and a designed alternative [32]. In contrast, DeVito et al. aggregate and analyze over 100,000 tweets to determine user folk theories that contribute to resistance against changes in Twitter's algorithmic content curation system [29].

*2.6.3 Why combine user mental models and folk theories?* To understand the prior assumptions, expectations, and understanding that our participants brought into their interaction with our system, we captured their *explanatory models* in related areas. Specifically, we identified related areas as AI and AI creativity, human creativity in writing, and differences between humans and AI in creativity and writing. Our system is specific enough that they are unlikely to have encountered a substantially similar system before, and are accordingly unlikely to have developed intuitive theories or mental models of our system. As such, these contextually informative areas may have bearing on their experiences, writing and sense-making processes, and evaluations of the outcome.

Creative processes with AI allow varied creators to expressively produce diverse artifacts. We aim to similarly capture complex and multidimensional explanatory models, considering both cognitive and sociocultural factors to obtain extensive representations of participants' prior assumptions, expectations, and understanding. We build on the concept of mental models to consider aspects including the user's beliefs about and attitudes towards AI and creativity, about the production of creative writing artifacts, and consider how these might affect downstream evaluations of the process of interacting with our system. We believe this approach can work inform design processes to yield tools that have clear affordances in creative contexts, and support a diversity of needs and practices.

## 3 SYSTEM PROTOTYPE

Our experimental prototype consists of two writing interfaces: **Editor-Green**, a minimal "blank page" tool, and **Editor-Red**, our augmented multimodal tool. To minimize cognitive bias when conducting our user study, we chose to give names to the editors that would seem roughly equivalent. The system also contains a server that runs language models, as well as a real-time database to track inputs, responses from the server, and interactions, e.g., interface settings. Fig. 2 shows both interfaces, including an active multimodal response in **(B)** with images[1] and sounds. Fig. 3 shows the underlying data flow through the system architecture that makes these interfaces possible.

### 3.1 Writing Interface

**Editor-Red** contains a page-like typing environment, with two suggestion blocks adjacent to it that contain suggestions (these are below the writing page on mobile devices). The two blocks offer two different types of suggestions, corresponding to text generation models fine-tuned on two different datasets. These are returned and presented through images and sounds in addition to suggested text. There is a control panel at the top, which contains some basic formatting features (text styles including heading levels, etc. as well as font formatting), and controls for invoking language model suggestions and switching multimodal response displays. One switch selects between two image presentations: by default, (the "on" position), images are displayed as a "grid" with full opacity behind the writing and response display elements. When toggled off, images are displayed as an "overlay", with multiple images stacked on top of each other and their opacity set low enough that they form an environment together. Two modality switches, one each for images and sounds, turn on and off the inclusion of these modalities, respectively. A slider can be used to adjust the volume of retrieved sounds. Finally, a "Suggestion" button launches a query for a new suggestion, and associated images and sounds, based on the current text of the

[1]Photos by Alan Ren, Matus Karahuta, Javad Esmaeili, Cassie Lopez (1), Christopher Campbell, Brooke Cagle (1), Benn McGuinness, Cassie Lopez (2), Claudio Schwarz, purzlbaum, Matheus Ferrero, Cory Woodward, Tim Photoguy, 2 Bro's Media, Nature Uninterrupted Photography, Fabe collage, Brooke Cagle (2), Ronaldo de Oliveira at Unsplash.

userâĂŹs story). Suggestions can also be invoked via the tab key, and after about 10 seconds without any writing activity, a hint regarding suggestion availability appears (indicating that tab can be pressed for suggestions). The suggestion texts are colored with a gradient to clearly distinguish them from user-written text, and are virtually "typed out" over a small amount of time to visually illustrate their narrative structure. A text field at the bottom includes credits for the presented photographs.

Several design features of our writing interface are based on popular word processing platforms. For example, a paper-shaped writing area, and a toolbar at the top for text formatting and other controls. The other design choices we made, for the new features we proposed, were refined through early prototyping and pilot testing with fourteen users, through which we discovered several usability challenges and corresponding solutions. We added tab-based suggestion invocation in addition to re-positioning the button to avoid accidental triggering based on observations made during these sessions. While we initially designed the image display as an overlaid blend underneath the writing area, we found that the grid-style display allowed for more explicit idea borrowing when desired, due to the increased clarity of individual images, thus making this display the default. We selected a number of images (20) that we found yielded sufficient variety from individual searches, but maintained the individual clarity on typical screen sizes. Finally, we had originally placed suggestions at the bottom of the interface, but found that this constrained space for writing and required more scrolling. We moved it to the right of the writing area to both position it as secondary, and allow quick glancing. We additionally added a gradient to this text to firmly distinguish it from user-written text.

To parallel this augmented interface, we provide another with the same core features, design, and layout that we call **Editor-Green**. This editor includes the text formatting features and the page-like writing environment. We use this interface as a point of comparison, based on the core features of common writing tools. The additional **Editor-Red** features are turned on and off, effectively switching between the interfaces, by clicking the interface title in the top left corner. We did this to allow flexible switching in the study context, while reducing the likelihood of accidental switching, which we observed in early iterations.

### 3.2 Language Models

In order to produce relevant suggestions, we expected that a pre-trained language model would need to be subsequently fine-tuned on a dataset containing many useful examples. However, narratives develop simultaneously at multiple hierarchical structural levels, and single suggestions do not capture any variation in this important property. As noted earlier, prior work has provided and investigated multiple simultaneous suggestions to demonstrate different directions, which points to multiple suggestions being helpful. However, stories allow us to make some domain-specific assumptions that can make these parallel suggestion channels semantically relevant. As such, we produce two variants of the base language model to capture overall plot and local description respectively, offering multiple semantically distinct channels of suggestions. Stochastic variation is also available by simply requesting additional suggestions in sequence without any additional writing (though we note this does introduce additional delay).

We fine-tuned the same language model on two different datasets, producing two final models. The base model is a medium-sized GPT-2 architecture with pre-trained weights obtained from huggingface[2]. The first experimental model is fine-tuned on a corpus of movie summaries [6], which we observe tend to contain high-level plot components and event sequences. As such, we label suggestions arising from this model as "Plot" suggestions. The second is fine-tuned on a writing prompts dataset [33], which features prompts and story responses taken from a prominent online forum for amateur fiction. Following the observation by Fast et al. that amateur fiction "tends to be explicit about both scene-setting and emotion, with a higher density of adjective descriptors" [34]

[2]https://huggingface.co/

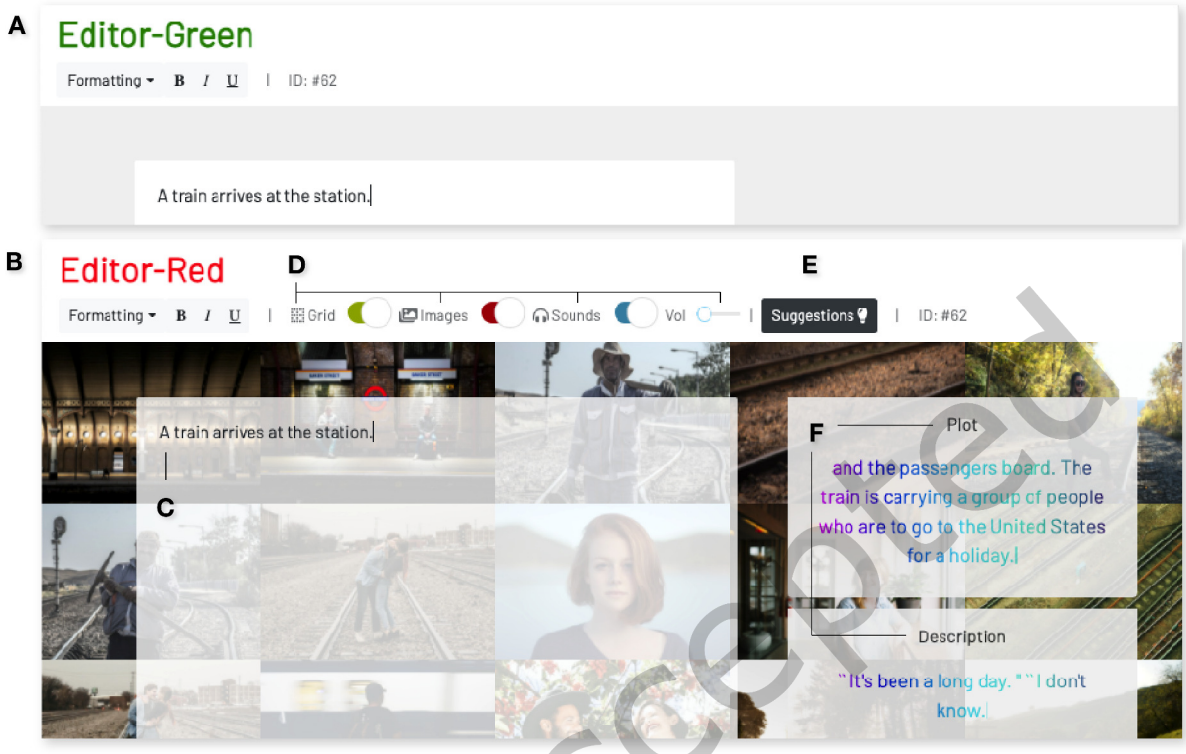


Fig. 2. **Our experimental writing interfaces. (A)** is a "blank page" editor with only basic formatting features, while **(B)** augments this with generated suggestions and multimodal feedback. In the second interface, users write text **(C)** and can request suggestion by invoking the *Suggestion* button **(E)** or using the tab key (a hint is shown after about 10 seconds of inactivity). Two types of suggestions, corresponding to text generation models fine-tuned on two different datasets, are returned **(F)** and presented through images and sounds in addition to suggested text. The user can turn on or off these stimuli, or change the image presentations to an overlay **(D)**.

as well as our own review of this dataset and the fine-tuned model, we label this second experimental model's outputs as "Description" suggestions.

For each query, our system produces responses from both models. When sampling from the models, we employ a top-k sampling strategy, with $k = 5$, temperature $= 0.5$, and a repetition penalty of 1.0; these parameter settings were based on initial experiments, i.e. looking at the models' outputs for different combinations of parameter values and making subjective judgements about quality, consistency, and relevance given a variety of input prompts. We decided on a maximum output suggestion length of 40 tokens (tokens are sub-word units, so there isn't a direct relationship to the number of words), finding that this suffices for many scenarios, and weighed against the time and computation needed for autoregressively sampling longer sequences. This cost-detail trade-off is one of many we needed to address in the design process of this prototype. Others included model size, i.e., number of parameters, for which more parameters typically result in greater coherence in modeling long-term semantic consistency but slower performance and consequently significantly greater latency, and the number of model options, with similar constraints. While other work has hosted multiple different-sized models in order to propagate this trade-off to a user decision at the interface's point of querying [9], we wanted to focus our

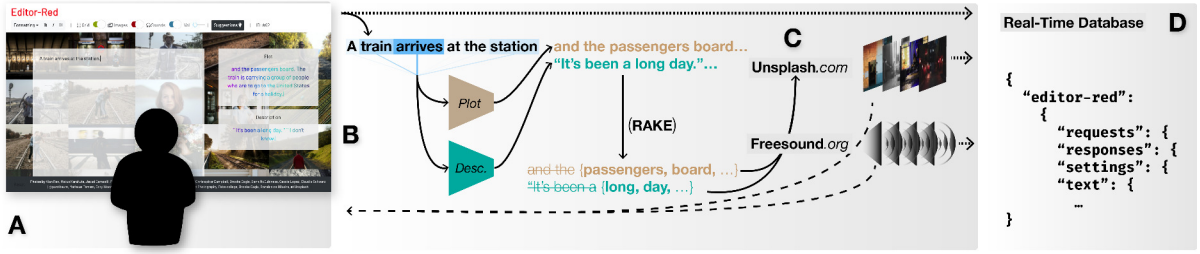


Fig. 3. **Flow of data through our system. (A)** The user enters text into the interface which is, upon request, transmitted to **(B)** a backend application. This operates two causal language models, fine-tuned for plot-level and description-level suggestions respectively. The text is tokenized and input to both, and generated suggestions are captured. Keywords are then extracted (using the RAKE algorithm) for use in the multimedia search queries: **(C)** calls to the Unsplash and Freesound APIs retrieve semantically associated image and audio content respectively, and these are sent back to the interface along with the suggestions to be presented to the user. In parallel, all use data is logged into **(D)** a real-time Firebase database. We track requests (including the state of the story at each request time), system responses (suggestions, links to media), the latest story state, and changes in settings (e.g., turning any specific modalities on and off). The logging system is replicated, for text only, in **Editor-Green** as well.

approach on the specific semantic channels of plot and descriptive detail development to support story writing and further reason about suggestion incorporation. As such, we opted to fine-tune two medium-sized models, which balance interactive responsiveness with expressive language modeling.

### 3.3 Multimedia Retrieval

Retrieving visual and auditory stimuli based on natural language descriptions is a challenging task. This is compounded for open-domain text, as in our case. Applications that do this typically need to defer to large internet media databases with search APIs to adequately support the range of possible queries with high-quality media objects. While some recent work focuses on learned approaches to semantic text-image similarity, these approaches are slow, require much data to train, and don't scale well to large databases, and so we opt for simple concept-based (rather than content-based) search of media platforms.

We use two such databases: Unsplash[3] for images, and Freesound[37] for audio. Concept-based searches for media on these platforms are typically performed with keywords rather than long-form text, and so we use the RAKE algorithm for automatic keyword extraction [87] as a preprocessing step, pooling the keywords from both model outputs (*plot* and *description*). We then query Unsplash with the output keyword list. We observe that Freesound is sensitive to multi-keyword searches and often returns no sounds in these cases, so to avoid rate limitation problems we supply only the first extracted keyword to its API to search for sounds. We apply three content filters to Freesound queries. First, we limit the duration to be between 10 seconds and 30 seconds to allow for sounds that are both long enough to contribute to an acoustic environment, and not so long as to extend retrieval time. Second, we filter out results marked as containing explicit content, after noting that these sometimes appear even when not necessarily suggested by the query. Finally, we apply a filter on the "dissonance" feature[4], which is extracted directly from each audio signal, so it is $\leq 0.4$. Since sounds need to be combined together, constraining the sensory dissonance [77] of each independent element helps to layer them effectively into a coherent soundscape.

[3]https://unsplash.com/
[4]https://essentia.upf.edu/reference/streaming_Dissonance.html

### 3.4 Data Logging

To allow detailed logging of user interaction with our prototype as well as generated suggestions, we use a real-time Firebase[5] database. This database keeps track of user-typed text in real-time as well as any changes to settings in the interface (e.g., turning sounds or images off or on), time-stamped requests with their associated input text, and time-stamped responses with text suggestions and links to retrieved media. This allows us to approximately reconstruct a sequence of writing events from a session, which we refer to in order to build the later section on *integrative leaps*.

## 4 STUDY

To investigate user interaction with our prototype and the role of user explanatory models of AI within that interaction, we designed a mixed-methods study consisting of two observed writing tasks, a four-part survey, and extensive logging of interaction data.

### 4.1 Formative Study

As part of our initial exploration, we conducted a formative study with 14 participants and an earlier version of the prototype. We called this interface MLVille, an homage to Herman Melville who famously struggled with writer's block before it was a well-documented phenomenon. With 4 out of 14 participants we conducted open-ended, interpretation-focused interviews, which allowed us to get in-depth data. These participants interpreted their actions and interactions with our prototype while performing the study tasks to provide additional context and insight. Through a broad set of survey questions, analysis of the produced text both computationally and qualitatively, and extensively documenting usability feedback, we performed several updates for our second study iteration, which is described in sections to follow. Specifically, we re-designed our interface, fine-tuned new language models on different datasets, re-oriented our study around thick description (noting the range of information and useful perspective it generated), and re-designed our survey to capture identified factors of variation and interest.

### 4.2 Recruitment

Potential participants were invited through large mailing lists associated with various departments and living groups at R1 universities, including one social sciences department and several Computer Science-adjacent lists, as well as a post on reddit. As a pre-condition for being recruited, applicants filled out a short survey to confirm that they meet the requirements of being fluent in English and being over 18 years old. The survey also contained a video tutorial that explained the features of **Editor-Red** and in order to filter for applicants who watched and paid attention to the video, they were asked to answer two screening attention check questions about the system's features. Those applicants who met the requirements were invited so as to have a balanced pool of participants who identified as native and non-native English speakers. We also made sure to have a balanced pool of participants with and without Computer Science backgrounds.

### 4.3 Participant Demographics

27 participants completed the writing task. Data from 4 had to be excluded due to firewall-related issues, mid-session server problems, and unwillingness to complete the task as instructed. All participants reported having at least a high school diploma. When asked about their disciplinary affiliation, 35% replied *Another STEM field*, 21% *Computer Science*, 13% *Life Science*, 9% *Business*, 13% *Humanities*, 4% *Social Sciences*, 4% *Medicine*. Participants' ages ranged from 18 to 45, with 48% of participants in the range of 18-22. 65% of participants reported that English is their first language. When asked "Do you struggle with writing?", 78% of the participants responded yes.

[5]https://firebase.google.com/


Fig. 4. **Study design.** Our study consists of two writing tasks, one each with **Editor-Green** (no augmentation) and **Editor-Red** (with augmentation) for 20 minutes; which interface participants used first was counterbalanced across subjects. For each task, the participant is given one of two prompts (in randomized order) to then create a story with. The two writing tasks are interlaced with sections of a four-part survey, with introductory and background components, as well as one for each writing task. The study takes approximately 75 minutes in total.

### 4.4 Study Structure

Our study design follows the structure shown in Fig. 4. We began by sending each participant a consent form in advance of the scheduled session, allowing them enough time to read and ask questions. They gave verbal consent at the beginning of the session with the interviewer and were then given an introductory overview of the study procedure, which took 3-5 minutes. Participants then completed a 10-minute introductory survey ($S^I$), designed to elicit the prior knowledge, conceptual frameworks, and beliefs that participants had about AI and its application in writing, human creative writing, and their own previous writing experience. Participants began the first 20 minute writing task, with either **Editor-Green** or **Editor-Red** depending on their group assignment. They were instructed to write a story using one of the following prompts: *The phone began to ring* or *A train arrives at the station* (alternating prompts between groups to control for the effect of the prompt). Both prompts were designed to be short, somewhat vague, and contain the beginning of some action (phone call and train arrival).

Most participants, once given the task ("write a story using the following prompt"), began writing without asking any questions. Some participants asked if there were any requirements in terms of genre, structure, or length, and we informed them that there were none. Participants were informed that they should use suggestions only if they find them helpful. Deciding when to stop writing was completely up to participants and we clearly stated that at the very beginning of the task.

After each writing task, participants completed the corresponding follow-up survey, i.e., $S^G$ (< 5 minutes) for participants who wrote in **Editor-Green** or $S^R$ (~ 10 minutes) for participants who wrote in **Editor-Red**. In accordance with standard order-counterbalancing, participants completed a second 20 minute writing task with whichever editor they had not yet experienced, followed by its corresponding survey.

Finally, all participants completed a survey that invited them to compare the two writing experiences they had during the session, as well as provide some additional demographics/background information ($S^C$). The overall duration was about 75 minutes, and participants were compensated with a $25 Amazon gift card.

Two researchers separately conducted study sessions via Zoom videoconferencing. The sessions were recorded with permission, and the researchers took notes throughout the session. Participants shared their screens during the writing sessions, and they were asked to switch this function off when answering survey questions. While writing, participants were explicitly encouraged to comment and react aloud as they wrote, processed information,

and responded to incoming suggestions and media. During the sessions, interviewers observed participants' interactions with the prototype and writing process. Additionally they prompted participants to communicate about their thought processes and experiences periodically throughout each writing session.

### 4.5 Survey

The survey consisted of four blocks. The Introductory block of questions ($S^I$) contained open questions and multiple choice questions. It was designed to elicit the prior assumptions, expectations, and understanding that participants had about Artificial Intelligence and the possibility of its application in writing and creative writing, specifically. Participants were also asked about their own writing and their thoughts about creativity in human writing.

The block of questions after writing in **Editor-Red** contained five open questions on the experience of the interaction, which was followed by a longer section that contained two grid sections with 7-point Likert-type items relating to general usability. After this, there were six multiple choice questions asking participants to provide more detailed information on their experience (e.g."When **Editor-Red** was giving its ideas, what were you paying attention to? Text, Images, Sounds, None", "I think I will enjoy using **Editor-Red** more, if..."). Finally, there was also one more 7-point Likert-type grid of items asking participants to rate statements on suggestions provided by **Editor-Red** (e.g."The suggestions made by **Editor-Red** were creative", "The suggestions made by **Editor-Red** were coherent", "I enjoyed co-writing with **Editor-Red**", "I enjoyed collaborating with **Editor-Red**"). The block of questions after writing in **Editor-Green** ($S^G$) contained two grid sections, with all items relating to the augmentation features omitted and the rest replicated.

The block on comparison ($S^C$) between **Editor-Green** and **Editor-Red** consisted of Yes/No and open questions on creative writing ("Do you consider the text that you wrote in **Editor-Red**/**Editor-Green** creative?", "If yes, in a few words, explain how it was creative. If no, explain in a few words, why not?"). There were also 7-point ordinal items asking to compare **Editor-Green** and **Editor-Red** in terms of creative writing (e.g. "In which editor was the text that you wrote more creative?") and four items on cognitive load adapted from the NASA TLX survey [47] (e.g. "Where did you feel more focused when writing a text?").

The final section contains demographic questions asking participants about their highest degree, disciplinary affiliation, age, and gender identification. Those participants who identify themselves as non-native speakers of English are asked to provide more information about levels of self-reported proficiency of various skills and depth of exposure, which we assess based on pre-existing instruments [49, 61]. In this block, there are also supplementary questions asking participants what kinds of writing they struggle with and how often they do creative writing.

All the questions throughout the four blocks are meant to elicit data for the key phenomena we were interested in: participants' prior understanding and anticipations of AI and writing using AI, how participants understand creativity and creative writing, participants' interpretation of the system's work and their explanation of engagement with the system's suggestions. Additional concepts of usability of the system, cognitive load, and agency were also included. The questions were strategically phrased in different ways (open questions, closed questions, multiple choice, Likert-type items).

### 4.6 Observation and Thick Description

Participants spent about 20 minutes writing in each editor (within each 75 minute session). This gave researchers an opportunity to capture a wide range of phenomena: participants would comment on how they usually write outside the study and how they are writing within the study, explain their process of coming up with ideas, their opinions and judgements of the systemâĂŹs suggestions, talk about how they were making decisions to incorporate or not incorporate suggestions, and give their reasons. The ability to be there with participants when

they were writing and to observe immediate reaction and, to the extent possible, raw and unmediated answers, allowed us to produce "thick description" [43], as noted earlier.

Observing the interaction with the system allows us to capture the reasoning of participants for incorporating or ignoring suggestions, and also glean how participants make sense of the interaction with the system and their strategy on structuring this interaction to support their writing. In describing the interaction with **Editor-Red**, we note that the interaction is not reduced to just getting suggestions from the system. Participants perform a task of writing a text while existing within a particular space, which is defined not only by the interface of the system and its multimodal suggestions, but also those expectations and explanatory models that participants had prior to the study and were constantly adjusting during the study. As such, we seek to capture detail about aspects of their experience that go beyond just suggestion incorporation behaviors.

## 4.7 Data Analysis

*4.7.1 Writing sessions.* The data from writing sessions consisted of (1) logged data of the texts participants wrote and suggestions they received, (2) transcripts of sessions where participants thought out loud during the interaction with the system and answer interviewers' questions, and (3) the notes that interviewers made during the sessions. After we finished running the study, interviewers watched the session videos, making additional notes and comparing them to the notes they made during the sessions. Then we entered all the data from the writing sessions into a shared document and MAXQDA[6] (software for qualitative analysis). In MAXQDA, we first used a deductive approach to code the data: we employed pre-existing concepts from research questions (such as conditions for acceptance of a suggestion, creativity, agency and ownership, etc) as codes. In the second stage of the analysis, we applied an inductive approach to code the data: in particular, the *in-vivo* method (using the words of the participants to create codes), so as to let the voice of the participants and their actual concepts structure the themes. Two rounds of inductive coding were done, followed by a process involving rearranging codes and turning in-vivo codes either into new themes, or adding them to existing codes. At this point, a second researcher did their round of coding and partially re-coded the data. The two coders discussed and reached agreement on the codes. A third round of coding by a third researcher was done to align and streamline all the codes.

*4.7.2 Survey responses.* For the answers to open-ended survey questions (expectations), one researcher performed an initial open coding (*in-vivo* method, i.e. using the words of the participants to create codes), followed by a second cycle that involved deductively applying domain concepts associated with posed questions to the initial codes. For example, in asking about differences between human and AI text production, we relied on some concepts from prior literature (such as statistical vs. symbolic processing or novelty, value, and surprise in creativity) that were closely related to the *in-vivo* codes. This was accompanied with a values orientation (i.e. trying to infer participants' values and beliefs). Then a second researcher reviewed and partially re-coded the same data, and disagreements were resolved through discussion of instances and codes themselves (labels and definitions), as well as including secondary codes for individual responses where appropriate.

# 5 RESULTS

We detail findings from the three primary components of our study: our survey to capture participants' prior assumptions and pre-existing explanatory models, observations and responses during the semi-structured interview process that accompanied their writing, and questions posed afterwards about their final thoughts and experiences during the sessions. In this section, we focus on detailing each independently before examining the synthesis of their respective data. We explicitly review specific examples of how these data interact, but note

[6] https://www.maxqda.com/

| | N | Example |
|---|---|---|
| Sparse-Abstract | 14 | "I think AI is a software that attempt to resemble the way an intelligent brain works. I suppose it bases its decision on a set of situations that are used as possible scenarios." |
| Sophisticated-Operational | 5 | "Different kinds of AI work differently. If we're talking about machine learning systems, they are trained with large corpi of data that are curated by data scientists or machine learning engineers. The algorithms for these systems find and exploit patterns in the data to then accomplish tasks. If I ask an AI something, it will look for a way to take my input and compare it to patterns in the corpus of data it was trained on." |
| Sparse-Operational | 2 | "i think it works by first being given a set of instructions, or base algorithms, which are then trained by feeding it various data sets/ user inputs. For example IâĂŸm pretty sure visual captcha companies use user input data from instructions likeâĂIJselect all the traffic lights in this imageâĂIJ o train or test their own image recognition algorithms. The data is relayed to the computer in a format it can interpret, such as code for matlab, and the computer then recognizes which configurations of code evaluate to true given the desired condition. For text, it can also scan the input for key words/tags, that make it branch down a certain path in the algorithm." |
| Sophisticated-Abstract | 2 | "There are different kinds of AI. The most simple is a seriesof if/else statements, more complicated AI might use neuralnetworks and deep learning. AI could get information from anysource a computer can: file input, cameras, microphones, etc.It produces information by taking some input, processing it insome way, and outputting it.It does not "understand" anything in the same way a humandoes, but rather algorithmically processes data it is given." |

Table 1. **Labels for types of elicited explanatory models of AI systems.** N is number of responses, and Example contains a quote associated with the label. *Abstract* theories communicate what AI does vs. *Operational* theories which emphasize how AI works. *Sparse* and *Sophisticated* refer to low and high levels of technical depth and accuracy in elicited explanations reflectively.



that our broader findings are informed by all three sources as they represent different means of inquiry and perspectives on the experiment.

## 5.1 Prior Assumptions, Explanatory Models

*5.1.1 Detail in explanation vs. technical depth and accuracy.* We assessed the structure of participants' prior explanatory models of AI through one open-ended question, i.e., "How do you think AI works? (For example, where does it get information? How does it produce information? How does it understand what you ask it?)", expecting a range in the responses. We observed during the first coding cycle that the results seemed to actually vary in more than one way (rather than being more or less structured overall), and so we model this as a two-dimensional construct. During the second (deductive) coding cycle, we adjusted code labels to relate them to prior and parallel work:

(1) **Type of Explanatory Model**
  (a) *Abstract*: What AI does
  (b) *Operational*: How AI works
(2) **Technical Depth**
  (a) *Sparse*: Vague or inaccurate description
  (b) *Sophisticated*: Low-level and accurate description

We based these labels on prior and parallel work. Specifically, DeVito et al. describe *abstract* and *operational* algorithmic folk theories, noting that the former "do not include specific attempts to theorize how an algorithm might actually operate" [30] (their sub-codes for these are not applicable to our case). Interestingly, in a study of

mental models of adversarial machine learning, Bieringer et al. found that their participants' prior knowledge did not necessarily determine the technical depth of elicited mental models, pointing to a possibly multidimensional space. They apply the labels *sparse* and *sophisticated* to describe the technical depth in these mental models [12].

The majority of participants in our study gave *Sparse-Abstract* models ($N = 14$). For example, P3 wrote "I think it gets info from devices and uses language features to understand us." See Table 1 for the full distribution over these label combinations, and additional examples.

The second most common explanatory type was *Sophisticated-Operational* ($N = 5$). For instance, P1 alluded to both generalization and optimization in their explanation: "I think that it works by feeding it data. It is then able to use the data that it is fed and apply the given outcomes for the provided data to novel situations. It is accurate as it continues to learn and reduce loss between the real and given answer."

*Sparse-Operational* and *Sophisticated-Abstract* explanations each occurred twice. We identified P16's description as an instance of the former:

> " Usually it gets information from big datasets of training examples, similar to ones it needs to act on. There are different ways for it to produce information - it may do some clustering algorithm, use neural network, genetic algorithms or simply build decision tree based on previous answers. Most AI do limited number of tasks, so they recognise one of few commands. The ones which recognise speech likely try to represent sentence grammatical structure and use previous users' dictionaries and set of predetermined algorithms. But i really do not know "

P16's explanation is long and contains examples of machine learning methods and speech recognition systems, the description of the latter however is ambiguous and likely inaccurate (by comparison to most existing speech recognition approaches); moreover, the participant explicitly indicates that they don't know how it works despite offering an account.

By contrast, a *Sophisticated-Abstract* explanation provides accurate description coupled with description of what AI does but not how AI works (despite the questions specifically asking "How"). For example, P15 wrote:

> " AI is trained on a large amount of data. The training will usually tell the machine what it needs to know and it will then produce information based on its training. It will identify similar features from the training dataset and the test dataset to make an analysis. "

This participant, like P1, refers to generalization (similarity between train and test features), which requires knowledge about how machine learning models can be useful for real-world tasks, but provides no theory about how this is accomplished despite the posed question explicitly asking for it.

*5.1.2 Human creativity in writing.* Our two questions relating to human creativity in writing allow us to elicit unconstrained thoughts through open responses as well as anchor to classical constructs such as Novelty (historically new), Surprise (unexpected), and Value (useful to people) [13] via a multiple-choice item. We additionally included an *Other* field in the multiple-choice item, to allow participants to specify a different dimension if they felt that their concept was not adequately represented by these three, especially given the domain constraint. A majority of participants ($N = 12$) indicated that Novelty was most important, followed by Value ($N = 6$), Surprise ($N = 4$), and one custom response: "evocative use of language", a domain specific attribute.

In the open responses, participants identified several features of human creative writing they considered important, which we coded as follows:

*Freedom/Expression.* Five participants commented on personal expression and expressive freedom (P1, P3, P10, P20, P22). For example, P3 simply stated "freedom of expression", while P20 remarked on both aspects: "Creativity in writing for me is putting down a **personal**, immersive response to a prompt. So taking a spark of direction and going **wherever I want** from there."

| | N | Example |
|---|---|---|
| Freedom/Expression | 5 | "Creative writing, to me, is writing that embodies one's own novel artistic expression and is not primarily functional." |
| Imagination/Fiction/Inspiration | 5 | "It is an enjoyable activity that involves imagination and allows one to express his/her feelings." |
| Structure/Clarity/Goal-directedness | 4 | "Having words flow out and describe things in a satisfying way" |
| Novelty | 4 | "For me its coming up with new, innovative and engaging ways to write a story." |
| Unexpectedness | 4 | "Creativity in writing is using tropes and ideas in uncommon ways." |
| Truth/Reality | 1 | "Being able to capture truths about the real world through words" |

Table 2. **Codes re: human creativity in writing.** N indicates number of participants, Example shows a corresponding quote.

*Imagination/Fiction/Inspiration.* Five participants commented on imagination and fictive writing (P5, P6, P8, P9, P11), such as P5 who wrote "Letting my imagination go free, creating worlds and scenarios that don't exist." P6 illustrated this by comparison: "Creativity writing is the type of writing used in **stories, novels, poems, journals**. I keep it **separate from scientific writing**, which I don't consider as creative writing."

*Additional thoughts: Structure/Clarity/Goal-directedness, Novelty, Unexpectedness, and Truth.* Still others commented on form, structure, flow, and direction. P2 indicated that creative writing involves "having a clear goal and many possible ways to accomplish that goal." The familiar dimensions of Novelty and Unexpectedness (surprise) also appeared in several comments. Finally, one participant alluded to "truth", perhaps indicating not the idea of verisimilitude (or similarity to reality) but "truth in fiction."

These features of human creative writing participants considered important are also summarized in Table 2 with additional examples.

*5.1.3 AI creativity.* When asked whether they thought AI could be creative, the majority of participants ($N = 17$) indicated Yes. We assessed this through analysis of an open-ended item, and some participants did indicate uncertainty by using words like "probably." We obtained the following codes through our analysis:

*Human-based.* Five participants (P3, P6, P8, P17, P19) noted that ostensibly creative AI is somehow modeling human creativity, and used this point to indicate that AI can be creative. For example, P3 wrote "probably, because it can mimic other human features."

*Combinatorial Creativity and Uniqueness/Randomness.* Others pointed to the notion of combinatorial creativity [13] (P4, P9, P16, P21), suggesting that "It can be creative if it happens to combine things in a way that people wouldn't naturally consider" (P9). Relatedly, participants noted the opportunity for creativity arising from randomness. P15 notes that AI-generated ideas "can be completely illogical which is sometimes the best creativity."

*Novelty/Surprise.* Three participants (P7, P10, P12) implicitly made the connection to novelty and surprise, remarking that AI "can be creative in the sense that it can produce novel solutions to problems," but also that "this presupposes a narrow conception of creativity" (P12).

*Additional thoughts: Future creativity, and uncertainty.* Still others pointed toward future creative ability, due to the improvement of AI, such as P5: "Eventually, yes, but I'm not sure it can make big leaps in novelty in a single go." P18 expressed uncertainty about the question of whether AI can be creative, writing that they were "unsure what makes humans creative." Other participants expressed different kinds of uncertainty; P1 wrote that they believe that AI 'can be accurate" but they would have a hard time "imagining what a creative AI would look like."

| | N | Example | Yes | Unsure | No |
|---|---|---|---|---|---|
| Human-based | 5 | "Perhaps. The AI itself is a work of human creation. It might help a person to be creative in the same way that automatic writing, randomness, or writing constraints do." | 4 | 0 | 1 |
| Combinatorial | 4 | "I think that AI can create from a broad set of information that it has been given, but I do not think it can make up something new" | 3 | 1 | 0 |
| Uniqueness/Randomness | 4 | "Yes, it may suggest ideas or connections that a human might not usually make" | 4 | 0 | 0 |
| Other | 3 | "I do believe that AI can not be creative based off of my understanding. I believe that it can be accurate but I have a hard time imagining what a creative AI would look like." | 0 | 1 | 2 |
| Novelty/Surprise | 3 | "Yes, in the sense that AI can generate novel and surprising ideas without input from a human." | 3 | 0 | 0 |
| Gradually/Future | 3 | "Yes definitely! But to a certain extent. Because technology keeps on evolving and the internet is a very good example of it where it really helps in building creativity. But nonetheless, i think there is a limit to its creativity as compared to human but of course it will help a lot in enhancing creativity" | 3 | 0 | 0 |
| Unsure what makes humans creative | 1 | "I'm not sure because I'm not sure what processes allow humans to be creative" | 0 | 1 | 0 |

Table 3. **Codes from open responses about AI creativity elicited from participants.** N is number of responses, Example contains a quote associated with the label, and Yes/No/Unsure are counts of categorical responses, for each label, from participants about whether they think AI can be creative. Some responses are labeled with more than one.

P22 emphasised the fact that AI depends on "the input humans give to it" and noted that "if humans don't keep updating the inputs, it may not be creative anymore."

Table 3 provides other examples for codes mentioned above.

*5.1.4 Human-AI Differences.* We also elicited participants' thoughts about qualitative differences between human and AI text production. We assessed this through two items: one multiple choice to indicate the presence of a difference ("Do you think the way AI produces text is different from humans?"- Yes/No/Unsure), followed by an open-ended question prompting them to explain how and why (or why not) human and AI text production mechanisms are different.

For the multiple choice item, 16 participants answered Yes (they think the way AI produces text is different from how humans do it), 5 answered "Unsure", and 2 participants answered "No." In surveying qualitative examples, we found diverse concepts of what differs between how humans and AI produces text, as well as effects of such differences.

*Statistical/Data vs. Symbolic/Mind.* Eight participants (P1, P2, P7, P8, P, P15, P17, P18) pointed to the contents and mechanics of the text producers. They differentiated between data-driven and mind-driven text production, for example P1 wrote: "AI is taught to produce text based off of given **data**, an algorithm is used to produce text while a human creates text based off of their **mind**." Participants phrased this by contrasting generating text "statistically and linearly" with "using mental hierarchy of words" (P1), or indicating rule-based vs data-driven constraints. For example, P8 noted that "people can produce infinite correct and comprehensible outputs based on their knowledge of their native language's grammar and vocabulary, while AI will only be able to produce content based on its input."

We label this by combining the cues from participant responses and the traditional AI notions of symbolic processing vs. statistical modeling, two dominant paradigms in natural language processing.

*World model and understanding.* Another thought from Five participants (P4, P5, P9, P11, P12) was that AI is lacking sufficient understanding of context, which comes from experiences in the world. P5 points out that AI "has not learned language by interacting with society and cultures, learning from family and personal experiences, or have the ability to draw on memory when responding in the same way humans do" while P4 appeals to sensorimotor functions:

> " AI produces numbers based on drawing patterns and similarities from the numbers in the dataset that it has been fed. It **doesn't understand and have visual representations in the brain that it then produces into motor action**, it's just reproducing what it's already seen. "

*Not sure how humans do it.* Four participants (P9, P12, P16, P18) remarked that either they were unsure, or not much is known, about how humans produce text. They differed in whether they thought this would extend to similarity or difference between humans and AI. For instance, P9 answered: "I'm not really sure how people "produce text"" and continued saying that since "AI learns from previous patterns" then maybe people in a similar way "learn from past experiences and instructions." P16 noted: "Not that is known how humans produce texts, so mechanisms developed independently are unlikely to be the same."

*Complexity, performance.* Three participants (P6, P7, P11) commented on expected difference in the complexity of the produced text, or performance factors that might affect quality. P6 emphasized that "it depends on the level of development of the AI" explaining that "in the ideal case one should not be able to recognize a human produced text from a machine produced one." P11 used an example of google translate and it being unable to translate complex terms or understand the context, to point out that the difference between human and AI producing text might have to do with the complexity of the language .

*Additional thoughts.* P14 argued that AI lacks "intentionality" due to not having "desires or beliefs", while P19 relatedly noted that AI cannot be "spontaneous" or "irrational" in its behavior as compared with humans. Two participants also made comments about formality in language. For example, P21 noted that "AI can only use what it has been taught or can access via some database while humans may access more informal or colloquial writing patterns."

Two comments noted that AI language systems are based on human-provided data. For example, P6 didn't expect a difference "because the information is mainly fed by humans." Finally, three participants made seemingly contradictory or unclear statements. For example, P22 indicated no difference, but then expressed an opinion on the difference of a somewhat ontological difference:

> " I think in some ways, each AI and humans communicate with our own languages and it's a mean of mutual understanding between them, so it's not that different. **It's just, humans don't operate the way AI does, and vice versa.** "

Table 4 provides other examples for examples and ratings mentioned above.

## 5.2 Interacting with the system

All the data discussed in this subsection was received through observation of participantsâĂŹ interaction with the system and through verbal comments that participants made during the study. The comments were made either when participants were thinking aloud during a writing task, or as a reply to the interviewer's questions. Throughout the session interviewers asked general questions, like, âĂIJWhat are you thinking?" and "How is the writing going?âĂİ: either once every 2 minutes or, if the participant seemed to be disturbed by being interrupted often, when the participant stopped writing. interviewers also asked more specific questions, like, âĂIJWhat do you think about that suggestion?", "Why are you laughing?", and "Tell me how you incorporated that suggestion.âĂİ

| | N | Example | Yes | Unsure | No |
|---|---|---|---|---|---|
| Statistical/Data vs. Symbolic/Mind | 8 | "from my understanding, computers generate text statistically and linearly (vs. humans using mental hierarchy of words)" | 7 | 1 | 0 |
| World model and understanding | 5 | "It has not learned language by interacting with society and cultures, learning from family and personal experiences, or have the ability to draw on memory when responding in the same way humans do." | 5 | 0 | 0 |
| Not sure how humans do it | 4 | "Not that is known how humans produce texts, so mechanisms developed independently are unlikely to be the same. Also humans seem to be much better at text generating but may be it is because they have more and more diverse experience" | 3 | 1 | 0 |
| Other | 3 | "In a sense, yes, because itâĂŸs more or less about encoding connections in memory, then following the connections through to retrieve this memory, or making predictions, based on what you already know." | 1 | 1 | 1 |
| Complexity, Performance | 3 | "I guess it depends on the level of development of the AI. In the ideal case one should not be able to recognize a human produced text from a machine produced one." | 2 | 1 | 0 |
| No intentionality | 2 | "It seems not as structured as humans? And it doesn't seem to have "intentionality" (e.g. they don't appear to have desires or beliefs when they try to make an argument)" | 2 | 0 | 0 |
| Formal vs. informal language/behavior | 2 | "AI can only use what it has been taught or can access via some database while humans may access more informal or colloquial writing patterns" | 2 | 0 | 0 |
| Human-based | 2 | "Because the information is mainly fed by humans" | 0 | 1 | 1 |

Table 4. **Codes re: expected differences between human and AI text production, *before* writing.** N indicates number of participants, Example shows a corresponding quote, and Yes/No/Unsure are counts of categorical responses, for each label, from participants about whether they think AI text production is generally different than that of humans. Some responses are labeled with more than one.



We noted broadly different styles of overall participant writing and engagement with suggestions, described in detail below. We provide examples of participants who most clearly embodied the associated characteristics of each.

(1) *Reactive writing.* Through observation, interviewers identified four participants for whom suggestions were actively shaping their story and helping them decide where the story was going (P17, P7, P10, P23). They wrote in a way that looked like a reaction to either suggestions of the system or as a reaction to the task. There were clearly effects from the pressure of time, conditions of the task, and their habits of writing. Some participants also mentioned that what they wrote was more like a âĂIJstream of consciousnessâĂİ (P23).

(2) *Proactive writing (with suggestions).* Participants with clearly proactive writing (P2, P11, P15, P18, P20) wrote having a clear idea of what they wanted to write (having some horizon of their story) and how they wanted to do it (having their own process). They incorporated suggestions of **Editor-Red** at some particular points of their stories, either when there was the end of the scene or after they had exhausted the story horizon they had in mind. They did not let **Editor-Red** take over their process of writing. This type of writing was characterized by longer writing periods and hitting suggestions fewer times. For example, P18 requested suggestions only two times, and had a 15 minute writing process non-stop. P19 requested suggestions three times, and had longer writing periods, with one period being 9 minutes.

(3) *Actively opposed to suggestions.* Four participants (P6, P8, P16, P22) were generally *not willing to incorporate* the suggestions of **Editor-Red**. For example, P8 had their own idea of how they wanted to do things and explained the resistance to incorporate the systemâĂŹs suggestions, saying âĂIJIâĂŹm not a super suggestible person.âĂİ P16 and P22 did not like the suggestions and did not include them in the writing. P6

wrote so as to improve the suggestions by the system, waiting for this to occur, and so didn't engage with the suggestions in the duration of writing their story with it.

We begin by exploring overall reasons, given certain types of suggestions and contexts, to incorporate or not to incorporate suggestions in the view of our participants. Subsequently, we describe and characterize instances of suggestion integration, often from more reactive and proactive writers (who did accept suggestions), in detail through the lens of *integrative leaps*.

*5.2.1 Reasons to incorporate suggestions.* Making a judgment as to whether the systemâĂŹs suggestions are in line with the participantâĂŹs writing or too âĂIJout thereâĂİ seemed to be an important axis along which participants, in the process of writing, were constantly making decisions about suggestion incorporation. Five participants specifically commented on the systemâĂŹs suggestions being in line with their writing (P1, P3, P4, P20, P21, P5). For example, P1 explained their decision to incorporate a suggestion because it was âĂIJthematically accurate and kind of good-to-keep-the-story-going description." P20 set a scene in their story and then hit the suggestions, as they wanted to see how **Editor-Red** "would interpret that." One of the suggestions of **Editor-Red** was âĂIJ...IâĂŹm calling to let you know that youâĂŹve been selected to the next round of the lottery,âĂİ and P20 exclaimed: "Wow... it's a bit scary because I had thought of the lottery idea or just some other kind of news... yeah, so it's interesting that it immediately followed that train of thought about a lottery.âĂİ

At the same time, some participants appreciated that the system was providing suggestions that were unexpected and not immediately related to their previous writing. For example, one participant explained that some suggestions, even though they seemed âĂIJabsurd,âĂİ were also so âĂIJdetailedâĂİ and âĂIJspecificâĂİ that it was âĂIJinspiring.âĂİ As a result, even though some suggestions were âĂIJa little bit out thereâĂİ to the extent that they would make them laugh, the suggestions would still give them âĂIJsomething to goâĂİ (P1).

We observed a subtle and variable trade-off between how creative or unusual suggestions were thought of as being by participants and how easy it was to incorporate them. The possibility of an easy transition towards incorporating the suggestions seemed to be a crucial factor. For example, P1 commented on one of the suggestions:

> " Some of the suggestions, would be either so similar to what I wrote that it doesn't seem worth incorporating or too creative, or I guess too hard to transition to. But that [suggestion] would logically be the next thing that I write about. "

Along these lines, some suggestions that might have been incorporated under the right circumstances by the corresponding writers were not integrated because of the time and effort it would require. P9, when choosing one of the two suggestion types (*Plot* and *Description*) that they equally liked, acknowledged that though they liked the suggestion that was âĂIJmore fun, weird, crazy,âĂİ it was such âĂIJa divergent shiftâĂİ that the suggestion looked âĂIJtoo effortful to incorporate.âĂİ Another participant explained that if they were to take âĂIJa much longer routeâĂİ they might follow the systemâĂŹs suggestion of monster hunting (*Description* suggestion: âĂIJYou are a member of a group of monster hunters.âĂİ) as âĂIJit seems funâĂİ but since they were short on time they decided not to explore this plot line (P5).

*5.2.2 Reasons to not incorporate suggestions.* Participants gave a wide variety of reasons as to why they might have been unwilling to incorporate suggestions. Some participants commented that the textual suggestions of **Editor-Red** looked âĂIJbasicâĂİ (P15), âĂIJplainâĂİ (P6), âĂIJredundantâĂİ (P4), or were not âĂIJpicking up the toneâĂİ of the story they were writing (P6). At some points in their writing, six participants commented that suggestions were not in line with what they wrote (P2, P6, P7, P8, P12, P22), complaining that the system was not able to see that âĂIJthis is not where IâĂŹm goingâĂİ (P7). Some of the suggestions also did not make sense to participants and were repetitive (P2, P6, P17), whereas for some participants the fact that some of the suggestions were not coherent was not an obstacle to engaging with their content. For example, when P7 was interacting with the system, it experienced a number of delays in producing suggestions, and finally a plot suggestion came out as: "not sure where to end Train to MIT for the first time not sure where to end Train to MIT for the first

time not sure where to end Train to MIT for the first time not." The participant said laughing: "ThatâĂŹs fine. I donâĂŹt necessarily need it to be coherent" and expressed the readiness to carry on the writing.

One of the four participants who did not incorporate the systemâĂŹs suggestions found that the suggestions contained tropes and led to more âĂIJstereotype writingâĂİ (P8). Four participants also specifically pointed out that the systemâĂŹs suggestion distracted them from pursuing their own ideas (P2, P8, P12, P14). P8 explained that they felt they had âĂIJto tune outâĂİ the systemâĂŹs suggestions as they already had a picture of what they wanted to write in their head and it was easier for them to write âĂIJwithout the extra stuff.âĂİ Some suggestions also did not come at the right time in the narrative: âĂIJOh, wow, I would not think of cryogenic sleep! That's an interesting idea, but I didn't use it. I don't think it came at a good time in the story.âĂİ (P1).

Visual suggestions were not incorporated if they were perceived as unrelated to the current writing (P14, P15, P6, P4). Participants also commented that some of the images were not only unrelated to the writing but also seemed arbitrarily constrained or homogeneous (e.g., demographically): âĂIJIt's kind of strange, there's just a bunch of white guys staring at me and I don't know whyâĂİ (P2) and âĂIJIâĂŹm confused ...And IâĂŹm curious as to why all the suggestions are very similar, and they are all images of straight blonde Caucasian womenâĂİ (P5).

P14 considered the images âĂIJaestheticâĂİ and âĂIJcuteâĂİ but was following the idea they had in mind already. Some of the image suggestions, similarly to the text, did not come at the right time: âĂIJThe pictures are coolâĂębut this doesn't really fit with the character I have right nowâĂİ (P15). P3 commented that even though they were not using visual suggestions, having them seemed âĂIJless daunting than having a white space in front of you.âĂİ

Sound suggestions were the least used, sometimes due to a lack of relevance to the writing, either in content or in tone and style, and sometimes for other reasons. P5 described the sounds being not relevant and âĂIJrandomâĂİ (P5), and P9 explained why they were going to switch off the sounds:

> "The sounds are a bit dystopian. I feel if I use the sounds as an inspiration, I'd end up thinking of some sort of totalitarian government that's using lots of walkie talkies all the time and tracking people. It sounds very different to the vibe I was thinking in my mind. "

The sounds were also described as âĂIJaggressiveâĂİ making it hard to focus (P8), âĂIJtoo muchâĂİ (P9), and âĂIJdistractingâĂİ (P12). All participants, at least once, switched off the sounds at some point of their writing, although we observed instances where sounds were related to the story and/or resulted in incorporated suggestions.

*5.2.3* ***Editor-Red*** *as "support system".* Some participants described the overall experience of the system, which is not limited to just the relevance of the suggestions. For instance, some reported that they felt **Editor-Red** supported the writing process. P3 commented that the system was giving âĂIJgood linesâĂİ and admitted that it helped âĂIJto continue along, where otherwise I think I will just stop writing.âĂİ P3 continued saying that when they had absolutely no idea what to write, taking a word or a line from **Editor-Red** gave them âĂIJsomething to addâĂİ and then they "kept going, and kinda went from there.âĂİ Two participants explained that **Editor-Red** helped them to feel less stuck in their writing (P12, P16). Even though P16 did not incorporate any of the suggestions, and during the experiment commented on the suggestions being âĂIJdumb,âĂİ they expressed their surprise that, in the end, writing in **Editor-Red** did seem to help them feel âĂIJless stuck.âĂİ P12 explained their feelings about one of the suggestions: âĂIJAlthough I wouldn't word-for-word take that, it, at least, redirects my attention from just being stuck in the kind of the crucial little loop to having somewhere else to go. So that's helpful to get unstuck, I suppose.âĂİ

P23, reflecting on their experience of writing after the end of the session, admitted that they felt that inspiration came not directly from suggestions but rather suggestions made them think of something else and this is where ideas came from. P6 admitted that even though they did not use the suggestions of **Editor-Red**, writing in it

actually âĂIJrelieved some of the stress of writing.âĂİ P6 further explained that even though the suggestions were not helpful to them personally, the system was still âĂIJcreating that distraction, that was good for making the task a little bit more relaxing.âĂİ They noted that interacting with **Editor-Red** really helped in mitigating the stress of writing, comparing it to a feeling of âĂIJpetting a cat.âĂİ

P1 described how interacting with the system changed their process of writing as they would âĂIJwrite for the suggestions.âĂİ P1 explained that when they didn't want to continue writing as they could not think of what to say, being in the system would be a motivation to âĂIJwrite a few additional sentences in order to get a better suggestionâĂİ and to continue writing if they were unhappy with the suggestions that I received till the get a better suggestion. P1 found it "very helpful."

All in all, observing participants' interaction with the system enables us to map out the multiplicity of tactics that participants engage in while performing the task of writing a text. Participants borrow words and lines from the system's suggestions, get inspired by the system's suggestions directly or indirectly, use ideas from the suggestions as reference points to produce their own ideas, or receive psychological support from the system (e.g. reducing the stress of writing by, for example, giving them something to focus on besides their own feelings of getting stuck).

*5.2.4 Willingness to cooperate with **Editor-Red**.* Participants practiced different patterns of engagement with the system hoping that it would provide suggestions that better served their purposes. Some were willing to wait longer for the system to start giving better suggestions, occasionally under the assumption that allowing more time would give the system an opportunity to catch up with the participantâĂŹs writing. For instance, P1 said: âĂIJThis makes me think I didn't wait long enough because now everything's about phones. So maybe I should wait a little longer.âĂİ

Some also decided to keep writing in order to give more information to the system (P1, P19, P8). P1 reasoned that the information that they were "feeding it" might be "not substantial nowâĂİ. When P7 received another round of suggestions, one sentence that particularly got her attention was the phrase âĂIJI could feel the horn blaring in the distance.âĂİ This phrase was almost identical to what the participant had already written, with the system having changed precisely one thing: the original sentence was âĂIJI *hear* the horn blaring in the distance." P7 pondered:

> " I see it changed some of the words around. *I could feel the horn blaring*, which is interesting. ItâĂŹs much more visceral than *I can hear it blaring*âĂę Yeah, *hear* is not quite the right wordâĂę I will just put *feel* for now. "

This is an example of how a participant is willing to cooperate with the system and make sense of its contributions, even when someone else might consider the re-phrasing to be overly subtle (merely a lexical substitution) and not valuable to developing the story further.

*5.2.5 Integrating the system's suggestions.* To provide a more granular exposition of the suggestion integration patterns we observed, we enumerate and detail a collection of *integrative leaps*. These leaps describe the different kinds of interpretation and expression involved in incorporating aspects of suggestions into the developing story, in particular how and how much participants alter the meaning and structure of their narratives when doing so. We use them as fine-grained windows into the mechanics of the most visible examples of this incorporative process, those we are able to access through our observational methodology.

Our data on suggestion integration contains 47 instances of integrative leaps from 19 (out of 23) participants; P6, P8, P16, and P22 did not appear to incorporate **Editor-Red**âĂŹs suggestions in any identifiable way). These are examples that the researchers conducting study sessions identified of participants engaging with and actively incorporating suggestions from the system. Participants often explicitly commented and explained why and how they incorporated suggestions, as they were encouraged to do, and we report on their interpretation of this process in addition to our observations and analysis.

*5.2.6 Types of integrative leaps.* The integrative leaps can be analyzed along a number of axes. First, we consider the "edit" distance (e.g. lexical, semantic, etc.) between the suggestion as presented to the user and as incorporated into the story. We broadly characterize these as *direct integration* ($N = 30$; e.g., verbatim or restructured verbatim for a textual suggestion or a textual analogue of the object or idea represented in a visual or auditory suggestion) or *indirect integration* ($N = 17$), where it often would be impossible to capture this integration if we did not have the participantsâĂŹ explanations, due to the modifications they made in the process of suggestion incorporation.

Second, we look at how incorporated suggestions relate to global aspects of their story's direction and most prominent elements. When participants used suggestions to explore new lines of narration, we call it *exploratory integration* ($N = 28$, shown on left half of both figures), in contrast to taking suggestions to continue with their chosen narrative by adding more details, which we call *confirmatory integration* ($N = 19$, shown on right half of both figures).

Finally, with the view that suggestions are intended to ease cognitive inertia in the writing process, we attend to the role they play in creative problem solving. Do they simply solve a localized problem by "closing" some aspect of the narrative in a necessary, analytical, or expected way? For example, naming a character that has already been described, or explaining why a character went from place A to place B if both of those events have been established. Or do they "open" up options to consider, resulting in abstract, novel, or unexpected events, patterns, or directions? We describe these as *convergent integration* ($N = 31$, shown on the bottom half of both figures) and *divergent integration, shown on the top half of both figures* ($N = 16$) respectively. While these often overlap with confirmatory and exploratory integrations respectively, there were a few cases in our coding process where we found it useful to explicitly make a distinction between these two dimensions, in order to better explain behaviors that we observed. For example, two of the six integrations we detail in the following section are ones we labeled, through an iterative process, as *exploratory* and *convergent*. In these leaps, participants may use suggestions to both pivot at a narrative level, and solve a local problem within this context. Although our categories are still relatively broad and cannot cover all the differences between integrations that we observed, we sought to sufficiently represent the most prominent aspects of integrations with these labels.

*5.2.7 Integrative leaps.* In this section we review several examples of *integrative leaps*, identifying them along the aforementioned axes as well as describing the participants' interpretation and comments. We summarize each instance in a discrete box that clearly identifies the input text (before the suggestion), the suggestion at hand, the text after integration, the participants' explanation, and our labels (for example, Integration 1). When participants identified that they were prompted by visual or auditory suggestions, we include thumbnails or links for the reader to review.

P3, following the "The phone began to ringâĂİ prompt, was writing an intense story of a mother getting a phone call from her estranged son. Through a number of previous suggestion interactions, the participant wrote a story where the son on the phone call was in trouble, as some people were holding a gun to his head and demanding some information he didn't have. The next round of suggestions contained "IâĂŹm just a normal person who is in a hurry to get home." Following that, the participant wrote "She freezes. What is he talking about? This isnâĂŹt making any senseâĂę yes, she has an estranged relationship with her son, but they are normal people." As the participant explained, the phrase in the suggestion "I'm just a normal person" stood out to them and prompted them to develop it into the mother's inner thoughts trying to come to terms with the fact that her son and she herself are probably in big trouble. We labeled this example as *direct* (almost verbatim integration: **normal person** to **normal people**), *exploratory* (the participant did not have a clear idea of the narrative) and *convergent* (solving a local question of how the main character reacts to the news that her son is in trouble). See integration 1 for more details.

P21 was developing a story from the prompt "The phone began to ring" and was describing a call from the best friend of the main character. P21 wrote the first part of the dialogue "âĂIJWait why were you in the hospital?âĂİ

*Integration 1*

**Input (summary)**: [Emotional dialogue, son is held captiveâĂę] âĂęâĂIJWhat?âĂİ She replied back. âĂIJWho are you talking about?âĂİ âĂIJItàĂŹs them,âĂİ he whimpered. âĂIJBut I-I donâĂŹt have anything to tell them. I donâĂŹt have the information theyâĂŹre looking for.âĂİ

**Suggestion**: *Plot.* IâĂŹm just a **normal person** who is in a hurry to get homeâĂę

**Integration**: She freezes. What is he talking about? This isnâĂŹt making any senseâĂę yes, she has an estranged relationship with her son, but **they are normal people.** âĂIJYouâĂŹre not making any sense.âĂİ âĂIJItâĂŹs **not normal. None of this is normal**âĂİ he responds shakily. She hears a scream and the phone cuts out.

**Explanation**:

> "âĂęI'm just thinking about how to continue this story but I don't really have muchâĂę but the suggestion under *Plot* is giving me someâĂę you know, "**I'm just a normal person**" lineâĂęI still don't have any sort of direction with the storyâĂę this feature seems to be good to help me, like, continue along, where otherwise I think I will just stop writingâĂę "

> " âĂęit just kinda stood out to me in relationship to this story... cause this story, it seems like again the mom is just a **normal person**, so if she is getting this phone call from her son, it doesn't make any sense, we are just **normal people**, so I thought I would incorporate that" "

**Our labels**:

- ***Direct***: the incorporation is almost verbatim (**I'm just a normal person** to **they are normal people**).
- ***Exploratory***: the writer, from their own remarks, does not have a clear narrative direction that this suggestion would reinforce. Rather, it gives them a possible next step to build on.
- ***Convergent***: the suggestion helps to solve a local problem in a concrete way (continuing the story further).

I asked my friendâĂİ and the subsequent round of the suggestions contained images with cars. The participant immediately took on the idea: âĂIJI'm seeing cars, so maybe he was in a car crash.âĂİ and to continue the dialogue, P21 wrote: âĂIJMy sister was in a car crash. SheâĂŹs okay, but she broke a rib.âĂİ Since the suggestions helped to keep the writing going and did not prompt the participant into a new avenue of thought, as well as being a textual representation of a suggested visual object, this entry is labeled as *direct* (**images of cars** to **car crash**), *confirmatory* (reinforces the existing narrative), and *convergent* (closes a local question of why the person is in the hospital). We report details in integration 2.

*Integration 2*

**Input (summary)**: [Best friend phone callâĂę] âĂęâĂIJI ran into your ex-boyfriend at the hospitalâĂİ. I was in shock. I hadnâĂŹt seen him since 4 years ago when he left me to run away to Cuba with some new woman.

**Suggestion**: (Images) {  ,  , ... }

**Integration**: âĂIJWait why were you in the hospital?âĂİ I asked my friend. âĂIJMy sister was in **car** crash. SheâĂŹs okay, but she broke a rib.âĂİ I completely forgot about what she said about my ex being in the area, assuming it was hours ago, and rushed to the hospital. We were neighbors growing up, so I was pretty close with her sister too.

**Explanation**: âĂIJI'm seeing **cars**, so maybe she was in a **car** crash.âĂİ

**Our labels**:

- ***Direct***: direct representation of visually represented object.
- ***Confirmatory***: reinforces the existing narrative.
- ***Convergent***: closes a local question: i.e. "why?", "what happened?" regarding a character in the story.

P4, following the prompt "The phone began to ring," was developing a story about a police detective who called the main character and asked to come to the police station because their sister was in trouble. P4 felt unsure as to how to continue and what it could be that the detective could have been accusing their sister of. This participant was really perplexed with what in their previous writing could have prompted the subsequent

suggestions involving zoos, animals, and tropical places (these were in the retrieved images) but still decided to go ahead and integrate the suggestions into their story. In the end, P4 wrote (about the police detective): âĂIJHe appeared a bit nervous. He told me that he suspects my sister may have stolen an elephant from the zoo when she was studying abroad in India. I felt shocked.âĂİ P4 explained their reasoning in integrating the system's suggestion: âĂIJI donâĂŹt know why these images popped up and how they are related to what I wrote before. But I saw the elephants and some kind of more tropical places and so it kind of made me think of ...I don't know I was thinking what could it she possibly have done wrong that she could be in trouble and so the elephant was standing out to me, so I chose to say that âĂIJshe stole an elephantâĂİ and I was thinking where elephants are and I know that there are a lot of elephants, like Indian elephants, so thatâĂŹs why I said that.âĂİ See integration 3 for more details.

The participant concluded their story by writing, in an attempt to rationalize and make sense of the participation of the elephant in their story:

> " I knew my sister loved animals, especially larger ones, but I never would have expected this. Where would she have left it? I had so many questions. I asked if I could talk to my sister. "Did you steal an elephant??" "I don't know what he's talking about. I've never seen it before." "

*Integration 3*

**Input (summary)**:[Going to meet detective] âĂęI took the train to get to the police station. When I arrived, the detective met me at the door. He appeared

**Suggestion**:

*Plot.* to be a little **bit nervous** but seemed to calm down when I asked himâĂę

{  ,  , ... }

(Images)

**Integration**: He appeared a **bit nervous**. He told me that he suspects my sister may have stolen an **elephant** from **the zoo** when she was studying abroad in India. I felt shocked.

**Explanation**:

> " I guess I used the first section of the plot to write âĂIJhe appeared a **bit nervous**"âĂę these images I donâĂŹt know why they popped up and how they are related to what I wrote before. But I saw the elephants and some kind of more tropical places and so it kind of made me think of ...I don't know I was thinking what could it she possibly have done wrong that she could be in trouble and so the elephant was standing out to me, so I chose to say that âĂIJshe stole an elephantâĂİ and I was thinking where elephants are and I know that there are a lot of elephants, like Indian elephants, so thatâĂŹs why I said that "

**Our labels**:

- ***Direct***: **elephant** (from images), **nervous** (from text)
- ***Exploratory***: the elephant, India, studying abroad are substantially new aspects of the plot at this point
- ***Divergent***: does somewhat "close" a local question (what did she do?), however in a very unexpected way that raises many more questions than it answers

Later on in P21's story (previously described in integration 2), they were describing a character driving to the hospital and the system gave auditory suggestions that P21 described as chanting and explained: âĂIJThere is chanting happening, it makes me think she got into traffic because thereâĂŹs a protest happening, ...or a parade.âĂİ So P21 wrote in their text: âĂIJIn my mad dash to get to the hospital, I forgot that the 4th of July parade was happening today just blocks down from the hospital. IâĂŹm stuck at an intersection where the parade is passing by...âĂİ In this example, sound suggestions prompted the participant to think about what could have caused the traffic, so call the integration indirect. The integration of this suggestion also significantly altered the course of the plot (exploratory) creating new avenues of the story development (divergent). More details are in integration 4.

*Integration 4*

**Input (summary)**: [Continuation (see integration 2] âĂęWe were neighbors growing up, so I was pretty close with her sister too.
**Suggestion**: Sound 1 (crowd call and response); Sound 2 (crowd cheering)
**Integration**: In my mad dash to get to the hospital, I forgot that the 4th of July **parade** was happening today just blocks down from the hospital. IâĂŹm stuck at an intersection where the parade is passing by, so I have no choice but to watch the high school band and floats made by various organizations go by.
**Explanation**: âĂIJThere is chanting happening, um It makes me think she got into traffic because there's a protest happening, âĂęor a parade.âĂİ
**Our labels**:

- ***Indirect***: no words or concepts are directly applied; abstract link must be explained by participant (**sound of crowd chanting** to **4th of July parade**)
- ***Exploratory***: altered the course of the plot significantly; narrator eventually turned around and went home after several experiences in the parade
- ***Divergent***: not necessary or expected; creates a twist to develop further; opens up new questions for story

P5 was writing a slow-paced descriptive story using the prompt "A train arrives at the station." At some point, the protagonist was stopped by an officer and told that the train would not be boarding as there was some issues. P5 requested a suggestion and one of the suggestions was âĂIJI had been waiting for this moment for years." The participant continued developing their story and wrote: âĂIJThe train was already late and now this; who knows how long before I get on board?! I canâĂŹt be lateâĂę maybe if I start now, I can drive over toâĂę no, no, no. IâĂŹll never make it that way.âĂİ To the interviewer who ran the session, there was no obvious connection between the suggestion and what the participant subsequently wrote. However, P5 explained that the suggestion "I had been waiting for this moment for years" made them think "more of a frustration for the train being late" and they imagined that there was something that the character was supposed to get to on time in another city. So this idea was translated into making the character impatient.

*Integration 5*

**Input (summary)**: [Train is running lateâĂę] "âĂęThere is a matter we have to attend to first before we will let anyone be checked in,âĂİ said the officer calmly.
**Suggestion**: *Description.* I had been waiting for this moment for years.
**Integration**: The train was already late and now this; who knows how long before I get on board?! I canâĂŹt be lateâĂę maybe if I start now, I can drive over toâĂę no, no, no. IâĂŹll never make it that way.
**Explanation**: âĂIJSo in this case rather than âĂIJwaiting for this moment for years,âĂİ I'm thinking more of, like, a frustration for the train being late and now more delays and there's like something that character was supposed to be trying to get to on time in another city. So it's going to [make] the character impatient.âĂİ
**Our labels**:

- ***Indirect***: **waiting** for **years** to frustration, impatience
- ***Exploratory***: switches from describing scene and events to narrating internal dialogue about the character's feelings
- ***Convergent***: an expected reaction to the situation that describes the effect of the train's lateness

Following the prompt "A train arrives at the station" P9 started writing a fantasy story about frogs waiting for their tadpoles to get back from Tadpole Kindergarten. Another round of suggestions read: âĂIJsound of a bell ringing and the frog who was holding the bell was holding a tray of frogsâĂęâĂİ As the participant explained, the specific "sound of a bell ringing" in the suggestion made them think about sounds in general and what kind of sounds can be in the setting of their story. The participant wrote âĂIJOnce inside the parlour they were all taken back by the ringing and clattering of dishes and trays. âĂİ Here, the participant took a concrete description of sound (sound of a bell ringing) and then made a shift from concrete description to the general concept of sound and made a decision about what kind of particular sound will be in their story (âĂIJclattering of dishes and traysâĂİ).

*Integration 6*

**Input (summary)**: [Tadpoles taking the train back home from KindergartenâĂę] âĂęOnce inside the parlour they were all taken back by the
**Suggestion**: sound of a bell ringing and the frog who was holding the bell was holding a tray of frogs and he was holding a tray of tadpoles who were all waiting for the new tadpoles
**Integration**: Once inside the parlour they were all taken back by the ringing and clattering of dishes and trays. Frogs and toads were excitedly gulping down the various fly filled delights inside. âĂIJGeorgia! Barry! Tadette!âĂİ beamed Mr Willeker. âĂIJYou all look so well!âĂİ Please take a look at the menu.
**Explanation**:

> " I found that interesting as I guess it made me think more of like the sounds that could be inside this parlor or something âĂę because, basically, I was going to end up doing another long description that's probably quite boring. Probably similar to my previous thing I was writing, but I could then think about the sounds like clattering plates. "

**Our labels**:

- ***Indirect***: **sound of a bell + tray** to **ringing and clattering of dishes and trays**; tray is shared, but most of it is indirect
- ***Exploratory***: participant uses it to lead in a different description (see their explanation)
- ***Divergent***: a level of description that opens up commentary about the food and environment, etc.; not necessary or expected

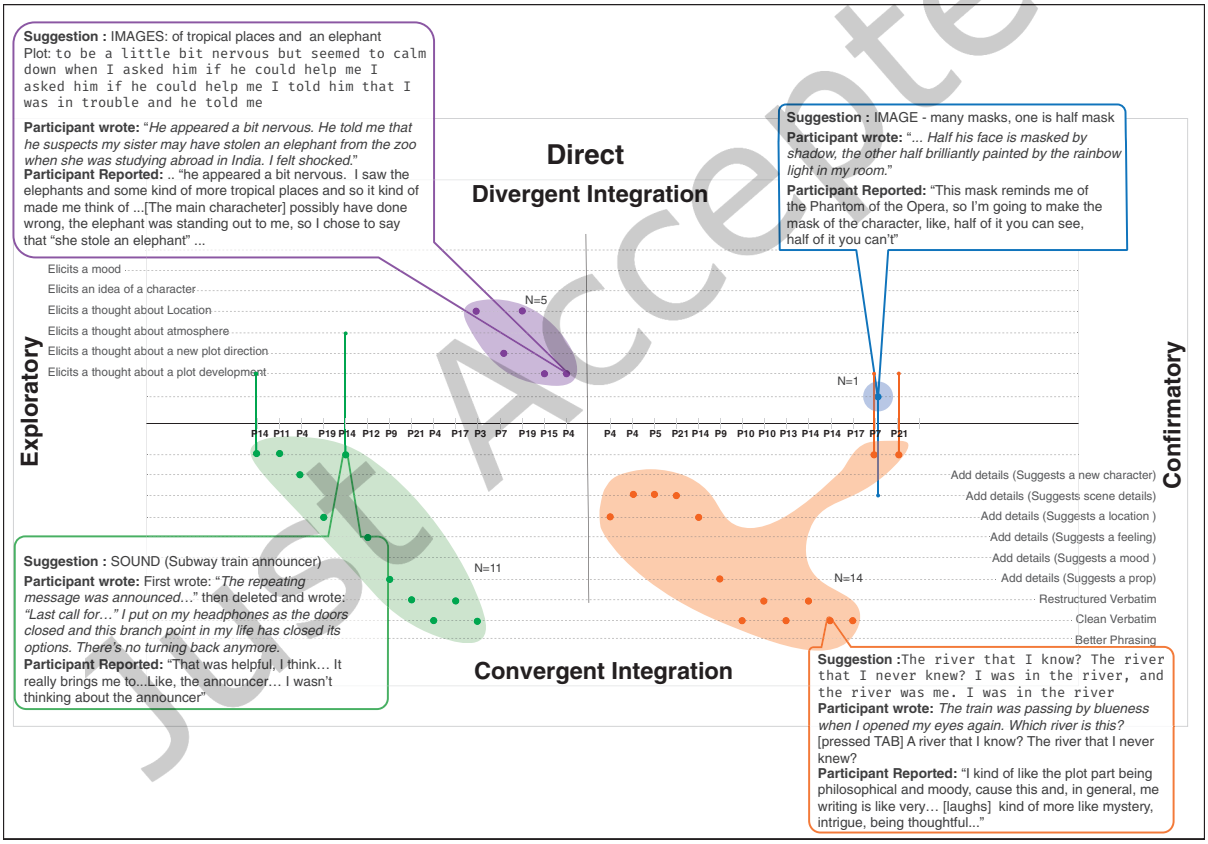


Fig. 5. Diagram of exploratory/confirmatory and divergent/convergent *direct* integrative leaps made by the participants.

We summarize these axes of integrative leaps in two figures. Fig. 5 shows *direct* integrative leaps, and Fig. 6 shows *indirect* integrative leaps. In both figures, left is *exploratory*, right is *confirmatory*, top is *divergent*, and

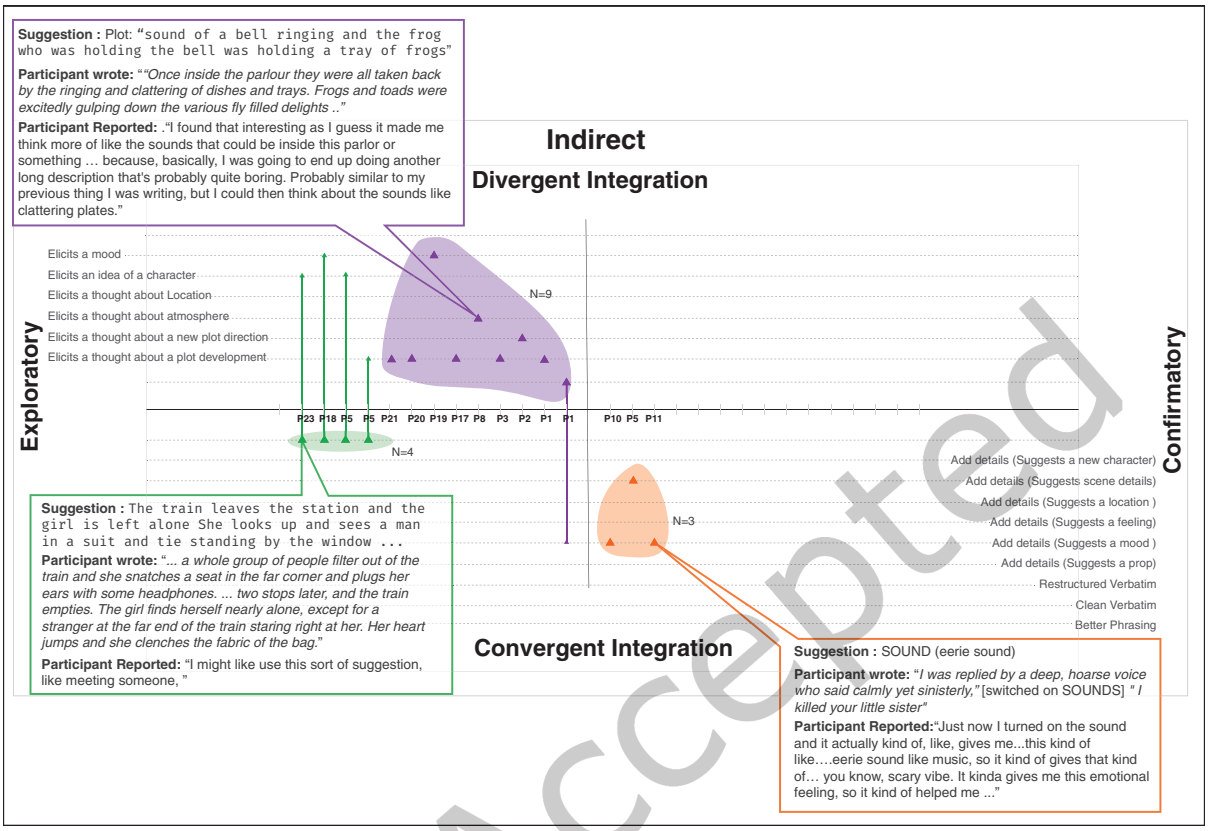


Fig. 6. Diagram of exploratory/confirmatory and divergent/convergent *indirect* integrative leaps made by the participants.

bottom is *convergent*. Participant IDs are noted along the horizontal axis, aligned to the corresponding instances. We can see a few patterns when surveying these leaps in total. For example, participants generally made more *direct* leaps than *indirect* leaps, but these are also related to the other dimensions: most direct leaps were also *convergent*, addressing necessary and local narrative features, though there are several exceptions. Conversely, *indirect* leaps are slightly biased toward *divergent* integrations. Similarly, the *exploratory* label often coincides with *divergent*, but we can see several exceptions to this. On the sides, we include high-level descriptions of what each integration contributes to the developing story, with illustrative examples provided for each quadrant.

## 5.3 Outcome Evaluations

In addition to participants commenting on their experience during the interaction with the system, we were also interested in capturing overall impressions and specific thoughts post-writing related to aspects of the prior assumptions and participant explanatory models we captured.

*5.3.1 General Impressions.* We obtained general impressions with one open-ended question ("What are your impressions from using **Editor-Red**?"), and subsequently tagged these with overall sentimental valence (summary in Table 5). We found that a majority of participants ($N = 13$) noted largely positive experiences with the interface, offering considerably different reasons. Some participants felt the suggestions were impressive or surprisingly

relevant; for example, P20 noted that they were "pleasantly surprised by how spot on the predictions were at times," and P12 wrote "I was impressed by its ability to generate sentences based on the context."

Some participants indicated that suggestions were directly helpful. P1 made an explicit connection to writer's block, saying "I really enjoyed the visuals and suggestions. I'm not a very visual person and struggle with writer's block, so this was very helpful."

Others found suggestions indirectly helpful, due to mood enhancement or some other affective effects. P11 wrote that it was "Very helpful in bringing the mood up in writing. It helps create the ambience and emotions needed for the writing." P5 and P9 both noted that the interaction was "fun" in addition to noting other effects beyond helpfulness of suggestions. P9 wrote:

> " It was fun/funny. The plot suggestions often didn't make sense but the description ones **were either useful/thought-provoking or amusing to read** even if I didn't use any part. The sound wasn't directly influencing my ideas but having background noise was **relaxing**. The pictures sometimes were relevant and sometimes not, so I didn't stare at them too long when they changed. "

P15 indicated that suggestions weren't always relevant, but that they were willing to do the work to find ways to incorporate them, which ultimately did make them helpful:

> " It was definitely a place to draw inspiration from. **I would not use the suggestions as they are, but with some modifications, the ideas presented were definitely helpful.** The sounds are calming and peaceful. The background images are a hit or miss because they don't tell the same story as the text I am writing, but **if I thought about the images a bit more creatively rather than literally, they were a bit more helpful.** I don't ever see myself using the overlay version of the images though. "

We annotated several responses as being overall negative ($N = 6$). Most of these also included comments about suggestions being less relevant, but participants did not indicate that they were necessarily able to integrate them despite this (in contrast to P15). For example, P16 wrote:

> " good idea but implementation worse than I expected. Images and text suggestions did not match storyline well, some proposed options were too exotic like "ustrophobia", selection tool to find images was highlighting parts of word not whole word, sounds were not very relevant, "

P18 also commented on suggestion relevance, but focusing on the writing context and timing rather than the overall relevance, noting that "I think it's interesting, but I don't see the plot or descriptions as being particularly helpful unless it's for a writing prompt. Once you get into the story, the suggestions are not very relevant."

Still other participants seemed neutral about their experience with the interface ($N = 4$), either providing little direction in terms of what worked for them and didn't, or explicitly accounting for both strengths and weaknesses in a balanced manner. P14 wrote:

> " It was sometimes helpful when I don't have any ideas, but not super helpfulâĂę Sometimes the text might not make senseâĂę The images are usually slightly off topic, but that could be helpful in giving me new ideas (they are very aesthetic/instagram-feel)âĂę I liked the sounds when they existed (it really does bring me to the place) "

*5.3.2 Suggestion Helpfulness.* Participants' overall impressions contained general indications of suggestion helpfulness, but we were also interested in obtaining fine-grained reflections to better elucidate the conditions and motivations involved in this. Again, participants diverged in what they found helpful and why. Here, we coded responses as indicating suggestions were **D**efinitely helpful ($N = 2$), **H**elpful ($N = 5$), **S**omewhat helpful ($N = 11$), **R**arely helpful ($N = 1$), or **N**o**t** helpful ($N = 4$), see Table 6 for full details. In addition to this, we found several different ways that suggestions were or were not helpful.

*New ideas, phrases, words.* A common indication was that suggestions yielded new ideas, possibly in phrase form but sometimes even as a word that was contextually useful. Although the suggestions typically contained

| | N | Example |
|---|---|---|
| Positive | 13 | "In Editor-Red I felt like I was writing in a time travel machine rather than staring at a blank page. I think it helped me feel a lot more grounded, present in the moment and in my body rather than just a disembodied brain trying to force words on a page." |
| Negative | 6 | "I didn't find it very helpful, but I could see how some features might be useful if developed further. The suggestions didn't seem to take into account the total content of what I had written, and so they seemed irrelevant and even distracting (for example, a phone scrolling instragram on a beautiful summer day when the phone in my story was a handset phone in a hotel room)" |
| Neutral | 4 | "It seemed fairly easy to use. I primarily was looking at the Plot section, possibly because it was more relevant to the given prompt while the Description section seemed to be more like different prompts. " |

Table 5. **Our sentiment labels for overall impressions from participants of writing with Editor-Red.** N indicates number of responses, Example shows a corresponding quote.



full sentences, subsets of these were more often described by participants as being helpful. The helpfulness of these ideas also extended to their presentations in other modalities (images and sounds). P21:

> " The suggestions were helpful. At one point, I felt that there had not yet been enough action in the story and the Editor suggested an event to happen. Another time, the sounds that were playing triggered an idea for the next setting of the story. Other times, even if I didn't like the suggestion I was given, it would still give me an idea for how to proceed. "

*Relevance.* Another reason suggestions were or were not helpful had to do with how relevant they were regarded as being in the context of writing. Most participants were mixed on this. For example, P8 wrote that the sounds seemed in-tune with the tone of the story they were writing, but described challenges to overall relevance of suggestions:

> " I was impressed that the sound suggestions seemed to pick up on the creepy, suspenseful tone of the story right away, and it could be helpful if the image suggestions followed the tone more closely. It kept showing me pictures of smart phones, which was not helpful. It would have been more helpful to see images of places and people for inspiration about how to describe their features. it would also be more helpful if the suggested images were more varied rather than all being pretty similar. That way the writer could choose from potentially useful images (and maybe even indicate which ones were more useful to see more like that?) "

P2, by contrast, found the sounds very distracting but the plot-level suggestions occasionally helpful:

> " the plot suggestions were occasionally helpful, but the descriptions were usually completely off the mark; the background images were aesthetic but not totally related, and the sounds were very distracting "

*Mood, continuity, and flow.* Some participants expressed effects on process rather than on content, as we also observed in their comments during the interaction. They suggested benefits beyond the direct application of suggestions, for example P1 noted:

> " For the most part I think that they were helpful. Even if I didn't use every idea that was suggested to me, I was inspired by their mood. "

P11 also pointed to this, writing about flow, mood, and ambience (the latter of which did seem to depend on relevance to the general topic of their progressing story):

> " Yes! They are definitely helpful! I think it helps to prompt me to think about what to write next and keep the mood on writing. It helps to keep the ambience according to the topic Im writing as well. "

*Redundant and repetitive suggestions.* Several participants noted the presence of redundant or repetitive suggestions, either repeating the content of the text in some form, or containing internal repetition. P15 wrote:

| | N | Example | D | H | S | R | Nt |
|---|---|---|---|---|---|---|---|
| New ideas/phrases/words | 10 | "The plot descriptions and text suggestions were helpful and creative. Some of the images, however were a bit generic. I mentioned a phone and the grid overlay just shoved several iterations of smartphones/, it would be nice if it could show different types of telephones, and in that way, allow the writer to have a visual reference, in case the writer wants to describe the object in more detail." | 1 | 3 | 6 | 0 | 0 |
| Relevance | 6 | "They were sometimes helpful and sometimes not. The suggestions were sometimes incoherent sentences, and they sometimes would not fit in well with the rest of the story." | 0 | 0 | 4 | 0 | 2 |
| Mood, continuity, and flow | 4 | "The suggestions were helpful. At one point, I felt that there had not yet been enough action in the story and the Editor suggested an event to happen. Another time, the sounds that were playing triggered an idea for the next setting of the story. Other times, even if I didn't like the suggestion I was given, it would still give me an idea for how to proceed." | 1 | 3 | 0 | 0 | 0 |
| Redundant/Repetitive | 3 | "I took some of the suggestions, but there was a lot of repetition (it kept wanting me to include a "hospital" scene, for instance). I mostly used it before a change in the narrative and it helped me think about plot development. " | 0 | 0 | 2 | 0 | 1 |
| Other | 2 | "For the suggestions menu, it didn't really work with me, only 1 out of 3 chances did I get to use it but it's quite similar with what the introduction video explained and I think they're helpful too in times of writer's block appear. " | 0 | 0 | 0 | 1 | 1 |

Table 6. **Codes from whether and how suggestions were helpful.** N indicates number of participants, Example shows a corresponding quote. Here, we coded responses as indicating suggestions were **D**efinitely helpful ($N = 2$), **H**elpful ($N = 5$), **S**omewhat helpful ($N = 11$), **R**arely helpful ($N = 1$), or **N**o**t** helpful ($N = 4$). Some responses are labeled with more than one.

> " The suggestions could definitely be improved. For example, the first suggestion I had kept repeating the same things in the plot box and the description box was very plain and essentially what I had already written. However, the second time I used a suggestion went better and I was able to draw inspiration from the images and plot box. "

*Additional thoughts.* Two participants found suggestions not or rarely helpful during the course of their writing, even when they might identify their potential value. Table 6 contains such an example, as well as a summary of all the other themes.

*5.3.3 Outcome Ownership.* Almost all participants ($N = 22$) indicated that they felt the outcome text was primarily or entirely theirs, citing a few different factors to reason about their ownership. No participants reported not feeling ownership, while one expressed a little uncertainty.

*Only took some phrases/words/ideas.* Most participants pointed to their cognitive and creative work in absorbing suggestions in the form of phrases, words, and ideas into their stories. P5 made a comparison to a collaborative process with another writer:

> " I would because I didn't take suggestions word-for-word except for one short phrase. Otherwise, it was like me bouncing ideas off a friend rather than the friend actually writing prose for me. "

*Ideas were primarily theirs.* Other participants made the perhaps related argument that general, global aspects of the stories were their own. From P13:

> " Yes because while I used the suggestions somewhat, the general storyline was my own. "

*Autonomy and authorial discretion.* P22 pointed to authorial discretion, i.e. "final cut", as the source of their ownership over the outcome: "I think whatever platform you use to brainstorm, in the end of the day, you are the decision maker to put that into your writings or not."

P21 referenced their work in not only deciding *how* a suggestion might be integrated, which we have discussed earlier, but perhaps whether or not to integrate it altogether:

> " I would call it my own. While I did receive inspiration from the Editor, there was nothing that I took verbatim from the Editor. I didn't include anything without putting thought into whether or not it would add to the story "

*Similar to real-world encounters.* Some participants compared the references obtained through working with **Editor-Red** to real-world encounters or analogous explorations of open domains like the internet. P11 wrote:

> " âĂęJust like when we try to find ideas through browsing the internet, or just **having a walk outside to create the mood and inspiration to write.** But editor Red makes it more efficient and **easier to find idea since it is all in one platform.** "

*Suggestions were not helpful.* Some participants felt ownership due to not incorporating any suggestions. P8 very simply stated: "I didn't really use any of the suggestions."

*Additional thoughts.* P1 indicated "I would call it my own, but acknowledge the suggestions that were made to me." They were the only participant that suggested such an acknowledgement. All these codes are summarized in Table 7.

*5.3.4 Differences from Initial Expectations.* To capture the ways in which writing with our interface qualitatively differed from their expectations of it, we posed an open-ended question. We coded the responses both for overall difference (Yes/No/Unsure) and themes that explain how or why, if so. A majority of participants ($N = 13$) indicated a difference from their expectations, with many also indicating that it was overall similar ($N = 9$). We coded one response as being unsure: P10 wrote "It was a novel experience. I had no expectations because I wasn't sure what to expect."

*Similar.* 6 participants noted that the experience was overall similar to their expectations. P14 wrote:

> " I have worked with language models before (I've played around with Writing with Transformer-type websites, using GPT2 for applications, and I've actually done an NLP externship that involved making image recommendations based on text haha using Unsplash too) so it was similar to what I expected in that it sometimes doesn't make sense, says things that are not super related, but could be coherent/interesting sometimes. "

*Less relevant.* Participants remarked that suggestions were less relevant than they imagined initially, if not always then some of the time at least. P4 remarked:

> " Sometimes I had no idea where the pictures in the grid came from, because sometimes they seemed relevant to what I had written and then sometimes it seemed like they were completely random. I expected the plot suggestions to be a little less repetitive. "

*Less out-there.* P2, among others, noted that suggestions were less big-picture or less out-there (which we use to imply further from the written narrative) than they anticipated:

> " fairly differently... I guess I was expecting some bigger-picture feedback, like larger plot suggestions or thematic images (like outer space for a space-related story ... though how clear would it be to AI that the story is about space?) "

*More creative/intelligent.* Two participants noted suggestions being more creative or intelligent than they expected. P1 wrote:

> " It was surprising to see the intelligence of the AI and the creativeness of the suggestions, for example "cryogenic sleep" was a very novel idea suggested to me. "

| | N | Example | Yes | Unsure |
|---|---|---|---|---|
| Only took some phrases/words/ideas | 10 | "Since some of the suggestions were pretty far out there, I would say that what I wrote is my own. However, there were interesting moments where I did copy a phrase that the system proposed. This felt more like plagiarizing than, say, picking a different word from a thesaurus. But I don't think the system can take too much credit since it was generating ideas based on my writing." | 10 | 0 |
| Other | 3 | "It depends. If I only used it to provide visual references and sounds for describing a scene, then would still call it my own, but if i got vital plot points from the editor suggestions, then I wouldn't fully call it my own work." | 2 | 1 |
| Ideas were primarily theirs | 3 | "It definitely feels very much like my own because almost all the words were mine, the story and progression is mine, the tone is very much mine." | 3 | 0 |
| Autonomy and authorial discretion | 3 | "I would call it my own because even though I did use the features to brainstorm, but I mostly write it on my own and I think whatever platform you use to brainstorm, in the end of the day, you are the decision maker to put that into your writings or not. " | 3 | 0 |
| Similar to real-world encounters | 2 | "Yes. Because it helps me find an idea, but I was the one who developed the story and make the story coherent. I think Editor Red is just a platform that helps a lot in creating ideas and mood in writing, not to help in writing the whole story itself. Just like when we try to find ideas through browsing the internet, or just haviing a walk outside to create the mood and inspiration to write. But editor Red makes it more efficient and easier to find idea since it is all in one platform." | 2 | 0 |
| Suggestions were not helpful | 2 | "Yes, because it was produced by me entirely " | 2 | 0 |

Table 7. **Codes from open responses regarding ownership over outcomes *after* writing.** N indicates number of participants, Example shows a corresponding quote, and Yes/Unsure are counts of participants reporting a feeling of ownership or being unsure (no participants responded no).



*Less subtlety/control.* P7 expected and desired differences in what parts of their writing the system attended to, indicating that they might like to do this through some control input:

> " I thought it would just look at the last few words that I had written and it would ideate on those ideas. Sometimes that was the case, particularly for the image suggestions. However, the text suggestions would sometimes go all the way back to the beginning of the story. I think that I wanted more control over where the AI was paying attention, but it otherwise did what I thought it would do. "

*Slower.* P16 remarked that the interface was slower to respond than they expected, in addition to suggestions being less relevant: "its interface acted as expected but I expected it to give suggestions faster and those be more relevant"

*Not as directly usable/helpful.* P9 and P3 noted that suggestions were not as directly usable or helpful as expected. P3 wrote: "I thought it would give me more suggestions/sentences that I would just copy into my writing directly. It was more of abridged words/phrases."

*5.3.5 Human-AI Differences.* We assessed differences in participants' practical expectations of our system and human writing partners with a counterfactual item: how did they think writing with such a partner might be different?

| | N | Example | Yes | No | Unsure |
|---|---|---|---|---|---|
| Similar | 6 | "Yes, it did. It is similar with the introduction video and I can use it easily by watching that." | 0 | 6 | 0 |
| Less relevant | 4 | "Sometimes I had no idea where the pictures in the grid came from, because sometimes they seemed relevant to what I had written and then sometimes it seemed like they were completely random. I expected the plot suggestions to be a little less repetitive." | 4 | 0 | 0 |
| Less out-there | 3 | "I expected the sounds to be more ambient sounds that contributed to a certain vibe of the story, but they seemed much more random and chaotic, which maybe had to do with the content of my story. I didn't expect that the editor would be able to draw upon elements of the story I had already written. I expected it to imagine new storylines that may have taken me in a new direction." | 2 | 1 | 0 |
| Other | 2 | "It was a novel experience. I had no expectations because I wasn't sure what to expect." | 1 | 0 | 1 |
| More creative/intelligent | 2 | "It was surprising to see the intelligence of the AI and the creativeness of the suggestions, for example "cryogenic sleep" was a very novel idea suggested to me." | 1 | 1 | 0 |
| Less subtlety/control | 2 | "The plot suggestions were not as subtle as I expected, like I thought it could help lead into plot points but instead the suggestions were often far off things I'd have to work toward and might take a bit of time to write to that point to incorporate the plot suggestions and make it make sense." | 2 | 0 | 0 |
| Slower | 2 | "see answer above for details. its interface acted as expected but I expected it to give suggestions faster and those be more relevant" | 1 | 1 | 0 |
| Not as directly usable/helpful | 2 | "The sounds were not what I was expecting. They sounded quite eery and scary, and I was trying to write something more lighthearted. Some of the suggestions were surprisingly good (like including characters that I'd mentioned and dialogue). Some of the suggestions also looped around though, and didn't really make sense (& I maybe expected them all to be kind of ok). I liked that the suggestions were slightly longer than I expected though." | 2 | 0 | 0 |

Table 8. **Codes from how using Editor-Red compared with each participant's expectations.** N is number of responses, Example contains a quote associated with the label, and Yes/No/Unsure indicate whether the experience was overall different than expected.



*More collaborative/communicative.* Most commonly, participants expected greater communication and a more collaborative interaction from a human co-writer. P7 noted this:

> " I think I would want another human to be more coherent. I would expect them to make more meaningful contributions to the work and I think that would feel more like a collaboration. This felt like I was immersing myself in a writing environment where the things were tangentially related to what I was writing, but not exactly relevant. "

*More understanding/experience.* Other participants alluded to experience of the world or understanding about more general aspects of narrative development. P6 wrote that "I could explain to the person the story I am thinking about. I could convey the tone and the feelings I am trying to infuse in my text"

*Speed (slower/faster) or effort.* Participants either thought writing with a human would be slower or faster, or require less or more effort in comparison. More commonly, participants thought it would be slower and require more effort. P22 reasoned: "I think, it will require more efforts because mutual agreements between the other humans are needed and the outcome really depends on how you and that person's relationship, their background, mutual understanding, etc."

P11 similarly expected writing with a human to be slower and/or require more effort, especially one who isn't a professional author. They attributed this to time needed for information processing, searching with other tools, etc.:

| | N | Example |
|---|---|---|
| More collaborative/communicative | 10 | "I think that another human being would have asked questions along with suggestions in order to better tailor their suggestions." |
| More understanding/experience | 8 | "Yes. I think the human would be more helpful because they could suggest if more description should be added to the setting or if the character needs to be developed more or suggest a direction for the plot, none of which I felt I got from the editor-red suggestions." |
| Speed (Slower/Faster) or Effort | 4 | "Slower going. Plot suggestions would have made more sense, but I don't think description ones would have been much difference in help. A human might have been more helpful with naming things in the story." |
| More questions | 3 | "I think we probably wouldn't asked each other questions back and forth like "why is the character doing this? What does it sound like?" etc. Writing with humans tends to involve a more "question-based" approach." |
| Other | 2 | "Another human might have challenges coming up with ideas just like the writer. The suggestions might have been different which would have geared my story in a different direction altogether." |
| Less self-driven | 2 | "Writing with another human would have definitely changed the course of the story. It also would have felt less like my writing because of the other person's ideas. " |

Table 9. **Codes re: expected differences from writing with a human co-writer, *after* writing.** N indicates number of participants, Example shows a corresponding quote. N.B. some responses are labeled with more than one code.

> " I think it would take more time if I write with another human since he/she would have to think for ideas and suggestions as well, or if not, he/she might also use another artificial intelligence like Google to find more ideas. Unless, the human is a professional author. If not, I think it will take more time to write as I would have to discuss as well. In addition, another human would not be able to create the mood/ ambience that I would like to have while writing. "

*More questions.* P3, among others, expected more questions along the writing process:

> " I think we probably [would've] asked each other questions back and forth like "why is the character doing this? What does it sound like?" etc. Writing with humans tends to involve a more "question-based" approach. "

*Less self-driven.* P14 and P15 expected the process to be less self-driven or self-owned, P14 wrote: "It might also feel less introspective (I enjoy the space from being alone)," while P15 alluded to ownership (see Table 9.

*Additional thoughts.* P23 expected that another human might face similar challenges to the writer (see Table 9 for quote) and P16 simply indicated a general difference, without specifying their reasoning about why.

## 5.4 Relating participant expectations, processes, and outcomes

Our three-fold study data collection generated a great deal of information from relatively few participants, describing each one's interaction with our system in substantive detail. Although our study design's primary goal is for these three types of data to collectively provide insight, here we review a few examples of instances where explicitly combining these sources of data at the level of the participant or sample provides additional information.

*5.4.1 Anchoring to prior expectations.* Regarding the possibility of AI creativity, P6 noted that they thought it "depends on the amount and type of data that will be available to the AI to create something new," emphasizing that the "broader and [more] various the set of data the more creative the AI could be." During the interaction, we observed P6 not engaging with the suggestions to advance their story, but rather, as discussed earlier, attempting to improve the system suggestions with their writing instead. In this case, an inaccurate expectation resulted in

the system being unhelpful to them, due to their behavior anchoring to this expectation rather than adjusting to the system's behavior during the process.

By contrast, some participants who were optimistic about the ability for AI to be creative managed to find utility in suggestions that may even have reflected poor or less coherent language-modeling behavior. For example, P15 wrote that "âĂęinformation/ideas provided by AI can be completely illogical which is sometimes the best creativity" and, after writing, indicated their willingness to make sense of and incorporate possibly irrelevant suggestions: "with some modifications, the ideas presented were definitely helpfulâĂę images are a hit or miss because they don't tell the same story as the text I am writing, but if I thought about the images a bit more creatively rather than literally, they were a bit more helpful."

*5.4.2 Adjustments to prior expectations.* Some participants appeared to adjust their prior expectations after interacting with the system. A particularly clear case is P1, who initially expressed a belief that "AI can not be creative," but could be "accurate." During the interaction with the system, having received a suggestion, P1 was impressed with it and was contemplating whether to characterize it as âĂIJaccurateâĂİ or "creative," finally coming to the conclusion that "this is a really good plot and it's creative enough." After the interaction, they noted that "It was surprising to see the intelligence of the AI and the creativeness of the suggestions." This participant initially identified that *Value (useful to people)* was the most important aspect of human creativity to them, and described the suggestions afterwards as "helpful. Even if I didn't use every idea that was suggested to me, I was inspired by their mood." The suggestions were useful to them in their process of writing, perhaps demonstrating the characteristics of *Value*.

At a sample level, a majority of participants ($N = 14$) initially responded that they would consider the final text to be "co-written by myself and AI," however afterwards, almost all participants ($N = 22$) indicated that they would call the written text their own (with one participant unsure). In addition, all those who responses Unsure or No to differences between human and AI text production initially ($N = 7$; P23, P17, P10, P6, P22, P9, P3) were able to communicate clear expected differences after the interaction. For example, P6 initially indicated that they were unsure, suggesting that "âĂęit depends on the level of development of the AI", but finally wrote that with a human they "could explain to the person the story I am thinking aboutâĂę convey the tone and the feelings I am trying to infuse," which is about explicit communication and intuitive influence rather than modeling performance. Participants had viewed a video of our system before the initial response, indicating that their perception was informed by actually interacting with the system rather than its overall design and method of suggesting.

*5.4.3 Do more accurate mental models of AI improve the experience or outcome?* To examine this question, we consider the two opposed categories of mental models of AI: Sparse-Abstract and Sophisticated-Operational. These groups respectively had the least and most detailed and accurate expectations of AI. We can examine how their outcome evaluations varied across the dimensions of overall experience, suggestion helpfulness, and differences from expectations.

*Sparse-Abstract.* 8/14 participants reported overall positive experiences, with 2 neutral and 4 negative. 10/14 total reported suggestions being at least sometimes helpful (1 rarely helpful, 3 not helpful). 9/14 indicated that the experience was different from their expectations, with 1 unsure and 4 reporting no or not much difference.

*Sophisticated-Operational.* 3/5 participants reported overall positive experiences, with 1 neutral and 1 negative. All 5 indicated that suggestions were at least sometimes helpful. 3/5 indicated that the experience was different from their expectations, with 2 reporting no or not much difference.

The data in this case indicate a complex relationship between the depth and accuracy in explanatory models of AI and experiences with our interface. A simple assumption might be that having more well-calibrated expectations of AI systems might arise from technically deeper and more accurate reasoning about how it generally works, which may appear to be supported by how helpful participants thought suggestions were.

However, our observations and participants' comments about suggestion helpfulness suggest that the differences have more to do with styles of writing and openness to narrative change than accurate expectations of the system's behavior. This is reinforced by the lack of clear difference in whether the system behaved as expected or not between these two groups.

Even so, we can explore this further by considering whether accurate expectations themselves were clearly associated with positive impressions or indicated suggestion helpfulness. 9/13 of those who reported a difference from expectations indicated an overall positive experience, compared with 4/9 who didn't. For helpfulness of suggestions, 9/13 who reported a difference from expectations found suggestions at least somewhat helpful, as compared with 8/9 who didn't. Again, there is a divergence between the two outcome variables, suggesting a complex relationship between expectations and outcomes.

## 5.5 Usability and overall experience

*5.5.1 Usage data.* In all, participants requested 165 suggestions and received 162 (3/165 requests were not resolved, for example due to requesting a subsequent suggestion while one was already processing). The median number of suggestions requested and received per participant over the 20-minute session was 7 (min = 2, max = 15). Participants wrote 229 words on average in **Editor-Red** compared with 296 in **Editor-Green**. With regard to suggestion modalities, on average participants had images toggled "on" for 99.8% (min 97.5%, max 100%) of the duration of their active engagement with the system, and sounds for 77.1% (min 25.2%, max 100%), with this duration measured from either the first suggestion request or alterations to these controls, to the last. During this period, participants toggled images on and off between 0 and 9 times (median 1) and sounds between 0 and 11 times (median 1.5).

*5.5.2 System usability.* As part of the post-task survey, we presented the participants with a battery of questions tailored to understand their experience using our interface, as shown in Fig. 7. The post-task survey contained 42 questions and produced more data than could be analyzed within the scope of this paper. We report those questions that examined the issues that are the focus of this paper.

Participants responded to **Q6.4** where they were asked to score using the likert scale Strongly Disagree (**DDD**)=1, Disagree (**DD**)=2, Somewhat Disagree (**D**)=3, Neutral(**N**)=4, Somewhat Agree (**A**)=5, Agree (**AA**)=6, Strongly Agree (**AAA**)=7 to the statement "The pictures used in **Editor-Red** were helpful." The responses were almost evenly distributed with one more participant disagreeing than agreeing (*AAA*=3, *AA*=3, *A*=3, *N*=4, *D*=8, *DD*=1, *DDD*=1). However, when asked to rate the statement **Q6.5** "The pictures used in **Editor-Red** distracted me from my task," the majority of participants disagreed (*AA*=3, *A*=2, *N*=2, *D*=3, *DD*=6, *DDD*=7).

We also asked participants to rate "The sounds used in **Editor-Red** were helpful." **Q6.6** (*AAA*=4, *AA*=2, *A*=5, *N*=1, *D*=2, *DD*=6, *DDD*=3), as well as "The sounds used in **Editor-Red** distracted me from my task," **Q6.7** (*AAA*=2, *AA*=4, *A*=3, *N*=5, *D*=1, *DD*=6, *DDD*=1) on both of which participants were somewhat evenly distributed between overall agreement and disagreement, with 5 participants vs. 1 participant neutral respectively. The distribution to these questions can be found in Fig.7A.

Participants responded to **Q6.11** "Using Editor-Red was intuitive" and **Q6.11** "Using Editor-Red was easy" we found that **Editor-Red** was considered intuitive to use (*AAA*=4, *AA*=9, *A*=6, *N*=3,*DD*=1) and easy (*AAA*=9, *AA*=10, *A*=3, *N*=1).

Finally, We were also curious to know how participants felt towards the different modalities available to them in the interface shown in Fig. 7B. When asked to rate **Q14.5** "I mostly used the textual suggestions and not pictures or sounds" participants agreed (*AAA*=8, *AA*=2, *A*=4, N=1, *D*=3, *DD*=4, *DDD*=1).

*5.5.3 Effort and cognitive load.* To understand the cognitive load imposed by writing with our system, we sub-sampled a group of 4 items from the NASA TLX (removing physical effort and performance, which are less

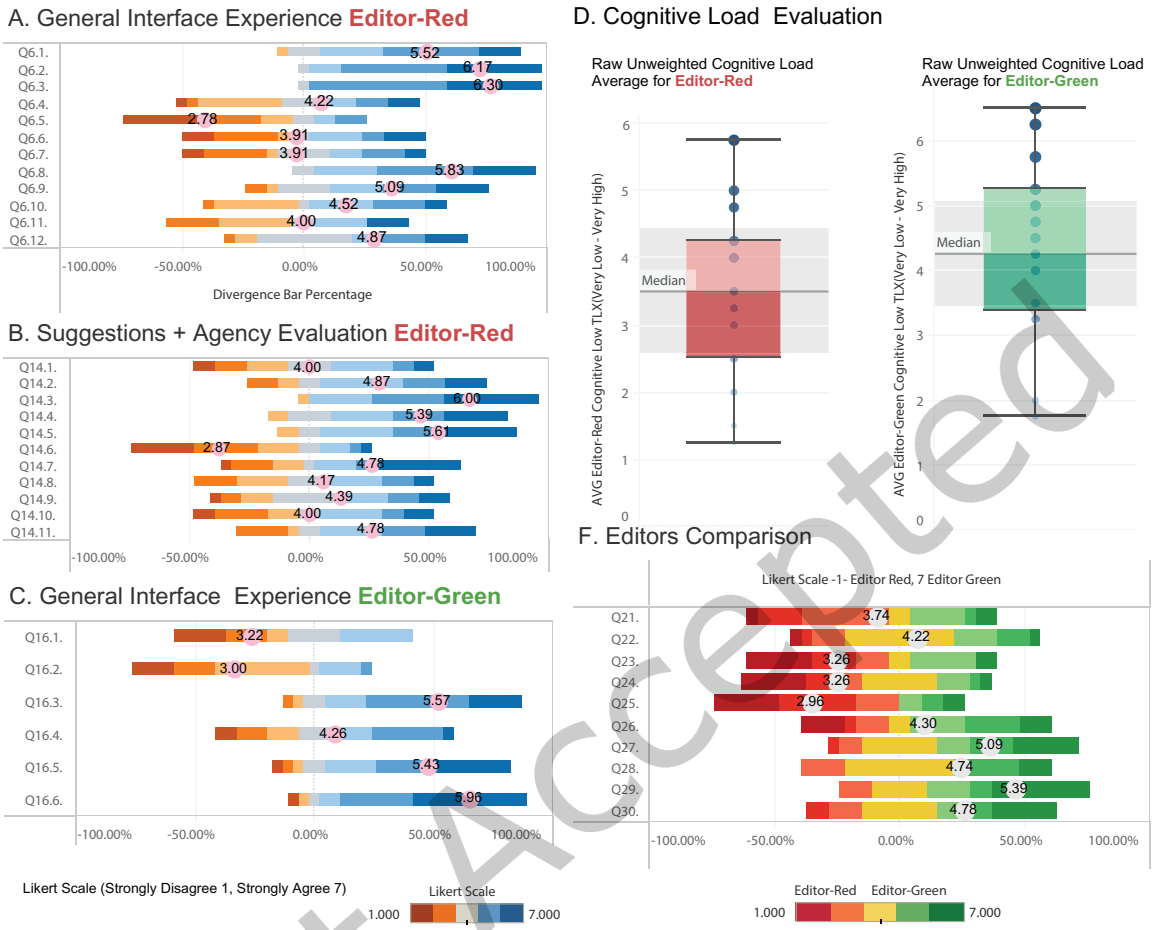


Fig. 7. **Post-task questionnaire focusing on user experience and editors comparison**. Full list of questions in appendix A.

relevant in our case) and then took the raw TLX score, which is simply the mean rating across all items per participant. We found that 17/23 participants rated the cognitive load of **Editor-Red** ($\mu$ = 3.41, $\sigma$ = 1.15, *Mdn.* = 3.50, *Min* = 1.25, *Max* = 5.75) lower than **Editor-Green** ($\mu$ = 4.13, $\sigma$ = 1.45, *Mdn.* = 4.25, *Min* = 1.75, *Max* = 6.50) Fig.7D.

*5.5.4 Evaluation of outcome creativity.* We captured the participants' perception of the creative output and of **Editor-Red**'s creative support by asking them to rate the following statement: **Q14.1** "I did most of the creative writing, using **Editor-Red** just for suggestions." Almost all the participants (22/23) agreed with the statement with the exception of one participant (*AAA*=9, *AA*=7, *A*=6, *DD*=1) . When asked to rate **Q14.8** "The suggestions made by **Editor-Red** were creative" we found that 16/23 participants agreed the suggestions were creative, 2/23 were neutral and 5/23 disagreed (*AAA*=4, *AA*=4, *A*=8, *N*=2, *D*=2, *DD*=3) Fig.7C.

We further asked for them to evaluate whether or not the final text created during the task was creative, using **Editor-Red**: **Q17** "Do you consider the text that you wrote in **Editor-Red** creative?" Most participants responded "Yes" (N=20), with 3 participants answering "No." Similarly, we asked them to rate the text they wrote with **Editor-Green**: **Q19** "Do you consider the text that you wrote in **Editor-Green** creative?" 14/23 participants responded "Yes" and 9/23 responded "No." We found that 13/23 participants preferred **Editor-Red** as the text generated during the task was perceived as more creative, and when asked: **Q21** "In which editor was the text that you wrote more creative?" (from **Editor-Red**=1 to **Editor-Green**=7), 2/23 indicated Neither or Both, with 8/23 participants towards **Editor-Green** (1: (N=1), 2: (N=4), 3: (N=8), 4: (N=2), 5: (N=5), 6: (N=1), 7: (N=2)). Participants show a preference towards **Editor-Red** for using that type of editor for writing creative text. **Q25** Which editor did you prefer for writing a creative text? **Editor-Red** (17/23) **Editor-Green** (6/23) (1: (N=6), 2: (N=7), 3: (N=4), 4: (N=0), 5: (N=2), 6: (N=2), 7: (N=2)) Fig. 7C.

### 5.6 Agency and ownership

We found that participants generally enjoyed writing with the help of suggestions from **Editor-Red** and were enthusiastic about the concept of writing with a âĂIJcollaborator,âĂİ especially once natural language generation capabilities improve and the suggestions are closer to their own writing style. From observing the writing session and post-survey conversation, it was unclear that the issue of ownership and agency in co-writing with AI was something that participants were at all concerned with or gave any thought to.

When prompted to rate the statement **Q14.2** "I enjoyed co-writing with **Editor-Red** " in our post-task survey, participants responded with 17 participants agreeing, 4 participants neutral, and 2 participants disagree (*AAA*=6, *AA*=5, *A*=6, *N*=4, *DD*=2). A similar response when asked **Q14.3** "I enjoyed collaborating with **Editor-Red**" was shown with 19/23 participants agree, 2/23 neutral, and 2/23 disagreeing (*AAA*=8, *AA*=4, *A*=7, *N*=2, *DD*=2) and when asked to rate the following statement **Q14.6** "The final product of writing is a result of joint efforts of **Editor-Red** and myself" participants responded with 10/23 agreeing, 4/23 neutral, and 9/23 disagreeing (*AAA*=2, *AA*=2, *A*=6, *N*=4, *D*=4, *DD*=3, *DDD*=2). Almost all the participants (22/23) responded that they only relied on **Editor-Red** for suggestions and did most of the creative writing **Q14.1**. (*AAA*=9, *AA*=7, *A*=6, *DD*=1) Fig. 7C.

Two participants admitted they were primed by questions in the pre-session survey to think about agency and ownership in writing using the suggestions of the system. P3 explained they had thought that âĂIJusing AI interface would make me feel that I wasnâĂŹt even doing my own writing," but ultimately P3 felt that the system âĂIJhelped alongâĂİ but âĂIJdidnâĂŹt tell me what to write.âĂİ P22 (one of the four participants who didnâĂŹt visibly use **Editor-Red**'s suggestions) explained that even if they had used the suggestions, they would still call it âĂIJusing my own creativityâĂİ as they believed that âĂIJeven deciding to use it or not, is actually really a choice for meâĂİ and is seen as a part of âĂIJcreativeness.âĂİ The participants seem to think that this type of system can improve creative writing by being supported by the system and less cognitively demanding than the simple text editor.

## 6 DISCUSSION

### 6.1 Suggestion quality: relevance, coherence, and variety

Several participants rejected suggestions for a perceived lack of coherence or relevance to their developing texts, which comports with prior work on language model assisted writing [18, 22]. Building on this, we have also shown that several others in our study did not see this as an obstacle to working with the system and in some cases appreciated less immediately semantically relevant suggestions and were able to incorporate ideas from less linguistically coherent suggested sentences. While we attempted to trace this difference to expectations, model behaviors, and other potential predictors, our expectation from observing participants is that this has primarily to do with a difference in participants' approach to creative writing. As such, the relevance of suggestions may not

be a simple variable to always aim toward maximizing; rather, the optimal level of relevance might vary by writer. Sometimes, it might also vary depending on other circumstances; for example, some participants noted that less relevant suggestions likely required more time to integrate, and that they might do so given additional time to write. This may also be reflected in the fact that on average, participants wrote less text in **Editor-Red** than in **Editor-Green**, though we note this is also related to other aspects of the interaction in our study (the novelty of the interface, participants talking more while using **Editor-Red**, etc.).

The ambiguity in assessing relevance extended to the multimodal concept representations; even when they were not used directly, their contribution to the environment might vary with their relevance. For example P8, who didn't visibly incorporate any suggestions, noted that they were "impressed that the sound suggestions seemed to pick up on the creepy, suspenseful tone of the story right away, and it could be helpful if the image suggestions followed the tone more closely" as compared with P5 who wrote that the "sound wasn't directly influencing my ideas but having background noise was relaxing."

Balancing relevance with variety is likely to be important in making suggestions useful to participants, in our assessment. Participants especially noted the homogeneity of images: "I mentioned a phone and the grid overlay just shoved several iterations of smartphones, it would be nice if it could show different types of telephones" (P20). This homogeneity also extended demographic factors: âĂIJthere's just a bunch of white guys staring at me and I don't know whyâĂİ (P2) and âĂIJthey are all images of straight blonde Caucasian womenâĂİ (P5). We noted that these instances were not directly related to query material, indicating that these might reflect broad biases in available images.

Technical approaches to generative modeling and information retrieval to support creative processes should, in our view, be intentional in handling these parameters (relevance, variety) and consider individual and situational variation in their optimality criteria. Modeling this is likely non-trivial and raises questions such as: what is relevant when and to whom? When are precise, logical suggestions needed, and when are surprising, unusual suggestions needed? The integrative leaps we have reported on suggest the practical challenges in automatically inferring this trade-off, or even reducing it to a simple, one-dimensional control. A helpful source of information in our case is the writers; they often have strong intuitions about both. Finding channels for writers to communicate their personal stylistic and contextual narrative needs to both interfaces and the underlying models, for example in natural language or by providing examples, may help these systems robustly support creative expression by both being flexible and allowing users to clearly and naturally communicate their needs and intentions.

## 6.2 **Editor-Red** beyond writing suggestions

*6.2.1 A supportive writing environment.* Though what **Editor-Red** seems to be doing on the surface is providing words, lines, and ideas to borrow and rely on, we observed much more than just that in the interaction settings we studied. A wide range of participants' comments highlight that the system acted as a support tool in diverse ways. Those participants who actively integrated the system's suggestions admitted that **Editor-Red** was structuring their process of writing. For instance, P1 admitted that they found themselves at a certain point "writing *for* the suggestions," seeing **Editor-Red** as "a form of motivation to continue writing" in order to get better suggestions. P3 commented that **Editor-Red** helped them "keep going" and "continue along" with their writing when they otherwise would have stopped.

**Editor-Red** redirected attention from being stuck (P12) and helped feel "less stuck" even when the participant was not taking the system's suggestions (P16). Writing as a process is fraught with self-doubt, anxiety, and feeling overwhelmed, systems like **Editor-Red** can mitigate stress by being a comforting distraction like "petting a cat" (P6).

Some participants used suggestions as just a starting point for the participants' own individual creative journey. For example, P23 explained that often **Editor-Red** suggestions gave them a different idea rather than taking

the suggestion right directly. As P23 further explained, getting inspiration from something can be unrelated to what that inspiration was. The integrative leaps (see §5.2.5) that participants made when engaging with **Editor-Red** illustrates a wide range of examples of what users are capable of and willing to do when integrating with a writing system.

*6.2.2 Personal and cultural references versus AI-generated references.* In the "blank page" writing with **Editor-Green**, 10 participants out of 23 visibly relied on cultural (books, TV shows, music videos) and personal references (memories, personal experiences, and immediate surroundings, e.g. describing what one can see from the window). For example, P8 writing in **Editor-Green** with the prompt "A train arrives at the station," explains that they are thinking about "the train station and Anna Karenina, kind of thing." P9 writing in **Editor-Green** with the prompt "The phone began to ring" explains that "the phone" made them think about a landline, a landline made them think about a hotel, and that, in turn, made them think about the last trip they had when they were staying in a hotel, which prompted a subsequent description they made in **Editor-Green** writing (after completing **Editor-Red** writing).

When writing in **Editor-Red**, 5 participants out of these 10 did not visibly use any cultural or personal references in their writing, but instead relied on the **Editor-Red** suggestions. Their story development was structured and oriented by the suggestions they incorporated. This is not that surprising on its own since participants were asked to use the features in **Editor-Red**. However, it is possible to imagine that writing with systems like **Editor-Red** allows a user to rely less on one's own cultural and personal references and one's "self" (an individualâĂŹs interiority, a means and ends of one's own actions [51, 52]). Rather, it provides an opportunity for a user to interact with the "self" of a system, deriving references from its suggestions.

We hypothesize that systems like **Editor-Red** can be used also when users for situational or psychological reasons do not want to engage with their own experience and inner thoughts and can be used for building systems that can provide therapeutic support for a user. This type of psychological support function has been recently identified in other human-AI creative interaction domains [91].

## 6.3 Dynamics of suggestion integration

Ideas for writing often came not through directly applying **Editor-Red**'s suggestions, but as a result of active engagement with the system from the participant's side and their readiness to do cognitive work in extending, adjusting, and altering suggestions and/or prior text to better suit the combination of text they had written and either any thoughts in their mind about how to proceed (confirmatory) or ideas about altering the narrative to lead in a new direction (exploratory).

The comments that participants made explaining these integration examples provide an insight into the multiplicity and multidimensionality of practices involved in human interaction with generative language systems, and especially how users create new meaning through this interaction. We specifically did not aim to do a linguistic or semiotic analysis of the integrative leaps that we documented, which we argue would require a great deal more data in order to yield generalizable insights. Instead, we aimed to document some orientation points that reflect structural differences in participant behaviors during our study.

*6.3.1 Creativity, inefficiency, and synthesis.* When interacting with **Editor-Red** , participants' concepts of creativity in suggestions often seem to be constrained by the possibility of an easy transition. The possibility of an easy transition, in turn, is individually and contextually varying. Those participants whom we identified as willing to cooperate with **Editor-Red** and incorporate its suggestions, did not seem to mind suggestions being "absurd," "crazy," and "out there" (we will refer to these as "out of sync"). These suggestions sometimes led to considerable changes to the subsequent and prior narratives; participants made decisive creative moves when they were willing to engage in this way.

Monster hunting, cryogenic sleep, a detective in 1890 Austria, and the first human to die on Mars were just some of the ideas that participants received as suggestions. Incorporating these suggestions depended on whether participants were mentally and emotionally ready to make the necessary efforts to synthesize an easy transition from the prior text and the given suggestion. It was "a much longer route" for monster hunting (P5), and didn't come "at a good time for the story" for cryogenic sleep (P1), and so these suggestions were not integrated. However, "you are a detective in 1890 Austria" made the participant think about the concept of time in their story (P1). P2, having completed almost the whole story that took place on a London farm, received a suggestion saying "you are the first human to die on Mars." P2 did comment that the description "is not very accurate" to their situation but then changed their mind: "You know, letâĂŹs make it about Mars, why not?" and rewrote the story in four places to fit the premise of taking place on Mars.

We conclude from this assortment of complementary and contradictory behaviors that the incorporation of a suggestion that is "too creative" does not depend just on the content of the suggestion, but rather on the possibility of transition which is influenced by individual and situational factors. The transition towards a suggestion that is unexpected and unrelated to the input text is dependent on the readiness and motivation of a user to the requisite cognitive and/or emotional work toward a meaningful synthesis of elements. These observations align with Freiman's characterization of the writer's drafting process, involving a "state of unknowing", a "kind of faith" that something will emerge from the drafting, and ultimately how "something that perhaps lacked cohesion or structure now becomes more concrete or coherent in the making of the text" [39]. Freiman suggests this happens by the writer making cognitive, affective, linguistic, and other creative decisions through a series of drafts and changes. We also expect that cognitive work done on drafting and revising to achieve such a synthesis may also become a path to support ownership of the text and creative endeavor, in our context of AI-generated suggestions.

*6.3.2 Cognitive reorganization and expectations of non-human writing systems.* What are the underlying cognitive mechanisms by which distant suggestions are able to be meaningfully integrated into users' existing narratives? Participants of our study were actively aware of the task environment [36]: writing a story using **Editor-Red** (non-human, AI system), a rhetorical problem (write a story given a prompt), integrating **Editor-Red** suggestions (which they had preconceptions of being based on human language, rule-based, possibly random, illogical, and creative), and the text itself that is evolving and changing. Since it was possible to get suggestions multiple times and the suggestions were different every time (both in content but also in terms of the level of relevance or detail), every new suggestion created a micro-moment of interaction and adjustment.

Attending to **Editor-Red** suggestions, building up all the missing cognitive links or not immediately visible links so as to update the story sometimes involves a considerable amount of cognitive reorganization of narrative information, in the sense of reorganizing what one already knows (e.g. Piaget's equilibration [75]) or, in this case, has already written. One possible mechanism for this is self-explanation, which is an attempt to make sense of new information by explaining it to oneself [20]. Unlike self-explanation in learning, wherein the central inferential process needs to construct new knowledge at the level of "the world", here self-explanation may provide an inferential process to reorganize the narrative by finding possible connections and associations, similarity, extracting abstract properties, or making referential links (for example, as we described earlier with P4 having the precondition of a crime, seeing an elephant that seems irrelevant, and explaining the presence of the elephant by making it the object of the crime involved). Other possible mechanisms for combining distant concepts have also been described in prior literature, such as causal reasoning [54], comparison and construction [97], conceptual integration or "blending" [25, 92], and satisfying constraints like diagnosticity, plausibility, and informativeness [24].

In the case of writing with our system, the willingness of participants to engage in this process may come from user expectations, due to the non-human source of the suggestions. For example, we noted earlier that participants may expect suggestions to be random, illogical, having connection to the real world yet often being

situationally "out of sync." In that way, any faults of the suggestions and difficulties that arise from those become not only something that has to be looked past but actually act as inherent features of the system and accepted as part and parcel of this interaction. These suggestions can be seen as not a bug but an inherent feature and a necessary condition of this interaction, as humans perform integrative leaps, engaging in cognitive reorganization of narrative because they accept the premise of interacting and co-writing with a non-human system, and the implications that come with it.

*6.3.3 How can "out of sync" suggestions be helpful for writing?* Earlier work has illustrated how completely unrelated ideas and unusual word combinations can be evocative and productive for creative writing [19, 31, 94]. In the case of causal reasoning, the surprisingness of combinations may provoke additional and exploratory processes and thereby the production of creative ideas [54]. We hypothesize that another mechanism by which semantically distant suggestions might be useful is by explicitly prompting more critical evaluations of written content, i.e. what Flower and Hayes call "evaluating" and "revising" [36]. By contrast, we might consider highly probable and user-adapted word predictions, which can be absorbed into a writing task with minimal effort (e.g. a click) to accomplish well-defined goals more efficiently (e.g. respond to a work email). We can model distant suggestions with such semantic difficulties as we observe as being useful inefficiencies which prompt critical evaluations of drafts and suggestions, metacognitive reflection about narrative development, and ultimately axes for more substantial narrative reorientation, where otherwise there would be no prompt or incentive to re-engage with and reconsider prior thoughts and writing. More work is needed to examine this possibility in detail.

## 6.4 Design recommendations

We observed that participants are capable of making leaps to integrate suggestions into their writing when presented even when the suggestions were unrelated to their current writing. However, there seems to be a general need to have these suggestions build on, refer to, or otherwise be relatable to aspects of their writing/story for many writers to have a more helpful experience. There is a need for details and descriptions of objects and important places when developing the story, and for systems to attend to the right parts of stories, which vary, when making suggestions. In our study, we observed most users rely on the system to enhance their writing when adding supporting material. When the system was not helpful in either introducing supporting material or helping them think of new directions, frustration and lack of trust in the tool often began to arise. However, as we have explored, suggestion quality is a multidimensional property which varies individually and contextually. To make suggestions useful to participants does not always mean maximizing their immediate relevance, but rather requires supporting the *process* of suggestion integration. Here we consider what that may mean for different technical and design considerations for creative writing support tools at every level of the process.

*6.4.1 Datasets for creative writing.* Participants in our study had different experiences with the two suggestion channels (*Plot* and *Description*), despite the commonality in the modeling method. Mirroring calls from other domains for data-oriented rather than model-oriented progress in AI [28, 88], we argue that well-curated datasets oriented towards domain constructs can support diversity and relevance, two factors we identified earlier as especially salient in machine contributions to creative writing work. Larger and more diverse pretraining sets can also result in greater coherence, if matched with an appropriately parameterized model, which we find would be helpful to several users in a variety of contexts.

*6.4.2 Language modeling.*

*What can better models help with?* As noted, more modeling power can result in increased coherence and relevance, especially as processed sequences get longer, if pretrained on appropriately large and diverse datasets, as well as fine-tuned on downstream datasets that provide creative value. These properties are desirable in many

cases, as pointed out by our study participants. In parallel, models with implicitly richer knowledge bases [74] may also extend more diverse suggestions to users, finding interesting relations with aspects of their writing, and assisting them in performing contextually appropriate and creatively fulfilling integrations.

*What can't better models help with?* Larger models are typically slower, more difficult to fine-tune and host, and increasingly closed-source, expensive to obtain access to, and private. Additionally, we noted many instances in which the cognitive work done by participants was the operative force in making suggestions helpful and ultimately able to contribute to their writing. For these participants, writing styles, and situations, larger language models may not necessarily help much, but would incur costs in interactivity, which were already pointed out by some participants in our current prototype. In our case, suggestions typically took 3-5 seconds after requests (given that we were running two separate fine-tuned models, extracting keywords, etc.), depending on the length of the input text; larger models may take significantly longer (one writer estimates GPT-3's Davinci model's typical speed at 147 words per minute [14]) and are very challenging to host and serve interactive requests with due to the resources needed.

Even the best possible language models have an extremely limited capacity to understand our intentions. They cannot reason about human internal cognitive processes, implicit judgments, and novel forms of creative exploration and expression that intentionally disregard convention. Better language models, better for different purposes, can support the process, but a great deal of what makes human creativity successful is outside of their purview.

*Semantic influence.* Some participants indicated a desire to influence or control this facet of suggestions with prior information, e.g. high-level story goals, moods, feelings, and ideas. While relevance can already be expressed to language models at *sampling* time to some extent, through stochastic decoding methods and controls like *temperature*, the ability to semantically "steer" relevance towards more fruitful integrations, rather than expressing it as a numerical value, might also better support diverse writers' diverse needs. Such steering can be explicitly enabled [50, 53, 58], for example, by conditional modeling, or, in the absence of specialized approaches, even discovered by so-called "prompt engineering" which has been successfully used[7] by many for language-controlled visual art generation [73] with general-purpose vision+language models [80].

#### 6.4.3 *Interface design.*

*Overall goals of interfaces.* Based on the behaviors observed in our study, we recommend that creative writing suggestions be designed to prompt and support cognitive processes that lead to suggestion integration and narrative engagement, rather than auto-complete style continuation. This seems to additionally support participant ownership over the outcome, as we observed in our study. There is a great deal of cognitive effort involved in writing with external stimuli, in order to make sense of them, recognize the possibilities for their contributions to the work, and perform effective integrations.

We argue that the focus of designing new creative writing support tools with intelligent augmentation should be on supporting this cognitive effort while preserving writer autonomy, authorial discretion, and creative flow. In our interface, we do this by implicitly discouraging directly absorptive behaviors; suggestions are presented in a different graphical environment rather than overlaid on the text, and the familiar tab key invokes new suggestions rather than directly integrating them into the writing. The corresponding reduction in cognitive load for most participants (17/23) by a small amount overall may reflect both the helpfulness of external suggestions in easing the cognitive burden of blank-page style writing, as well as the additional load introduced by the additional stimuli in context.

[7] https://ml.berkeley.edu/blog/posts/clip-art/

*Multimodal support for a unimodal task.* Additionally, visual and auditory suggestions cannot be simply inserted into a textual story, and we expect that the process of resolving these morphological differences to create meaningful semantic connections may also contribute to making creative leaps in writing stories. Our results suggest these features be made easy to turn off: this was a feature our participants used extensively to account for both individual and situational variation. Future work might examine the methods for communicating these parallel channels of information.

*6.4.4 Evaluation criteria and methodologies.* Our Expectation-Process-Outcome model, which guided our study design that combined surveys, behavioral observation, and semi-structured interviews, allowed us to capture several things: a rich representation of conceptually relevant background which participants brought into their interaction with a novel system, their interpretive reasoning through the course of the interaction, and their evaluative judgements and impressions afterwards. Additionally, through capturing prior assumptions and explanatory models, we were able to begin to obtain a fuller picture of how the interaction is framed by and adjusts expectations, as well as some effects this may have on the experience and outcomes.

We recommend that designers of complex, novel tools to support open-ended creative tasks similarly consider the conceptual priors of their users in conjunction with evidence from their experiences, behaviors, and *a posteriori* thoughts. Through this, we might begin to better characterize the significant level of individual and situational variation, and design tools that not only practically accommodate this but actively benefit from it.

## 7 CONCLUSION

This research presents an extensive study of machine-in-the-loop creative writing, centered around a new interface that makes writing suggestions through sight, sound, and language. Through collecting data on participant expectations, processes, and outcomes of interacting with this system, we discussed how individual writing approaches and narrative circumstances influence the interaction. By eliciting user explanatory models of AI, human and AI creativity, and creative writing, we explored how expectations might influence and be influenced by the interaction. We additionally reported on users' responses to suggestions through the lens of *integrative leaps*, by which participants incorporate suggested ideas into their writing process by performing cognitive work to make transitions possible.

As AI-based systems increasingly engage in traditionally human creative capacities, building stronger and more adaptive human-centered foundations for human-AI creative interaction will be increasingly important. Modeling advances in the systems periphery of everyday life have made it increasingly plausible that AI can be creative, but the more challenging work is to make it plausible that it might broadly extend our creative faculties by understanding our needs differently than other human creative partners. We believe that deep and wide-ranging investigations such as those we described in this work can inform design methodologies and yield powerful and useful tools that extend our abilities.



## ACKNOWLEDGMENTS

This material is based upon work supported by the NSF under Grant No. 2107391.

## REFERENCES

[1] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara Martin, and Mark Riedl. 2019. Guided neural language generation for automated storytelling. In *Proceedings of the Second Workshop on Storytelling*. 46–55.

[2] Kenneth C Arnold, Kai-Wei Chang, and Adam T Kalai. 2017. Counterfactual language model adaptation for suggesting phrases. *arXiv preprint arXiv:1710.01799* (2017).

[3] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.

[4] Kenneth C Arnold, Krzysztof Z Gajos, and Adam T Kalai. 2016. On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 603–608.

[5] Alan D Baddeley and Graham Hitch. 1974. Working memory. In *Psychology of learning and motivation*. Vol. 8. Elsevier, 47–89.

[6] David Bamman, Brendan OâĂŹConnor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 352–361.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[8] Jeffrey Bardzell and Shaowen Bardzell. 2016. Humanistic Hci. *Interactions* 23, 2 (2016), 20–29.

[9] Eden Bensaid. 2021. *Multimodal generative models for storytelling*. Master's thesis. Massachusetts Institute of Technology.

[10] Eden Bensaid, Mauro Martino, Benjamin Hoover, Jacob Andreas, and Hendrik Strobelt. 2021. FairyTailor: A Multimodal Generative Framework for Storytelling. *arXiv preprint arXiv:2108.04324* (2021).

[11] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both complete and correct? multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2297–2306.

[12] Lukas Bieringer, Kathrin Grosse, Michael Backes, and Katharina Krombholz. 2021. Mental Models of Adversarial Machine Learning. *arXiv preprint arXiv:2105.03726* (2021).

[13] Margaret A Boden et al. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

[14] Gwern Branwen. 2020. GPT-3 creative fiction. (2020).

[15] 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. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[16] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[17] 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*. 1–13.

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

[19] Orson Scott Card. 1990. *How to write science fiction and fantasy*. Writer's digest books Cincinnati, OH.

[20] Michelene TH Chi. 2000. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology* 5 (2000), 161–238.

[21] Hwan-Hee Choi, Jeroen JG Van Merriënboer, and Fred Paas. 2014. Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educational Psychology Review* 26, 2 (2014), 225–244.

[22] 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.

[23] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *arXiv preprint arXiv:2107.07430* (2021).

[24] Fintan J Costello and Mark T Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science* 24, 2 (2000), 299–349.

[25] Barbara Dancygier. 2006. What can blending do for you? *Language and Literature* 15, 1 (2006), 5–15.

[26] Marie-Catherine De Marneffe and Christopher D Manning. 2008. *Stanford typed dependencies manual*. Technical Report. Technical report, Stanford University.

[27] Leslie A DeChurch and Jessica R Mesmer-Magnus. 2010. The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of applied psychology* 95, 1 (2010), 32.

[28] DeepLearningAI. [n.d.]. *A Chat with Andrew on MLOps: From Model-centric to Data-centric AI*. Youtube. https://www.youtube.com/watch?v=06-AZXmwHjo

[29] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3163–3174.

[30] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can hci use lay understandings of algorithmic systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[31] Stephen R Donaldson. 2008. *The gap into conflict: The real story*. CNIB.

[32] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.
[33] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).
[34] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4647–4657.
[35] Ronald A Finke, Thomas B Ward, and Steven M Smith. 1992. Creative Cognition: Theory. *Research and Applications: MIT press Cambridge, MA* (1992).
[36] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.
[37] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. 411–412.
[38] Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2015. Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 649–658.
[39] Marcelle Freiman. 2015. A âĂŸcognitive turnâĂŹin creative writing–Cognition, body and imagination. *New Writing* 12, 2 (2015), 127–142.
[40] Megan French and Jeff Hancock. 2017. What's the folk theory? Reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)* (2017).
[41] Richard P Gabriel, Jilin Chen, and Jeffrey Nichols. 2015. InkWell: A Creative Writer's Creative Assistant. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 93–102.
[42] William W. Gaver, Randall B. Smith, and Tim O'Shea. 1991. Effective Sounds in Complex Systems: The ARKOLA Simulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana, USA) *(CHI '91)*. Association for Computing Machinery, New York, NY, USA, 85âĂŞ90. https://doi.org/10.1145/108844.108857
[43] Clifford Geertz. 1973. *The interpretation of cultures*. Vol. 5019. Basic books.
[44] Susan A Gelman and Cristine H Legare. 2011. Concepts and folk theories. *Annual review of anthropology* 40 (2011), 379–398.
[45] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1âĂŞ12. https://doi.org/10.1145/3313831.3376316
[46] 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*. 1–12.
[47] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[48] Scott Barry Kaufman and James C Kaufman. 2009. *The psychology of creative writing*. Cambridge University Press.
[49] Margarita Kaushanskaya, Henrike K Blumenfeld, and Viorica Marian. 2020. The language experience and proficiency questionnaire (leap-q): Ten years later. *Bilingualism: Language and Cognition* 23, 5 (2020), 945–950.
[50] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).
[51] Paul Kockelman. 2006. Agent, person, subject, self. (2006).
[52] Paul Kockelman. 2013. *Agent, person, subject, self: A theory of ontology, interaction, and infrastructure*. Oxford University Press.
[53] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367* (2020).
[54] Ziva Kunda, Dale T Miller, and Theresa Claire. 1990. Combining social concepts: The role of causal reasoning. *Cognitive Science* 14, 4 (1990), 551–577.
[55] Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–34.
[56] Brian M Landry and Mark Guzdial. 2006. iTell: supporting retrospective storytelling with digital photos. In *Proceedings of the 6th conference on Designing Interactive systems*. 160–168.
[57] Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures.. In *LREC*. Citeseer, 2231–2234.
[58] Zhiyu Lin and Mark O Riedl. 2021. Plug-and-Blend: A Framework for Plug-and-Play Controllable Story Generation with Sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 17. 58–65.
[59] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse* 3, 2 (2012), 101–124.

[60] N. Macdonald, L. Frase, P. Gingrich, and S. Keenan. 1982. The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications* 30, 1 (1982), 105–110. https://doi.org/10.1109/TCOM.1982.1095380

[61] Viorica Marian, Henrike K Blumenfeld, and Margarita Kaushanskaya. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of speech, language, and hearing research* (2007).

[62] Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[63] Ravi Mehta, Rui Zhu, and Amar Cheema. 2012. Is noise always bad? Exploring the effects of ambient noise on creative cognition. *Journal of Consumer Research* 39, 4 (2012), 784–799.

[64] Philippa Mothersill and V Michael Bove. 2019. Design Daydreams: Juxtaposing Digital and Physical Inspiration. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. 265–269.

[65] Frieder Nake. 1994. Human-computer interaction: signs and signals interfacing. *Languages of design* 2, 193-205 (1994).

[66] Hugo Nicolau, André Rodrigues, André Santos, Tiago Guerreiro, Kyle Montague, and João Guerreiro. 2019. The Design Space of Nonvisual Word Completion. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 249–261.

[67] Ellen W Nold. 1981. Revising. *Writing: the nature, development, and teaching of written communication* 2 (1981), 67–79.

[68] Donald A Norman. 1983. Some observations on mental models. *Mental models* 7, 112 (1983), 7–14.

[69] Martin Nystrand. 1989. A social-interactive model of writing. *Written communication* 6, 1 (1989), 66–85.

[70] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase WritersâĂŹ Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[71] Sharon Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 576–583.

[72] Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.

[73] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[74] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).

[75] Jean Piaget. 1985. *The equilibration of cognitive structures: The central problem of intellectual development*. University of Chicago press.

[76] Ben Pietrzak, Ben Swanson, Kory Mathewson, Monica Dinculescu, and Sherol Chen. 2021. Story Centaur: Large Language Model Few Shot Learning as a Creative Writing Tool.

[77] Reinier Plomp and Willem Johannes Maria Levelt. 1965. Tonal consonance and critical bandwidth. *The journal of the Acoustical Society of America* 38, 4 (1965), 548–560.

[78] Paul Prior. 2006. A sociocultural theory of writing. *Handbook of writing research* (2006), 54–66.

[79] Philip Quinn and Shumin Zhai. 2016. A cost-benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 83–88.

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[81] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[82] David Ramsay, Ishwarya Ananthabhotla, and Joseph Paradiso. 2019. The Intrinsic Memorability of Everyday Sounds. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.

[83] Arie Rip. 2006. Folk theories of nanotechnologists. *Science as culture* 15, 4 (2006), 349–365.

[84] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. âĂIJI CanâĂŹt Reply with ThatâĂİ: Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

[86] Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.

[87] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory* 1 (2010), 1–20.

[88] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. âĂIJEveryone wants to do the model work, not the data workâĂİ: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[89] Phoebe Sengers and Bill Gaver. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems*. 99–108.
[90] Sidney L Smith and Nancy C Goodwin. 1971. Alphabetic data entry via the Touch-Tone pad: A comment.
[91] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.
[92] Mark Turner and Gilles Fauconnier. 1999. A mechanism of creativity. *Alternation* 6, 2 (1999), 273–292.
[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), 5998–6008.
[94] Thomas B Ward and E Thomas Lawson. 2009. Creative cognition in science fiction and fantasy writing. (2009).
[95] Thomas B Ward, Steven M Smith, and Ronald A Finke. 1999. Creative cognition. *Handbook of creativity* 189 (1999), 212.
[96] Max Weber. 1949. *Max Weber on the methodology of the social sciences*. Free Press.
[97] Edward J Wisniewski. 1997. When concepts combine. *Psychonomic bulletin & review* 4, 2 (1997), 167–183.
[98] Earl Woodruff, Carl Bereiter, and Marlene Scardamalia. 1981. On the road to computer assisted compositions. *Journal of Educational Technology Systems* 10, 2 (1981), 133–148.
[99] Amanda Woodward and Karen Hoyne. 1999. Infants' learning about words and sounds in relation to objects. *Child development* 70, 1 (1999), 65–77.
[100] Nan Zhao, Asaph Azaria, and Joseph A Paradiso. 2017. Mediated atmospheres: A multimodal mediated work environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–23.

## A QUESTIONS

| ID | QUESTIONS |
|---|---|
| 6.1 | I quickly figured out how to use Editor-Red |
| 6.2 | It was easy to come up with ideas while writing |
| 6.3 | It was easy to decide how I will continue this story |
| 6.4 | The more time I spend writing with Editor-Red, the better it gets. |
| 6.5 | The pictures used in Editor-Red distracted me from my task |
| 6.6 | The pictures used in Editor-red were helpful |
| 6.7 | The sounds used in Editor-Red distracted me from my task |
| 6.8 | The sounds used in Editor-Red were helpful |
| 6.9 | The story that I wrote in Editor-Red is coherent |
| 6.10 | The story that I wrote in Editor-Red is creative |
| 6.11 | Using Editor-Red felt intuitive |
| 6.12 | Using Editor-Red was easy |
| 14.1 | I did most of the creative writing, using Editor-Red just for suggestions. |
| 14.2 | I enjoyed co-writing with Editor-Red |
| 14.3 | I enjoyed collaborating with Editor-Red |
| 14.4 | I equally used textual suggestions and pictures and sounds |
| 14.5 | I mostly used the textual suggestions and not pictures or sounds |
| 14.6 | The final product of writing is a result of joint efforts of Editor-Red and myself |
| 14.7 | The suggestions made by Editor-Red were coherent |
| 14.8 | The suggestions made by Editor-Red were creative |
| 14.9 | The suggestions made by Editor-Red were grammatically correct |
| 14.10 | The suggestions made by Editor-Red were relevant |
| 14.11 | The suggestions made by Editor-Red were surprising |
| 16.1 | It was easy to come up with ideas while writing |
| 16.2 | It was easy to decide how I will continue this story |
| 16.3 | The story that I wrote in Editor-Green is coherent |
| 16.4 | The story that I wrote in Editor-Green is creative |
| 16.5 | Using Editor-Green felt intuitive |
| 16.6 | Using Editor-Green was easy |
| 21 | In which editor was the text that you wrote more creative? |
| 22 | In which editor was the text that you wrote more coherent? |
| 23 | Where did you feel more relaxed when writing a story? |
| 24 | Where was it easier to write a text? |
| 25 | Which editor did you prefer for writing a creative text? |
| 26 | Where did you feel more focused when writing a text? |
| 27 | Where did it feel more demanding when writing a text? |
| 28 | Where did it feel more rushed when writing a text? |
| 29 | Where did you feel you had to work harder when writing a text? |
| 30 | Which editor made you feel more discouraged or annoyed when writing a text? |

## PRIOR PUBLICATIONS

This paper has no relation to prior or concurrently submitted papers by the same authors.