Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing

Jeongyeon Kim
University of California, San Diego
San Diego, California, United States

Sangho Suh
University of California, San Diego
San Diego, California, United States

Lydia B Chilton
Columbia University
New York, New York, United States

Haijun Xia
University of California, San Diego
San Diego, California, United States

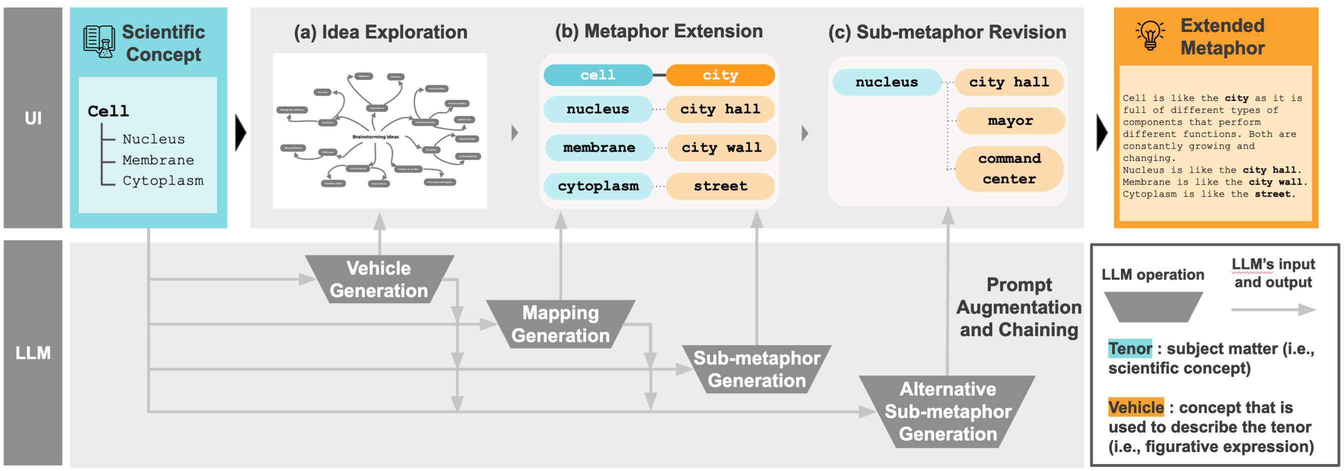

Figure 1: Overview of Metaphorian, an interactive metaphor creation support tool powered by the large language model (LLM). The system enables writers to (a) search, (b) extend, and (c) revise extended metaphors in science. The LLM-based workflow powers each writing phase via prompt augmentation and chaining.

ABSTRACT

Science writers commonly use extended metaphors to communicate unfamiliar concepts in a more accessible way to a wider audience. However, creating metaphors for science writing is challenging even for professional writers; according to our formative study (n=6), finding inspiration and extending metaphors with coherent structures were critical yet significantly challenging tasks for them. We contribute Metaphorian, a system that supports science writers with the creation of scientific metaphors by facilitating the search, extension, and iterative revision of metaphors. Metaphorian uses a large language model-based workflow inspired by the heuristic rules revealed from a study with six professional writers. A user study (n=16) revealed that Metaphorian significantly enhances satisfaction, confidence, and inspiration in metaphor writing without decreasing writers' sense of agency. We discuss design implications for creativity support for figurative writing in science.

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07.
https://doi.org/10.1145/3563657.3595996

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools.**

KEYWORDS

Metaphors; Creativity Support Tools; Science Writing; Writing Support; Large Language Model; GPT-3

ACM Reference Format:
Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3563657.3595996

1 INTRODUCTION

Science writers communicate complex scientific topics in a more accessible way to the general public. This includes introducing the fundamentals of science to children, explaining advanced topics to college students, and broadcasting new research findings in the news. In this regard, metaphors are effective tools in science writers' toolboxes because they explain unfamiliar topics in terms of familiar things. For example, a metaphor, 'a cell is like a city' can provide intuition for understanding what a cell is. However, because scientific concepts are often complex, a simple metaphor is rarely enough to explain a concept. Instead, science writers use extended metaphors to explain multiple components of the cell by extending the same original metaphor of a city. For example:

> A *cell* is like a *city*. The *nucleus* is the *city center*, where important decisions are made. The *mitochondria* are the *power plants*, providing energy for the cell. The *Golgi apparatus* is the *post office*, which sorts and delivers molecules. The *clysosomes* are the *garbage collectors*, breaking down and recycling old material.

Writing extended metaphors is difficult because it requires finding a familiar concept that can be extended in multiple ways to explain all the important aspects of a scientific concept. The subject of the metaphor such as the cell is called *tenor*, and the *vehicle* (e.g., city) must be found by creative means. The writer has to then find *sub-vehicles* (e.g., city center, power plants, etc.) for each *sub-tenors* (e.g., nucleus, mitochondria, etc.). This process is extremely challenging, as it requires finding vehicles and sub-vehicles that have congruent underlying structures as the tenors and sub-tenors. Given that professional writers already find it challenging to create a suitable simple metaphor [13, 76], the difficulty of creating extended metaphors that contain several sub-metaphors with strict structural constraints can be hard to bear.

The goal of this work is thus to facilitate the exploration and creation of extended metaphors to satisfy a desired structural congruence. To achieve this, a formative study was conducted to understand the practices and pain points of professional science writers while creating extended metaphors, as well as the desired qualities of said metaphors. The study showed that metaphor creation workflows involve the gradual expansion and iterative revision of sub-metaphors. Writers had significant difficulty brainstorming ideas without access to a pool of metaphor ideas and also finding sub-vehicles that satisfied the structural congruence.

To address their challenges, we support the creation of extended metaphors by leveraging a generative large language model (LLM). Generative LLMs, such as GPT-3, have been trained on terabytes of text and thus have the potential to offer a rich pool of metaphor ideas and develop the underlying links among sub-metaphors, which makes them compelling alternatives to metaphor generation relying on static embedding (e.g., word2vec) [14, 15, 33] or a knowledge graph (e.g., ConceptNet) [35]. We conducted an exploratory study of LLM—from which we selected GPT-3 from other popular LLMs[1]—and explored the capability of GPT-3 in generating metaphors that meet professional writers' requirements as well as derived heuristic rules that informed the prompt design to enhance the LLM to satisfy the writers' requirements. A technical evaluation showed that our enhanced prompt design improves the originality ($p < 0.05$), metaphoricity ($p < 0.005$), and coherence ($p < 0.05$) of metaphors compared to the vanilla prompt design.

[1]The set of "popular LLMs" was chosen from the pool of LLMs available in the mid-2022, which is when the exploratory study was done.

Based on the findings from the formative and LLM exploratory studies, we developed Metaphorian, an interactive system that enables writers to search, extend, and iteratively revise metaphors. The system supports the stepwise development of metaphor ideas using the LLM's prompt augmentation and chaining method (Fig. 1). Using Metaphorian, writers can explore the pool of metaphors as a source of inspiration. Metaphorian supports visual and semantic search by allowing writers to search for metaphors according to their originality, understandability, and thematic clusters. In addition, writers can iteratively expand and revise the metaphors as Metaphorian supports the further exploration of sub-metaphors while ensuring coherence among them.

The effectiveness of Metaphorian was then demonstrated via an empirical user study (n=16). The user study results indicated that Metaphorian provided an inspirational and immersive writing experience for the participants, enabling them to flexibly explore and iterate on the sub-metaphors while ensuring coherence with the main theme. Based on these findings, we discuss the design implications for building a creativity support tool by leveraging the LLM with human-AI co-creation. The main contributions of this work are thus:

- Practices and challenges of science writers while creating extended metaphors, identified from a formative study;
- An exploration of LLM's capacity, shortcomings, and augmentation strategies to generate extended metaphors;
- A dataset containing 600 extended metaphors consisting of 4,255 simple metaphors in science with ratings for fluency, creativity, metaphoricity, scientific precision, relevance, and coherence;
- A design and implementation of a metaphor creation support tool, Metaphorian;
- Results of an empirical user study demonstrating the effectiveness of Metaphorian in creating scientific metaphors.

2 RELATED WORK

2.1 Metaphors in Science Communication

Metaphors are used extensively in science communication because they act as crucial devices for funneling abstract scientific concepts to others. Science writers use metaphors to not only represent abstract scientific ideas in ways that are familiar to readers but also to give them names we can relate, to infuse life into them and afford a common language for seamless exchange of ideas [20]. While traditional literature on metaphors have focused on single metaphors, e.g., metaphors contained in a single sentence [75], researchers are beginning to examine extended metaphors [51], analyzing how they are processed differently by listeners than single metaphors and how their roles can be identified in various contexts [19, 60], such as writing [44, 68], marketing [32], teaching [20], and science communication [20, 78]. Extended metaphors are often found in science communication because even basic scientific concepts build on layers of other scientific sub-concepts. For example, if we describe hydrogen as 'a chemical element with the atomic number 1', the concept of an *atomic number* may also need to be explained because science writing is written for the general public who have limited knowledge of scientific terminology and concepts. Thus, it is common for science writing to extend new metaphors from a metaphor used for the main concept to explain its sub-concepts (or related concepts). A carefully selected set of metaphors (extended metaphors) allow metaphors to complement "parts of the explanation backgrounded by another metaphor" [20], making them a well suited and frequent pattern in science writing [20, 51]. However, despite their importance in science writing, no research investigated ways to facilitate the creation of *extended* metaphors for science communication—a gap this research aims to address.

2.2 Metaphor Generation

Because metaphors are essential in a variety of contexts—such as rhetoric [54, 68], poetry [34, 35], science writing [36], and teaching [20, 65], metaphor generation, along with metaphor identification, has long been a topic of interest in the artificial intelligence and natural language processing communities [35, 74]. Given a tenor and its attributes, some early work used probabilistic relationships between words to suggest candidate metaphors and produce simple "A is like B" expressions [10, 67]. Given a literal expression (e.g., "the party *ended* as soon as she left"), some researchers explored word embedding and seq-to-seq models to update verbs to transform them into a metaphoric expression (e.g., "the party *died* as soon as she left") [21, 64]. Using an open-source knowledge graph, Gero and Chilton [35] proposed an algorithm that calculated the semantic distance between vehicle features and a source word to find metaphorical connections between two words that users added to their system. Thesaurus Rex [72, 73], a web service, accepts two words (e.g., *soccer* & *basketball*) and returns shared category results such as *sport* and *game*.

Overall, metaphor generation research has focused on constructing a single metaphor (e.g., 'A is like B') as opposed to extended metaphors (e.g., 'A is like B, C is like D') [64]. This distinguishes our research from previous research on metaphor generation. Specifically, our approach addresses challenges and demands specific to the generation of extended metaphors, which have yet to be studied extensively. Therefore, we contribute a first step to understanding the constraints and challenges of creating extended metaphors as well as a comprehensive, annotated dataset of extended metaphors, which can serve as a dataset for training and testing in future research (like other metaphor datasets that have served a similar purpose [63, 69, 70]).

2.3 Large Language Models and Writing Support

LLMs have made substantial progress in recent years and demonstrated their effectiveness in a variety of tasks. Commercial services powered by LLMs now facilitate a range of tasks performed by copywriters [6], writers [9], and researchers [7]. They generate copy for copywriters, story ideas for writers, and summaries of papers for researchers conducting a literature review. A growing body of research has also explored a variety of writing-related applications. Goodman et al. [39] developed a mail writing system for adults with dyslexia that facilitated the process of writing emails by outlining the email's content and subject line, suggesting phrases, and offering rewriting features. Yuan et al. [84] designed Wordcraft, a text editor that enabled users to collaborate with generative language models while writing a short story. To provide inspiration to domain experts writing tweetorials, i.e., "short, technical explanations of around 500 words written on Twitter for a general audience", Gero et al. [36] used a mid-sized language model, GPT-2, that was fine-tuned to generate specific and diverse sentences. Although their system was also designed for science writers, our research focuses on the creation of extended metaphors and uses GPT-3 and prompt design tailored to extended metaphor generation.

Despite the rapid improvement of LLMs in understanding human instructions, prompt design can significantly affect the accuracy of the output on specific tasks, causing it to vary from almost chance to near state-of-the-art [85]. Thus prompt engineering—i.e, finding optimal prompt designs and strategies (e.g., chaining [77])—has been an important step in the process of inventing a workflow based on LLMs. Depending on the complexity of the task, prompts can be as simple as one sentence (e.g., "describe the concept of cell") or can span multiple sentences and contain task descriptions, examples, idiomatic phrases (e.g., "TLDR" for a summary), and the format of the output text. For tasks that can be broken into sub-tasks, Wu et al. showed that the chaining of prompts to each sub-task can be an effective strategy [77]. As explained in later sections, our work extends these works by developing prompting strategies and designs for generating extended metaphors and demonstrating how they can be integrated into science writers' workflow.

When using content generated by LLMs, it is crucial that humans remain in the loop to ensure the accuracy and understandability of the generated content. As our formative study will show, this is particularly important for metaphors in science communication, as the goal of the metaphors is to help the general audience accurately and easily understand scientific concepts. Therefore, the focus of this work is an interactive, human-in-the-loop system that enables writers to leverage LLMs to easily explore and create metaphors rather than improving LLMs' abilities in generating metaphors.

3 FORMATIVE STUDY

To understand the workflow, challenges, and support needed to write scientific metaphors, we conducted a formative study with six science writers (5 male, 1 female, F1-F6) recruited through the Upwork platform [4]. They had an average of 13 years (range: 4-25) of experience publishing scientific articles for the general public in newsletters, online blogs, and magazines. Interviews were conducted on Zoom, 90-minute long and semi-structured, and participants were given a $45 gift card as compensation. The interviews explored three themes: (1) the general writing process they used when creating metaphors, (2) the challenges they encountered, and (3) the support that they would like to have when writing scientific metaphors. The complete questions are in Supplementary Materials.

3.1 Analysis

The audio recorded from the interviews was manually transcribed and analyzed following an iterative coding process [42]. One author created a codebook for all transcripts using an inductive approach and refined the codebook through discussions with another author who analyzed half of the transcripts. After reaching a consensus on the codebook, they coded two randomly selected transcripts. Cohen's Kappa $\kappa = 0.76$, found good agreement, with an average of 88% agreement between the two authors. They then resolved conflicts and finalized the codebook. Each of the authors then independently coded the remaining interviews, met to discuss interpretations, and adjusted the coded data.

3.2 Usage and Workflow of Scientific Metaphors

We report the findings on the steps, challenges, and support that writers wish to have when creating metaphors.

3.2.1 Making Scientific Concepts Accessible and Engaging. Interviewees reported they used metaphors to make their writing understandable (i.e., making scientific concept easy to understand), memorable, and engaging, which aligns with prior literature on general metaphor writing [25, 52]. Interviewees elaborated that the primary goal of using metaphors was to increase the comprehensibility of complex and unfamiliar scientific concepts. F1 elaborated that "*I use metaphors for complex subjects where readers may not otherwise understand it.*"

3.2.2 Extensive Use of Extended Metaphors. While the notion of extended metaphors was not introduced to science writers during our study, all of them (6/6) stated that the metaphors they used in practice were mostly in an extended form that spanned multiple sentences or paragraphs and had multiple parallels to explain multiple concepts in a coherent narrative. F4 explained that he used a metaphor of a computer to explain the brain by mapping each component in the computer to the parts of the brain. F1 mentioned that the extended metaphors can even bridge across the entire article, stating that "*it [extended metaphor] is almost like webbing a spider web that connects different parts [of the article].*"

3.2.3 Interleaving Divergent and Convergent Thinking for Metaphor Creation. Based on the interviewees' responses, we could observe that the metaphor creation employed an iterative process of divergent and convergent thinking involving the following steps: (1) breaking down the main tenor into a set of sub-tenors (e.g., an atom consists of a nucleus and electrons); (2) identifying the properties of tenors (e.g., electrons orbit the nucleus of an atom); and (3) based on these properties, brainstorming vehicles that shared properties and structures with the tenors (e.g., planets spin around the sun). F4 said, "*the structure of scientific concepts should be parallel to that of metaphors. I mean, the two structures should be identical.*" Creating extended metaphors was not a one-shot process, but rather one that involved incremental expansion. For example, interviewees first came up with the main metaphor and gradually add sub-metaphors that were consistent with the main metaphor.

3.3 Challenges and Desired Support for Creating Extended Metaphors

The findings also identified the challenges that interviewees encountered while crafting metaphors as well as the desired support.

3.3.1 Lack of Ideas and Inspiration. Like many other creative tasks, the main struggle for interviewees was coming up with new ideas. For their writing to be impressive and engaging, they sought out original metaphors, e.g., "*worn-out metaphors just add noise to writing, having no impact as metaphors*" (F6). Due to difficulties finding original ideas, interviewees often experienced writer's block: "*I tried everything on the list [of possible metaphors], but nothing worked*" (F1). Some (4/6) mentioned that they referenced other writers' metaphors for inspiration by looking at science articles and research talks, but scientific metaphors were not easily accessible. F5 noted that "*high quality extended metaphors are not publicly accessible. I sometimes encountered them during an academic conference, which means [they are] not searchable and available in usual time.*"

3.3.2 Tension Between Understandability and Scientific Precision. Although leveraging metaphors to make the scientific concepts more understandable was a top priority, interviewees also found themselves wrestling with the need to champion scientific precision. For example, they used simple metaphors to make scientific concepts easier to understand but realized that such simplicity could result in leaving out important properties of the original concept, thus compromising scientific accuracy.

3.3.3 Ensuring Coherence in Extended Metaphors. As creating extended metaphors involved incrementally extending sub-metaphors, interviewees stated that metaphor creation was a cognitively demanding process because all vehicles must be coherent with each other. For example, when trying to create a metaphor for a submarine volcano, F3 related the volcano to a monster and the volcanic gas to the monster's breath. While continuing to write, he then sought out metaphors for hydrothermal vents that would relate to the already established metaphors; however, the established metaphor drastically reduced his explorable options. Interviewees said they would start over if they could not find a suitable vehicle for any one of the tenors to explain.

3.3.4 Desired Support in Tools for Metaphor Creation. Interviewees hoped to see a number of features that can support metaphor creation support, including semantic search, auto-completion of metaphors, and a search engine for metaphors. They also wanted to be able to filter their search results along dimensions, including (1) Theme: thematic groups that each metaphor was included in, (2) Distance: the semantic similarity between tenor and vehicle, (3) Difficulty: the level of difficulty of vehicles, (4) Property: the property of a scientific concept that the metaphor focuses on, and (5) Originality: the novelty of the metaphors.

3.4 Design Goals

Based on the interview results, the following design goals (**DGs**) were derived to guide the design of Metaphorian:

DG1. Enable semantic search for metaphors. Based on the formative study, science writers reference examples of metaphors for inspiration. They wanted to search for metaphors by the semantic relation between the tenor and vehicle. The semantic relations included, for example, how similar the vehicle and the tenor are. The system should provide users with semantic filtering or categorization for metaphor search.

DG2. Support the creation of understandable and original metaphors. Science writers' main concerns in metaphor creation were the understandability and originality of the metaphors. The system should be able to evaluate metaphors based on understandability and originality and enable users to explore metaphors based on these criteria.

DG3. Enhance chained extension of extended metaphors. The formative study revealed that writers do not create extended metaphors in one go but gradually expand them. In other words, sub-metaphors are added one by one based on their structural congruence with already created metaphors. Our system should enable users to incrementally extend the metaphors.

DG4. Enhance chained iterations of sub-metaphors. According to the formative study, writers consume a significant amount of time in iterating on sub-metaphor ideas, as even one invalid sub-metaphor can cause the whole metaphor to fail. Our system should allow users to explore and iteratively modify individual sub-metaphors when the initial candidate sub-metaphor is not apt.

4 EXPLORATORY STUDY OF LLM

The formative study informed the challenges science writers encounter and the support they need to write scientific metaphors, therefore how we should design the system. Since we sought to rely on LLM to generate extended metaphors, it was important that we explore (1) whether the LLM is capable of generating metaphors that meet writers' requirements and if it lacks, (2) how we can augment the LLM to generate metaphors that reflect the writers' requirements. To achieve these, we conducted an exploratory study to (1) compare the extended metaphors generated by the LLM with human-generated extended metaphors (Section 4.3) and (2) compare the extended metaphors generated by the vanilla LLM, with those generated by the LLM with the augmentation strategies that we devised (Section 4.4). The Discussion section includes further discourse pertaining to current LLMs, such as ChatGPT and GPT-4.

4.1 Model Selection

The first step when designing an LLM-based workflow is selecting a model. To determine which model would be appropriate for generating extended metaphors, we compared four representative transformers for text generation [41, 47, 71], i.e., RoBERTa [48], XLNet [80], GPT-2 [55], and GPT-3 (text-davinci-002) [18]. We compared the metaphors generated by these models based on the metrics introduced by the literature, distance of domains between tenor and vehicle [11], coherence of vehicles [21, 83], and diversity of vehicles [27]. Human rating is needed for evaluating creative tasks like metaphor writing, but the quantitative evaluation was enough in the model selection stage due to huge gaps in performance between the models. The evaluation results showed that there was no difference in the distance of domains between tenor and vehicle across models, but GPT-3 showed significantly higher scores for the coherence of vehicles and the diversity of vehicles compared to other models. Especially given no validated extended metaphor dataset on which the models can be trained and the open-ended nature of creative writing tasks, the comparative study showed that GPT-3 had the highest potential without further engineering to generate extended metaphors that would meet writers' requirements. Therefore, GPT-3 was selected as the NLP engine. The details of the study and hyperparameters are in Supplementary Materials.

4.2 Participants

Six science writers (5 male, 1 female, E1-E6) were recruited through Upwork [4], with an average of 5 years of experience publishing scientific articles for the general public in newspapers, blogs, or magazines (range: 2-9). They were paid $170 for taking part in a 5-hour study that was primarily conducted asynchronously, with the exception of a 30-minute introductory session.

4.3 Is the LLM capable of generating metaphors that meet writers' requirements?

In this section, we sought to quantitatively measure the quality of extended metaphors generated by GPT-3 by comparing them with human-made metaphors. We gathered qualitative responses from professional writers regarding which of their requirements for extended metaphors are not met by the GPT-3.

4.3.1 Constructing Datasets. To collect human-made metaphors as a baseline to which we can compare the LLM-generated metaphors, we asked participants to write extended metaphors for the given scientific concepts, selected from six representative domains [2], i.e., Astronomy, Biology, Chemistry, Computer Science, Earth Science, and Physics. We then considered three levels of difficulty, i.e., elementary school, middle and high school, and college level. Each participant chose three domains they had expert knowledge in and had expressed confidence in metaphor writing and evaluation. To prepare a set of LLM-generated metaphors, we used a default setting (i.e., a 0-shot) to generate extended metaphors, for the same domains and difficulty levels as the human-made ones. An example of prompt design is in Appendix. Since GPT-3 is non-deterministic (i.e., different outputs for the same input), we generated five different extended metaphors for each main tenor and selected the first five generation results to avoid cherry-picking the results.

4.3.2 Evaluation Criteria and Procedure. Once we had both human-made and LLM-generated metaphors, we asked the participants to score 18 metaphors (2 conditions * 3 domains * 3 difficulty levels) with the order of the metaphors randomized. Ratings were collected on both the vehicle and the mapping. For example, in "an atom is like the solar system because it is composed of a central nucleus with small particles orbiting around it", the solar system is the vehicle and the segment following the conjunction "because" was the mapping (i.e., "because it is composed of ... orbiting it").

To evaluate vehicles, we selected six criteria from the literature: originality [11, 27, 83], metaphoricity (i.e., literal vs metaphorical, e.g., literal expression: an enzyme is a biological catalyst; metaphorical expression: an enzyme is a lock and key) [11, 27], relevance of sub-vehicles to main vehicle [21, 83], coherence of sub vehicles (i.e., sub-vehicles are coherently connected being in the same domain) [21, 83], willingness to adopt [45], and inspirational effect [24, 57]. To evaluate mapping, five criteria were used: fluency [11], scientific precision, relatedness [27], willingness to adopt [45], and inspirational effect [24, 57]. The willingness to use and an inspirational effect was in binary scale and the rest of the evaluation criteria were on a scale of 1-5 (1: Very Poor, 5: Very Good). To control for the subjective nature of metaphor evaluation [35, 37, 82], we assigned three participants to each metaphor and used the average of the three scores. After collecting participants' ratings of the metaphors, we conducted 30-minute follow-up interviews to ask participants about (1) the deficiencies and failure cases of the LLM-generated metaphors and (2) their requirements for valid and inspirational metaphors to learn what is lacking in them.

4.3.3 Results. In this section, we report the results for comparison between LLM-generated and human-made metaphors along with writers' requirements for metaphors, not met by the LLM. Furthermore, we open-source the dataset[2] developed from this study to facilitate this line of research, which contains 600 extended metaphors that amount to a total of 4,255 single metaphors with ratings (e.g., fluency, creativity, and metaphoricity).

The results of comparing human−made and LLM−generated metaphors showed that for both vehicle and mapping, the LLM fell short of willingness to use and inspirational effect compared to human-made ones (Table 1). For other evaluation criteria (e.g., originality, coherence), we ran an unpaired two sample Mann Whitney U test, a nonparametric test for the unpaired data to investigate differences between conditions. Humans created significantly more original (U = 1404.0, p < 0.005, M_{human} = 4.0, SD_{human} = 1.24, M_{LLM} = 3.0, SD_{LLM} = 0.92) and metaphorical (U = 1800.0, p < 0.0001, M_{human} = 4.0, SD_{human} = 0.71, M_{LLM} = 3.0, SD_{LLM} = 0.77) metaphors than the LLM-generated metaphors. The sub-vehicles of the humans were also more relevant to the main vehicle (U = 1494.0, p < 0.0001, M_{human} = 4.0, SD_{human} = 0.91, M_{LLM} = 3.0, SD_{LLM} = 0.66) and there were no significant differences in coherence (U = 1233.0, p < 0.0001, M_{human} = 3.0, SD_{human} = 1.23, M_{LLM} = 3.0, SD_{LLM} = 0.68). There were no significant differences in fluency (U = 1089.0, p > 0.05, M_{human} = 4.0, SD_{human} = 0.98, M_{LLM} = 4.0, SD_{LLM} = 0.79) and scientific precision (U = 1251.0, p > 0.05, M_{human} = 3.0, SD_{human} = 1.03, M_{LLM} = 3.0, SD_{LLM} = 0.73) , but the LLM showed a slightly higher score for fluency. The human-made metaphors showed a higher relatedness (U = 1323.0, p < 0.05, M_{human} = 4.0, SD_{human} = 1.08, M_{LLM} = 3.0, SD_{LLM} = 0.76).

The follow-up interview results explained this huge gap between the LLM-generated and human-made metaphors. We identified five types of writers' requirements that were not met by the LLM. The writers' observations on the LLM's failure cases are summarized in Table 2 with quotations from the participants.

4.4 How can we augment the LLM to generate metaphors that reflect the writers' requirements?

Having identified the requirements (Table 2) for effective extended metaphors, we sought to augment the LLM to generate metaphors with fewer failure cases. Given the lack of a ground-truth dataset of high-quality extended metaphors and the lack of public access to the embedding space of GPT-3, it is technically infeasible for us to improve GPT-3 itself or employ common automated prompt generation techniques such as discrete (e.g., prompt mining [43] prompt scoring [29]) and continuous (e.g., prompting in the embedding space [46]) prompting. On the other hand, since GPT-3 is known for few-short learning and prior work has demonstrated the promise of simplifying the generation tasks with prompt chaining [77], we employed (1) few-short learning (10-shot) and (2) prompt design to improve the quality of generated metaphors. We conducted a comparison study to evaluate the efficacy of the augmentation strategies against the default setting of GPT-3.

[2]https://github.com/ucsd-creativitylab/metaphor_dataset

4.4.1 Augmentation Strategies. To improve the LLM's metaphor generation, we used two prompt designs: (1) **10-shot**, where the best metaphors by the ten professional writers (the metaphors that received the highest total score across the evaluation criteria) were used as the examples; (2) **10-shot with augmented prompt design**, 10-shot condition augmented by chaining prompts [77] and concatenating requirements to the prompt. The basic format of prompts was constructed based on the principles of prompt design for GPT-3 [58, 77] and observation on variations of prompts (e.g., paraphrasing, instructional style of questioning style). The details are in Supplementary Materials. Below we further explain the two augmentation strategies for 10-shot with augmented prompt design.

Prompt Chaining for Stepwise Metaphor Generation. The first method we used was prompt chaining, which is defined as the process of breaking up complex tasks into manageable steps, with each step being completed by an independent run of an LLM, and where the output of one or more steps is used as input for the next [77]. For example, Fig. 2 (a) and (c) illustrate this: *Main Vehicle* is generated from *Main Vehicle Generation* (Fig. 2a), it is then inputted into *Sub Vehicle Generation* (Fig. 2c) where it uses *Main Vehicle* to generate *Sub Vehicle* relevant and coherent to the main metaphor. An example of prompt design is in Appendix. As shown in Table 3, we can use a single prompt that includes multiple sub-tasks without prompt chaining. However, several factors motivated using prompt chaining. First, as shown by the output in Table 3, the reasons for each sub-metaphor (i.e., solar system - giant family, sun - father, planets - children, asteroids - cousins, comets - uncles) is not provided. Instead, only one reason—i.e., 'They all orbit around the sun ...'—is provided. Furthermore, without prompt chaining, the length of the prompt drastically increases in few-shot settings, and the LLM's inference time became too long to support real-time interactions in our system (Section 5). Thus, considering the performance and time efficiency of the LLM, the task was decomposed into four sub-tasks, i.e., generation of the main vehicle, main mapping, sub-vehicle, and sub-mapping (Fig. 2) to apply the prompt chaining.

Requirement Concatenation for Targeted Metaphor Generation. The second augmentation was adding the writer's requirements to the prompt. By adding the set of heuristic rules to the prompt, we sought to guide the LLM to generate metaphors that meet these requirements. As shown in Table 4, these rules, in the format of natural language text, were appended to the prompts.

4.4.2 Evaluation Setup. After generating metaphors using the different conditions, 0-shot, 10-shot, and 10-shot with augmented prompts, the same six professional writers rated a total of 810 metaphors (5 repetitive generations * 3 domains * 3 difficulty levels * 3 conditions * 6 writers). Like the previous evaluation session, willingness to use and inspirational effect were on the binary scale and the rest were on a scale of 1-5 (1: Very Poor, 5: Very Good).

4.4.3 Results. This section reports how the metaphor generation with the LLM was improved through a few-shot learning and prompt augmentation. As shown in Table 5, 10-shot improved the willingness to use and inspirational effect from 0-shot by 2 times

Table 1: The results of professional writers' ratings for LLM-generated and human-made metaphors. It shows the proportion of metaphors that are rated as willing-to-use and inspirational by the expert writers (# of metaphors = 54). For both vehicle and mapping, the human-made metaphors surpassed the LLM-generated metaphors.

Subject of Evaluation	Condition	Fraction of Willing-to-Use Metaphors	Fraction of Inspirational Metaphors
Vehicle	LLM-Generated	0.23	0.10
	Human-Made	**0.58**	**0.71**
Mapping	LLM-Generated	0.31	0.15
	Human-Made	**0.58**	**0.71**

Table 2: A list of writers' requirements in creating extended metaphors in science, along with interview responses as the rationale for the requirements.

Sub-task	Requirements	Rationale from Interview Responses
Vehicle Generation	**(Distant Tenor and Vehicle)** Vehicle should not be in the same domain as the tenor.	- *"When the scientific concept and metaphor are too similar, for example, the metaphor of 'the volcano is like a mountain' does not add any information"* (E5). - *"If the domain of scientific concept and metaphor is too close, it just becomes a literal expression, not a metaphor"* (E6). Example: "the volcano is like a mountain"
Vehicle Generation	**(Original Vehicles)** Vehicles should be original and novel.	- *"Unoriginal metaphors (from the dataset) are no more than noises to the writing"* (E1). Example: "the brain is like a computer"
Sub-vehicle Generation	**(Coherent Sub-Vehicles)** All vehicles should be in the same domain.	- *"To me, valid metaphors are the ones in which all the metaphors are in the same domain with high coherence"* (E5). Example: "the sun is like a light bulb. The planets are like family. The moon is like a sibling."
Mapping Generation	**(Closely Connected Tenor and Vehicle)** Every reason should closely connect the tenor and vehicles.	- *"When the link between the science concepts and metaphors are not clear, it was not helpful at all"* (E5). Example: "the event horizon is like the teeth of a monster since it is the point beyond which even light cannot escape"
Mapping Generation	**(Scientifically Precise Mappings)** Every reason should explain the tenor scientifically precisely.	- *"To be valid metaphors, they should explain the science concepts accurately and specifically"* (E2). - *"The metaphor should provide an overview of the scientific concept"* (E4). Example: "volcanoes are like the pieces of the puzzle being taken apart because they are the force that pulls the plates apart." Reason: *"when the plates pull apart, magma effuses through the gap and the volcanoes may form. But the causal relationship is not accurate in this metaphor"* (E2).

Table 3: An example of a prompt where multiple sub-tasks are described in one single prompt (as opposed to chains of multiple instructions). The inaccuracy of the output for the second instruction ("Write reasons...") motivated the prompt chaining.

Prompt	Output
"Write an extended metaphor in science for the main concept, solar system and sub-concepts, the sun, planets, asteroid, and comets. Write reasons for the metaphors."	"The solar system is like a *giant family*. The sun is the *father*, the planets are the *children*, the asteroids are the *cousins*, the comets are the *uncles*. They all orbit around the sun and have their own individual roles to play."

and 1.2 times, respectively. For mapping, there was no noticeable difference across conditions. We ran a Friedman test with a Nemenyi post-hoc test, a nonparametric test for three or more groups in which the same subjects show up in each group as one rater scored metaphors from the three conditions to investigate the statistical difference between conditions. We could observe significant differences between the three conditions except for scientific precision ($\chi^2(2) = 2.77$, $p > 0.05$) - originality ($\chi^2(2) = 31.47$, $p < 0.0001$), metaphoricity ($\chi^2(2) = 14.37$, $p < 0.001$), relevance to main metaphor ($\chi^2(2) = 36.25$, $p < 0.0001$), coherence of sub-metaphors ($\chi^2(2) = 48.51$, $p < 0.0001$), fluency ($\chi^2(2) = 9.80$, $p < 0.01$). This result aligns with literature that demonstrated that quality examples improve

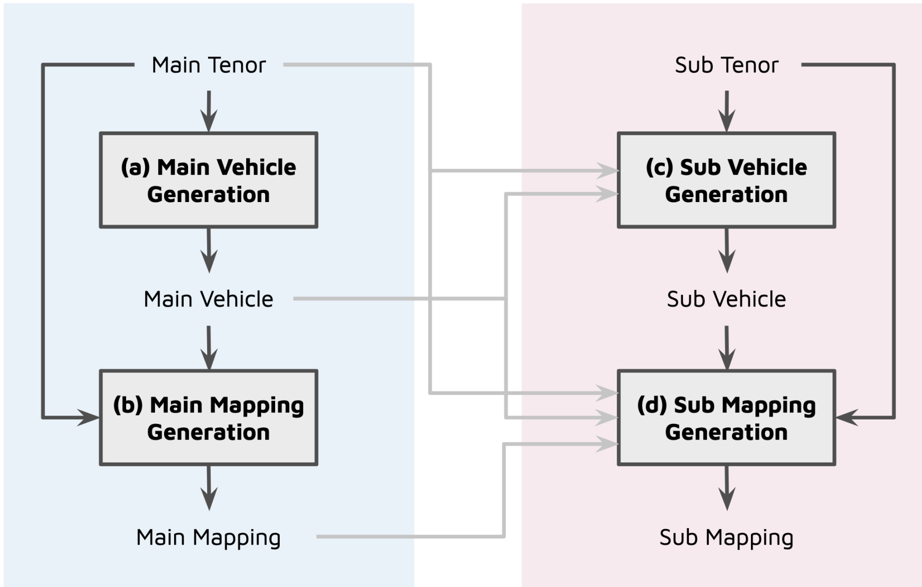

Figure 2: A structure of prompt chaining for extended metaphor generation consisted of four sub-tasks: (a) Main Vehicle Generation: generated the main vehicle based on the main tenor, (b) Main Mapping Generation: generated the main mapping based on the main tenor and vehicle, (c) Sub-Vehicle Generation: generated a sub-vehicle based on the main tenor, main vehicle, and sub-tenor, (d) Sub-Mapping Generation: generated a sub-mapping based on the main tenor, main vehicle, main mapping, and sub-vehicle.

the LLM's performance [18]. According to the post-hoc test, the 10-shot setting brought significant improvements compared to the 0-shot setting - originality ($p < 0.05$, $M_{0-shot} = 3.0$, $SD_{0-shot} = 1.24$, $M_{10-shot} = 3.0$, $SD_{10-shot} = 1.31$), metaphoricity ($p < 0.05$, $M_{0-shot} = 3.0$, $SD_{0-shot} = 1.55$, $M_{10-shot} = 4.0$, $SD_{10-shot} = 1.44$), relevance to main metaphor ($p < 0.001$, $M_{0-shot} = 3.0$, $SD_{0-shot} = 1.37$, $M_{10-shot} = 4.0$, $SD_{10-shot} = 1.31$), coherence of sub-metaphors ($p < 0.05$, $M_{0-shot} = 2.0$, $SD_{0-shot} = 1.44$, $M_{10-shot} = 3.0$, $SD_{10-shot} = 1.37$), fluency ($p < 0.01$, $M_{0-shot} = 4.0$, $SD_{0-shot} = 1.45$, $M_{10-shot} = 4.0$, $SD_{10-shot} = 1.31$), relatedness ($p < 0.05$, $M_{0-shot} = 3.0$, $SD_{0-shot} = 1.49$, $M_{10-shot} = 4.0$, $SD_{10-shot} = 1.29$). On the other hand, an augmentation of prompt design by appending the writers' requirements (Table 2) significantly improved the metaphors' originality ($p < 0.05$, $M_{10-shot} = 3.0$, $SD_{10-shot} = 1.31$, $M_{10-shot-aug} = 3.0$, $SD_{10-shot-aug} = 1.37$), metaphoricity ($p < 0.005$, $M_{10-shot} = 3.0$, $SD_{10-shot} = 1.44$, $M_{10-shot-aug} = 4.0$, $SD_{10-shot-aug} = 1.43$), coherence of sub metaphorical concepts ($p < 0.05$, $M_{10-shot} = 3.0$, $SD_{10-shot} = 1.37$, $M_{10-shot-aug} = 4.0$, $SD_{10-shot-aug} = 1.42$) compared to the 10-shot setting without prompt augmentation.

The analysis results also aligned with the participants' responses from the follow–up interview. For example, E5 compared the metaphors for the ecosystem: *"this one [10–shot setting] says ecosystem is like a machine and biotic components are like gears of a machine, and it [10–shot augmented] says biotic components are like the workers that make the ecosystem active and function. I think the worker metaphor is more precise."* E2 mentioned the originality of metaphors comparing two conditions. For superstring theory, 10–shot only generated superstring theory like 'a web' whereas the 10-shot augmentation condition provides more novel vehicles such as 'tapestry' and 'music' according to E2.

4.5 Takeaway

Through the explorative study of LLM, we sought to explore if the LLM is capable of creating metaphors that meet writers' requirements and find ways to improve the LLM to generate metaphors that reflect the writers' requirements. The results showed that the current LLM falls short of human writers in metaphor writing, and, in particular, we identified the LLM's five failure patterns that do not meet the writers' requirements (Table 2). Informed by the LLM's failure cases and writers' requirements, we enhanced the LLM's metaphor generation via few-shot learning, prompt augmentation and chaining. The improved LLM (willingness to use: 43%, inspirational effect: 47%) reached closer to the level of metaphors written by human writers (willingness to use: 58%, inspirational effect: 71%), compared to the vanilla prompt design (willingness to use: 23%, inspirational effect: 26%).

5 METAPHORIAN

Based on the formative and analysis study, an interactive metaphor creation support tool, Metaphorian was designed and implemented based on the GPT-3 model improved with 10-shot learning and prompt, which showed the strongest performance. The Metaphorian interface has two main areas: a Text Editor (Fig. 3a) where users can author an article and a Metaphor Creation Support area where users can search, extend, and iteratively modify metaphors (Fig. 3b).

Table 4: An example of an augmented prompt design using the heuristic rules revealed from the exploratory study of LLM. The writers' requirements for each stage were concatenated in the prompt.

Sub-task	Prompt with Concatenation of Requirements
Main Vehicle Generation	Write a metaphor for the {main tenor}, which covers all the requirements. **Requirements:** **Vehicle should not be in the same domain as the {tenor}.** **Vehicles should be original and novel.** Metaphor: The {main tenor} is like
Main Mapping Generation	Write reasons for the metaphor, which covers all the requirements. **Requirements:** **Every reason should closely connect the {tenor} and vehicles.** **Every reason should explain the {tenor} scientifically precisely.** Metaphor: {main tenor} is like the main vehicle. Reason:
Sub-Vehicle Generation	Given the main metaphor, write sub-metaphors for each sub-concept of the {main tenor}, which covers all the requirements. **Requirements:** **Vehicle should not be in the same domain as the {tenor}.** **Vehicles should be original and novel.** **All vehicles should be in the same domain.** Main metaphor: the {main tenor} is like the {main vehicle}. Sub-concepts of the {main tenor}: {sub-tenors} Sub-metaphors:
Sub-Mapping Generation	Given the main metaphor, reasons for main metaphor, and the sub-metaphors, write reasons for the sub-metaphors, which covers all the requirements. **Requirements:** **Every reason should closely connect the {tenor} and vehicles.** **Every reason should explain the {tenor} scientifically precisely.** Main metaphor: the {main tenor} is like the {main vehicle}. Reasons for main metaphor: {main mapping}. Sub-metaphors: {sub-vehicles} Reasons for sub-metaphor:

Table 5: The results of professional writers' ratings for 0-shot setting, 10-shot setting, and 10-shot augmented conditions. It shows the proportion of metaphors that are rated as willing-to-use and inspirational by the expert writers (# of metaphors = 810). For both vehicle and mapping, the human-made metaphors surpassed the LLM-generated metaphors.

Subject of Evaluation	Condition	Fraction of Willing-to-Use Metaphors	Fraction of Inspirational Metaphors
Vehicle	0-shot	0.23	0.26
	10-shot	0.40	0.32
	10-shot augmented	**0.44**	**0.35**
Mapping	0-shot	0.43	0.47
	10-shot	**0.49**	0.47
	10-shot augmented	0.43	0.47

To satisfy the design goals derived from the formative study (Section 3.4), there are four main features within the Metaphor Creation Support area: MetaphorMap, MetaphorFilter, MetaphorExtension, and MetaphorBranch.

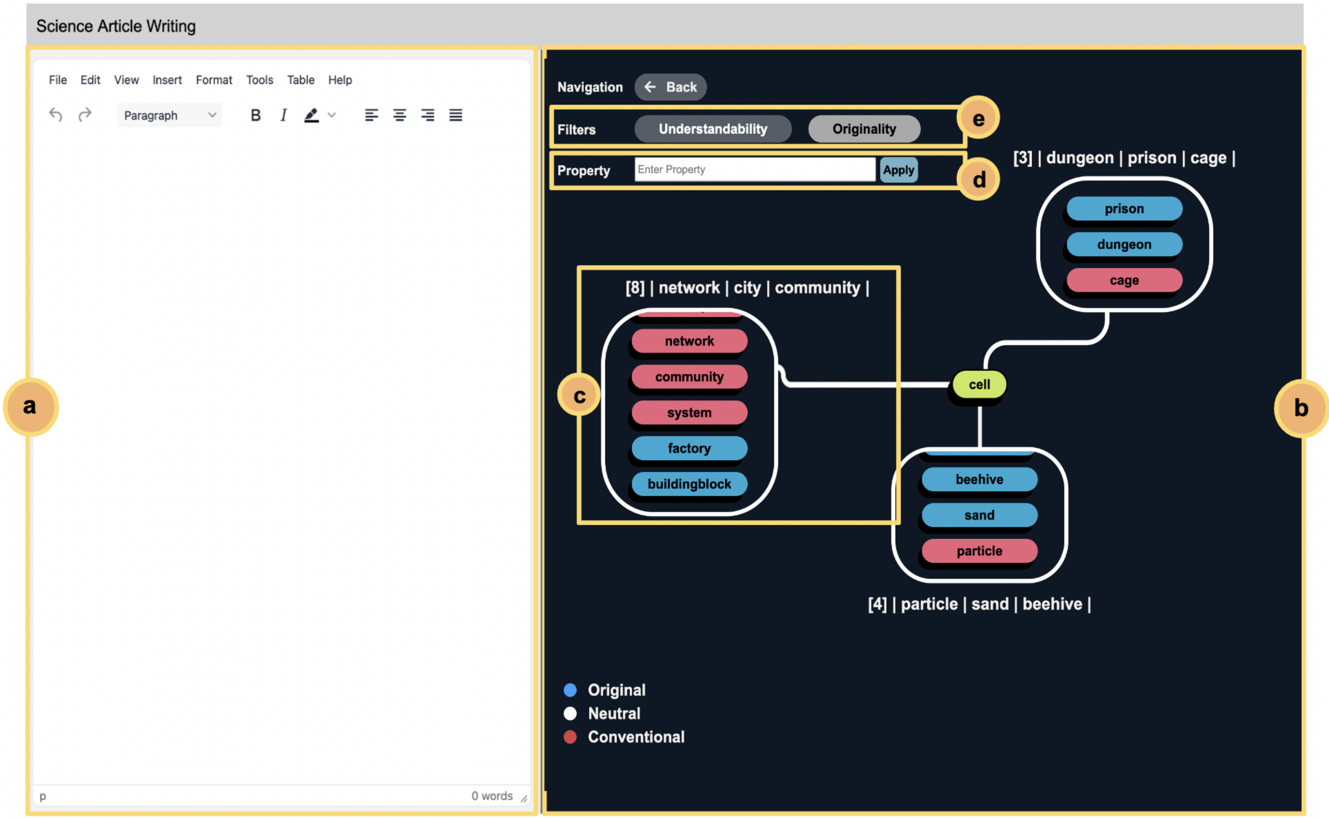

Figure 3: The Metaphorian interface is divided into (a) the Text Editor, where users can write text and (b) the Metaphor Creation Support area, where users can search, extend, and iteratively modify metaphors.

5.1 MetaphorMap - A Search Space for Ideas

MetaphorMap enables users to semantically search for metaphors (Fig. 3; **DG2**). The node at the center of the map represents the tenor entered by the user (e.g., cell in Fig. 3). The nodes branching out from the tenor node are potential vehicles (e.g., city, machine). Two designs are used to highlight the semantic relations between metaphors: (1) clusters of potential vehicles and (2) the distance between the tenor and vehicle clusters. Vehicles are grouped according to their semantic themes (e.g., network | city | community) (Fig. 3c). The numbers next to the themes (e.g., [8]) show the number of metaphor ideas included in each cluster. The spatial distance between the tenor node and clusters on the map represents the semantic distance between them. The farther they are from the tenor node, the more semantically distant the metaphors are. To group vehicles by semantic themes, hierarchical clustering [30] that is time-efficient and does not require several clusters was used. Keywords of the clusters (e.g., network | city | community) were implemented using the BERT embeddings [40]. The semantic similarity between the tenor and vehicles was computed based on cosine similarity in BERT-based embedding space [56].

The search by properties function (Fig. 3d) allows users to specify the properties of the tenor they want to focus on. For example, if they query the property of the DNA (e.g., 'DNA is passed down from parents to offspring'), focusing on the action of 'passing (something) down', then the system generates metaphors that highlight this property, such as 'DNA is like a recipe that is passed down from generation to generation.' If users input the solar system and its sub-concepts, the sun and planets, the system generates corresponding vehicles, such as family, parents, and children for each concept.

5.2 MetaphorFilter - Finding Original & Understandable Metaphors

MetaphorFilter supported the creation of understandable and original metaphors (**DG1**; Fig. 3e). If users clicked on the understandability button, the nodes of metaphor ideas would be color coded depending on the level of difficulty, blue for the scientific concepts that are easy to understand, white for neutral, and red for those that were comparatively hard to understand. To enable vehicle filtering by understandability, we adopted a frequency measure [17, 66] that estimated the understandability of vehicles depending on their frequencies in a large corpus [49, 62].

If users clicked on the originality button, the nodes would be recolored depending on the degree of originality. The novel ones were color-coded blue, neutral white, and conventional (unoriginal) red. To support vehicle filtering by originality, the degree of novelty of the metaphors was estimated by counting the number of overlaps between the LLM-generated metaphors with the existing ones from Metamia [5], an open database of metaphors to which anyone can upload metaphors. Although Metamia is not a validated dataset, it suffices to give an estimate of metaphors' degrees of novelty. The thresholds are in Supplementary Materials.

5.3 MetaphorExtension - Adding Vehicles

MetaphorExtension supports the enhanced chained extension of extended metaphors (**DG3**; Fig. 4). After the initial exploration of metaphor ideas using MetaphorMap, the system generates mappings (Fig. 4a) that explain the reasons for the metaphors that were chosen by users. If users chose multiple metaphor ideas, they would be displayed as multiple branches in the tree-structure diagram (Fig. 4b). Meanwhile, users can expand the metaphors by adding a new sub-tenor (Fig. 4c). The system then suggests the sub-vehicles that correspond to the new sub-tenor (Fig. 4d), in consideration of the relations to the already generated metaphors.

To generate a sub-vehicle that corresponds to a new tenor added by users, we used a chained prompt that included the already generated vehicles. We instructed the LLM to generate a new sub-vehicle in consideration of the current vehicles (Fig. 2) with writers' requirements appended (Table 2).

5.4 MetaphorBranch - Revising Sub-Vehicles

MetaphorBranch enhances chained iterations of sub-metaphors (**DG4**; Fig. 5). Users can iteratively revise individual sub-metaphors by referencing alternative ideas for them. If users select a sub-metaphor that they want to explore further (Fig. 5a), the system displays alternative sub-metaphor ideas as sub-branches in the tree-structure diagram (Fig. 5b). Meanwhile, if they click a magic wand button, a recommended metaphor idea is displayed, which corresponds to the input concept and its sub-concepts. The recommendation picks the idea randomly, with no overlap with the already chosen ideas, through simple string comparison.

To generate alternative metaphors for a sub-vehicle, we used the same prompt as the one used in MetaphorExtension. However, to avoid displaying the same metaphors, the prompt specified (1) which vehicles were already generated and (2) that newly generated sub-vehicles should not overlap with existing ones, i.e., with "Existing vehicles: {main vehicle}, {sub vehicles} *n* Requirements: The sub-metaphor should not be one of the existing vehicles" as the part of prompt.

6 EVALUATING METAPHORIAN

To test whether Metaphorian successfully supports science writers in creating extended metaphors, we conducted a user study with professional science writers to evaluate their writing experiences while using Metaphorian. It was designed as a controlled study with a within-subjects design where each participant compared the system to a baseline interface. To maintain uniformity in the appearance and feel of both interfaces, the baseline used the same interface design as our system, except that the Metaphor Creation Support area on the right was replaced with a Google search engine (Fig. 6). There are several reasons we designed our baseline with the Google search engine. First, the professional writers in the formative study revealed that the search engine is the primary tool they rely on for inspiration. Also, there are no tools for us to directly compare in either the academic or commercial market that support the creation of extended metaphors in science writing (e.g., Gero et al.'s work [35] focuses on metaphor creation in poetry and does not support the extended form of the metaphor). Furthermore, other language models fall short of science metaphor creation even to be used as the baseline, as we investigated in Section 4.1. Similarly, the simple adoption of the large language model without layers of features we designed for Metaphorian is not capable of generating extended metaphors nor far from the writers' current practice.

In the user study, we sought to answer the following questions:

- **RQ1**. How do writers use and benefit from MetaphorMap and MetaphorFilter when exploring metaphor ideas?
- **RQ2**. How do writers use and benefit from MetaphorExtension when expanding metaphors?
- **RQ3**. How do writers use and benefit from MetaphorBranch when iterating sub-metaphor ideas?
- **RQ4**. How does Metaphorian affect the creative writing experience of scientific metaphors compared to the baseline?
- **RQ5**. How does Metaphorian affect users' sense of agency compared to the baseline interface?
- **RQ6**. How do perceived qualities of metaphors created with Metaphorian compare to those created with the baseline?

6.1 Participants and Procedure

Sixteen science writers (11 male, 5 female; average age: 33, range: 24-59; P1-P16) were recruited through Upwork [4]. All had published scientific articles for the general public in newspapers, magazines, and online blogs. They had on average 8 years of experience (Std=4) as science writers. They received $30 for an hour-long study.

The study was conducted through Zoom. They were given two writing tasks, (1) write a short paragraph explaining the given scientific concept to the general public and (2) write a short paragraph explaining a topic of their choosing to the general public. The given topic was either 'cell' or 'DNA' with associated sub-concepts. These two topics were chosen because they (1) have a similar difficulty level (i.e., both are middle school level concepts [3, 8]) and (2) are likely to be familiar to every writer. The order of using the baseline and Metaphorian interfaces, in addition to that of the topics, was counterbalanced. Participants were encouraged to complete each task within 5-7 minutes but their time was not strictly limited. The study then concluded with a survey and exit interview. The survey consisted of questionnaires about participants' perceived writing experiences, their satisfaction with the metaphors, their willingness to use the system, their cognitive load, and their sense of agency.

6.2 Results

On average, it took participants more time to complete the writing tasks using the baseline (Scale = minutes, Baseline: Mean = 6.3, Std = 2.5, Metaphorian: Mean = 4.9, Std = 2.7, t (statistics) = 2.69, $p < 0.05$). Meanwhile, participants using Metaphorian submitted more content (Scale = number of words, Baseline: Mean = 82, Std = 36, Metaphorian: Mean = 105, Std = 60, t = 2.13, $p < 0.05$). In other words, participants wrote longer metaphors in a shorter time using

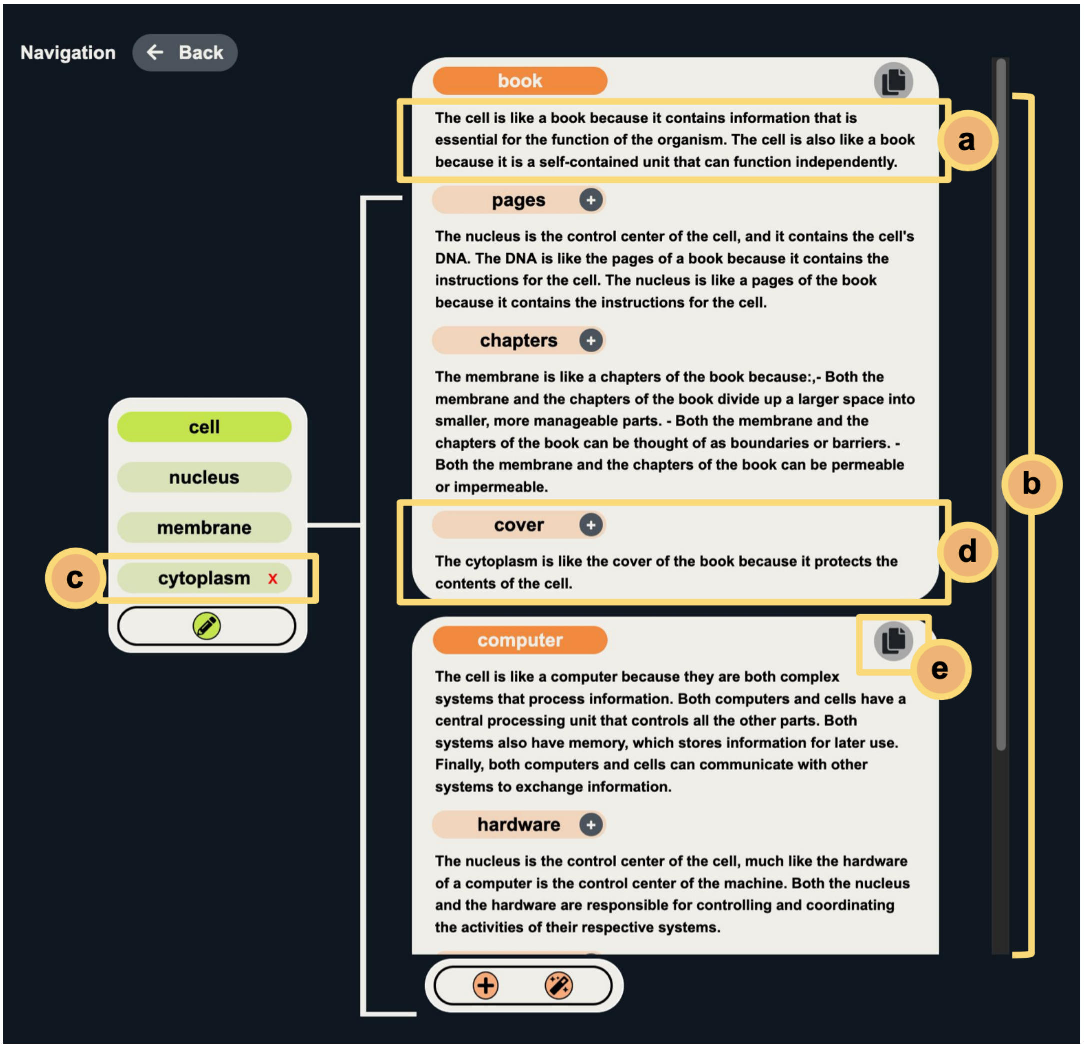

Figure 4: MetaphorExtension extends the metaphors: (a) Mapping: shows the reasons why a tenor is metaphorical for a vehicle, (b) Branch: in the diagram displays ideas chosen by users, (c) Sub-Tenor: can be added by users to extend the metaphor, (d) Sub-Vehicle: is generated, which corresponds to the sub-tenor, (e) button for adding the generated text into the Text Editor.

Metaphorian than when using the baseline. The summary of survey results can be found in Appendix.

6.2.1 RQ1. How do writers use and benefit from MetaphorMap and MetaphorFilter when exploring metaphor ideas? The survey data were analyzed using Wilcoxon signed-rank tests. The average rating on the question of whether it helps metaphor writing was 6.3/7 for MetaphorMap and 5.8/7 for MetaphorFilter. The participants responded that MetaphorMap and MetaphorFilter allowed them to explore metaphors depending on the similarity between the tenor and vehicle (Z = 1.5, p < 0.005) and themes of metaphors (Z = 1.0, p < 0.005), achieving **DG1** (i.e., enabling the semantic search for metaphors). MetaphorFilter enabled participants to explore the metaphors depending on the properties of scientific concepts (Z = 10.0, p < 0.05) and create understandable metaphors (Z = 14.0, p < 0.05). However, there was no indication (Z = 17.0, p = 0.08) that it helped participants create original metaphors, thus, **DG2** (i.e., support the creation of understandable and original metaphors) was only partially achieved.

P7 commented *"it provides a ready list for metaphor ideas, providing inspiration."* P10 said that Metaphorian was like a "*thesaurus for [scientific] metaphors*" as *"it's like a metaphorical dictionary for science concepts."* Participants also highlighted that the originality filter *"helps me add novelty to my writing to make it more compelling"* (P8) and *"gives your an idea about the credibility of the options that you are getting from the system"* (P14).

6.2.2 RQ2. How do writers use and benefit from Metaphorian's MetaphorExtension in expanding the metaphors? The results demonstrated that MetaphorExtension helped expand the metaphors (Z = 0.0, p < 0.0001) with an average rating on the helpfulness of 6.2/7, achieving **DG3** (i.e., enhance the chained extension of extended

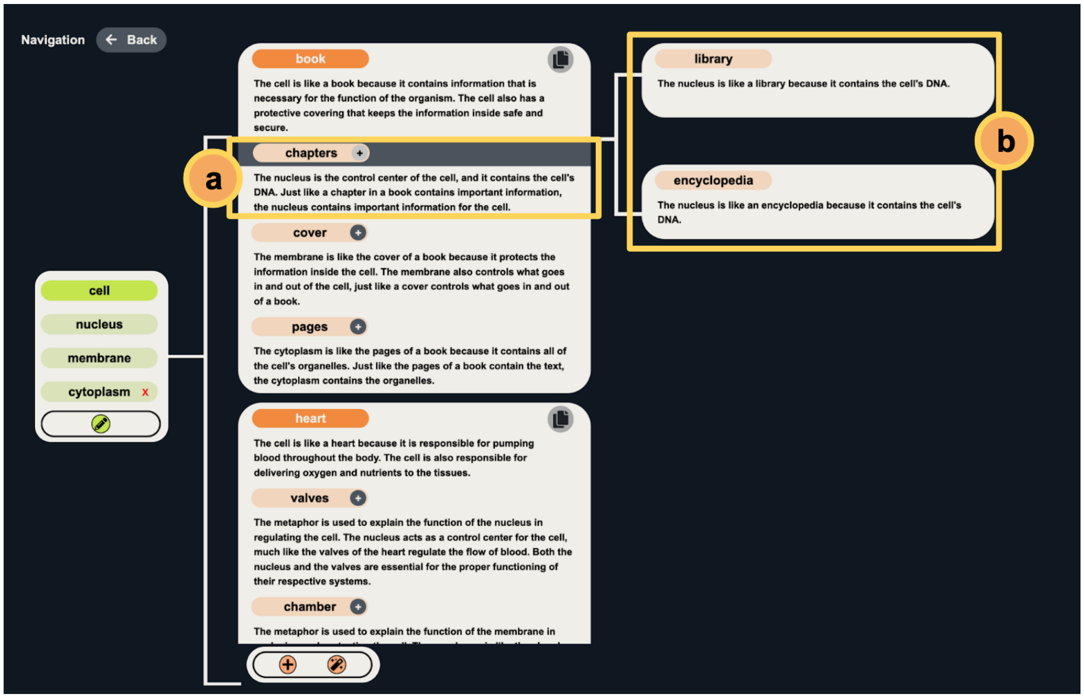

Figure 5: MetaphorBranch supports the iterative revision of metaphors: (a) Chosen Sub-Vehicle: is the sub-vehicle users choose to revise or ideate about, (b) Sub-Branch: displays candidate ideas to replace the chosen sub-vehicle.

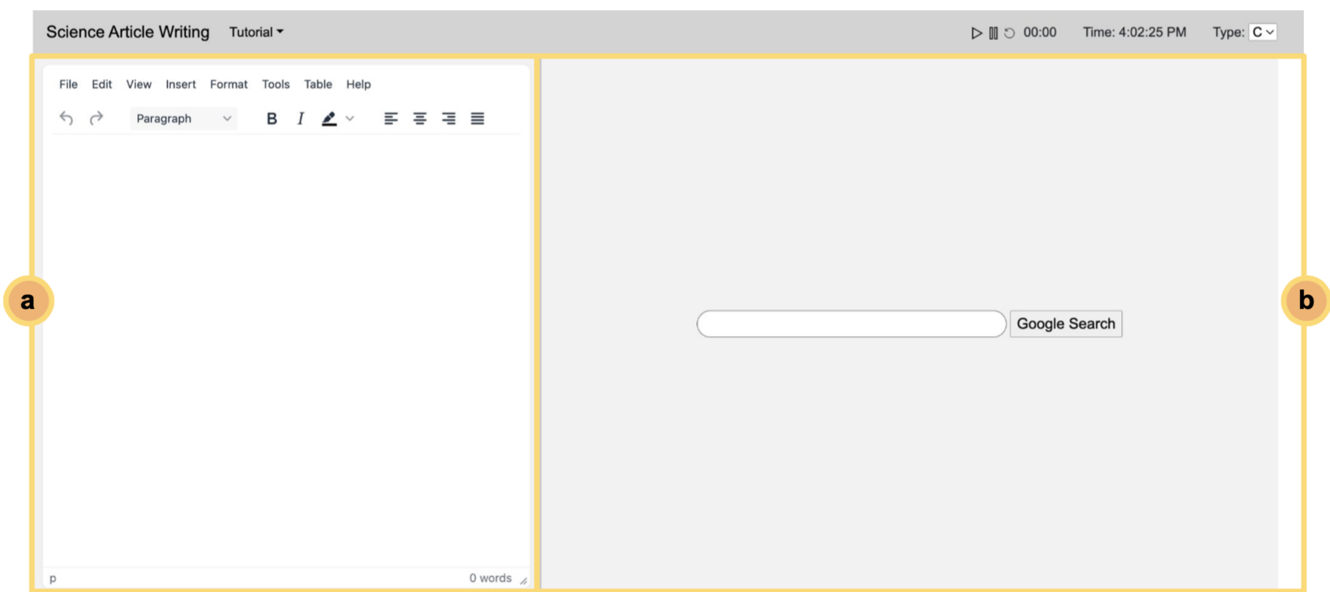

Figure 6: The baseline interface for the user study: (a) the Text Editor and (b) the Google search engine area, where users were free to search anything they want.

metaphors). Participants liked that the sub-metaphors generated by the system were coherent with the main metaphor, i.e., *"it's sometimes hard to relate sub-metaphors to the main one, but the system helped with extending metaphors while sticking to a single theme"* (P13). P18 liked how sub-metaphors were connected: *"it allowed me a step-by-step (metaphor creation) of the basic concept as well as the sub-concepts, and all of them were linked. It was like a logical map where everything fits together."*

6.2.3 RQ3. How do writers use and benefit from MetaphorBranch in iterating sub-metaphor ideas? Participants indicated that MetaphorBranch supported them in modifying sub-metaphors ($Z = 0.0$, $p < 0.0001$) with an average rating on the helpfulness of 6.2/7, achieving **DG4** (e.g., enhance chained iterations of sub-metaphors). They had positive feedback for the MetaphorBranch. P5 highlighted that the revision process is a key step for writers. P7 stated that *"it [MetaphorBranch] was beneficial from an editorial point of view to generate more compelling metaphors."* P1 said that the suggested sub-metaphors were sometimes scientifically inaccurate.

6.2.4 RQ4. How does Metaphorian affect the creative writing experience of scientific metaphors compared to the baseline interface? How the systems supported or did not support, creative writing experiences were analyzed from two aspects, i.e., perceived writing experiences and creativity support.

Perceived Writing Experience. When using Metaphorian, participants were satisfied ($Z = 13.0$, $p < 0.05$) and confident with the originality ($Z = 17.5$, $p < 0.05$) of their metaphors. There were no differences in if the metaphor integrated into their writing ($Z = 15.0$, $p = 0.057$) or confidence when creating scientifically accurate metaphors ($Z = 16.0$, $p = 0.069$). Like the formative study results, several participants mentioned the need for a balance between originality and scientific precision, e.g., *"I wanted to search some original metaphors (for the human heart), but common ones, like 'the human heart is like the engine motor' had more accurate explanations"* (P8). They were also willing to use this system for their daily writing ($Z = 15.0$, $p < 0.01$), e.g., *"Let me know when you are selling it, so I can buy it"* (P11) and *"this system will be a game-changer for science writing. I hope it becomes publicly available soon"* (P5).

Based on the NASA-TLX, the participants reported a significantly lower frustration level ($Z = 7.0$, $p < 0.05$) and temporal demand ($Z = 16.0$, $p < 0.05$) when using our system compared to the baseline. We did not observe a significant difference in mental demand ($Z = 25.5$, $p = 0.16$) and effort ($Z = 30.5$, $p = 0.16$).

Creativity Support. Analyzing the submitted metaphors for the given topics (i.e., cell and DNA), the participants submitted 12 unique vehicles using the baseline and 13 using Metaphorian, suggesting that the system did not lead them to adopt the same ideas and generated diverse metaphors. This could also be because the LLM returned different results every time and participants used their own input and selection strategies. In terms of the recommendation acceptance rates, all participants adopted the metaphors suggested by Metaphorian. Seven out of 16 participants made minor revisions to the suggested metaphors, correcting grammatical errors, while the rest of them made major revisions, such as rewriting sub-metaphors. The participants found the metaphor search results helped them get unstuck while writing ($Z = 11.5$, $p < 0.05$), come up with new ideas ($Z = 9.0$, $p < 0.01$), and inspired their writing ($Z = 11.0$, $p < 0.05$). P6 said, *"it gave me ideas I have never thought of."*

The ratings for the creativity support index [23] showed that Metaphorian provided an immersive writing experience. Participants responded that while doing the activity, the system disappeared ($Z = 1.5$, $p < 0.05$), they were absorbed in this activity ($Z = 6.0$, $p < 0.01$), and became expressive and creative ($Z = 21.0$, $p < 0.05$). For instance, P11 said, *"I find it hard to know I have to stop using this. I want to use it on a daily basis."* What they were able to produce was more worth the effort ($Z = 14.0$, $p < 0.05$), and it was easy to explore many different ideas without tedious interaction ($Z = 15.0$, $p < 0.05$). There was no significant difference in the perceived success for the given task ($Z = 25.5$, $p = 0.15$).

The metaphors the participants made demonstrated Metaphorian's usefulness for various scientific topics in addition to the cell and DNA. The custom topics chosen by the participants include fossils, chemical reactions, evolution, the digestive system, cancer, black hole, tooth, brain, and kidneys. P13 stated that, when using the baseline system, he struggled to make a metaphor for the animal, highlighting the difficulty of creating a sub-metaphor that is consistent with the main one. She was not satisfied with her sub-metaphor, as it had a weak relation to the main metaphor: "*the animal is like a city that contains all components to run normal functions, as the animal body contains different organs. Their heart is like the people in the city that make the city alive.*" On the other hand, using the Metaphorian, he expressed the ease of extending the sub-metaphors. The metaphor he created using our system was, "*the kidney is like a filter because it cleans the blood and removes waste products. The nephron is the functional unit of the kidney. It is responsible for the filtration, reabsorption, and secretion of wastes and ions. The nephron is like a sieve because it filters out wastes and ions from the blood. It demonstrates how the nephron can selectively reabsorb certain molecules while excreting others. The bowman's capsule is like a strainer of the filter because it helps to remove impurities from the water. The bowman's capsule is also like a strainer of the filter because it helps to keep the water clean and clear.*" Table 6 lists more instances of extended metaphors generated by Metaphorian and the final version submitted by participants.

6.2.5 RQ5. How does Metaphorian affect the sense of agency compared to the baseline interface? Participants indicated that they could understand why the metaphors were suggested to them ($Z = 0.0$, $p < 0.005$) more so than the baseline search engine results. However, there was no difference in terms of the sense of agency that they felt (i.e., questions: *"I feel like - I'm the author of the metaphors ($Z = 62.0$, $p = 0.78$), I'm in full control of creating metaphors ($Z = 24.5$, $p = 0.14$), the metaphor is mine ($Z = 68.0$, $p = 1.0$)"*).

During the interviews, opinions on the sense of agency diverged. On one side, participants like P5 stated that *"using the search engine [baseline], I could just reference them [metaphors] as they are, but with the system, I could add sub-concepts I wanted, and the search results are specifically the ones I was looking for. So I felt like I'm the writer of them."* On the other side, P7 explained that it was easy to write metaphors without much effort, making it difficult to feel a sense of agency, highlighting the trade-off between a sense of ownership and easiness while writing. Meanwhile, several participants explained that using this system was like a collaborative process, e.g., *"I kept revising the sub-metaphors, and the creation process was very participatory"* (P8). P9 stated that *"[when using the system] I felt the writing process was collaborative. I didn't feel like they [metaphors] are entirely mine, but I also contributed to it by selecting and revising it."* P12 suggested one way to improve the sense of agency: *"I would use the concepts [vehicles], not the whole sentences. Using the whole sentences makes you feel like you are copying someone else."* At the same time, they commented that

Table 6: The Metaphorian-generated metaphors are the ones that are generated by our system and the participants' final metaphors are the metaphors the participants came up with using the baseline interface. The main tenor and sub-tenors are tenors that are chosen by the participants.

Main tenor *Sub-tenor*	Metaphorian-Generated Metaphor	Participants' Final Metaphor
kidney *nephron, bowman's capsule*	The **kidney** is like a filter because it cleans the blood and removes waste products. The *nephron* is the functional unit of the **kidney**. It is responsible for the filtration, reabsorption, and secretion of wastes and ions. The *nephron* is like a sieve because it filters out wastes and ions from the blood. The sieve analogy is also useful because it demonstrates how the *nphron* can selectively reabsorb certain molecules while excreting others. The *bowman's capsule* is like a strainer of the filter because it helps to remove impurities from the water. The *bowman's capsule* is also like a strainer of the filter because it helps to keep the water clean and clear.	The **kidney** is like a filter because it cleans the blood and removes waste products. The *nephron* is the functional unit of the kidney. It is responsible for the filtration, reabsorption, and secretion of wastes and ions. The *nephron* is like a sieve because it filters out wastes and ions from the blood. It demonstrates how the *nephron* can selectively reabsorb certain molecules while excreting others. The *bowman's capsule* is like a strainer of the filter because it helps to remove impurities from the water. The *bowman's capsule* is also like a strainer of the filter because it helps to keep the water clean and clear.
brain *spinal cord, nerve*	The **brain** is like a computer because it is able to process large amounts of information quickly and efficiently. Like a computer, the brain is able to store large amounts of information in its memory. Additionally, the brain is able to retrieve information from its memory when needed. ... The *spinal cord* is like the hardware of a computer because it is the main highway for communication between the **brain** and the rest of the body. The *spinal cord* is protected by the bones of the spine, just as the hardware of a computer is protected by its casing. The *nerves* are responsible for transmitting information between the **brain** and the rest of the body. Just as the internet helps to transmit information between different computers, the *nerves* play a similar role in the body.	The **brain** is like a computer because it can process large amounts of information quickly and efficiently. Like a computer, the **brain** is able to store large amounts of information in its memory. ... The *spinal cord* is like the hardware of a computer because it is the main highway for communication between the **brain** and the rest of the body. The *spinal cord* is protected by the bones of the spine, just as the hardware of a computer is protected by its casing. The *nerves* are like the internet as they are responsible for transmitting information between the **brain** and the rest of the body, just as the internet helps to transmit information between different computers
evolution *mutation, challenge*	The **evolution** is like a game because it is a competition in which players try to outdo each other. The *mutation* is like a levels of the game because it is a change that occurs in the DNA of an organism that can be passed down to future generations. The *mutation* can be beneficial, neutral, or harmful to the organism. ... The *challenges* of the game, are like the phenotype of an organism because the phenotype is the physical and behavioral traits of an organism that are determined by its genes. The phenotype can be affected by the environment, but the genes are the ultimate determinant of the phenotype.	**Evolution** is like a game because it is a competition in which players try to outdo each other. The *mutation* is like levels of the game because it is a change that occurs in an organism's DNA that can be passed down to future generations. ... The ability to tackle the game's *challenges* is like the phenotype of an organism because the phenotype is the physical and behavioral traits of an organism determined by its genes.

they also could not feel a sense of agency while using the baseline as some of them used the metaphors in the search engine as is.

6.2.6 RQ6. How do perceived qualities of metaphors created with Metaphorian compare to those created with the baseline? Our user study results revealed that study participants generally found the metaphors created with Metaphorian to be more understandable (5.9 vs 4.8 out of 7; Z = -2.2, $p < 0.05$), original (5.4 vs 4.2 out of 7; Z = -1.7, p = 0.08), scientifically accurate (5.6 vs 4.2 out of 7; Z = -2.1, $p < 0.05$), and satisfactory (5.7 vs 4.7 out of 7; Z = -2.3, $p < 0.05$). However, since self-reports are susceptible to bias, we conducted an evaluation study with other expert science writers to validate this. Since Metaphorian is primarily designed to facilitate the exploration and not necessarily replace the writing and edits by the science writers, our hypothesis was that when other science raters evaluate, we would not find significant differences across Metaphorian and baseline conditions (although there were when comparing self-reports).

To have other science writers evaluate these metaphors, we randomly selected one pair of metaphors (1 Metaphorian and 1 baseline) from each participant to create a set of 32 extended metaphors,[3] which comprised extended metaphors from 8 fixed topics & 8 custom topics from the baseline interface and 8 fixed topics & 8 custom topics from Metaphorian. Four science writers (1 male, 3 female; average age: 29, range: 22-34) with an average of 4.5 years of (Std=3.5) experience in publishing scientific articles for the general public were recruited through Upwork [4] and were asked to rate the understandability, originality, scientific accuracy, and overall quality of the 32 extended metaphors. The order of extended metaphors was randomized to minimize the order effect. They received $30 for answering an hour long survey.

Although our sample size is small and we should therefore interpret this with caution, the Wilcoxon Signed-Rank Test did not yield significant differences across conditions on four criteria. This was expected as writers in the user study all modified metaphors created by Metaphorian based on their own preferences, resulting in extended metaphors of similar qualities. The discrepancies between the ratings reported by the participants themselves and the new group of evaluators can be due to the good-participant effect known in experiments that involve human subjects or that assistance from LLMs may have an effect on participants' assessment of the metaphors that they created.

7 DISCUSSION

In this work, we sought to support science writers in creating extended metaphors, which is a notoriously difficult task in that it requires the writers to search for concepts that are original and yet simultaneously conform to strict structural constraints. We leveraged LLMs to address this challenge as LLMs are capable of searching within a large idea space while following the guidance from the users via prompts. To our knowledge, no work has explored the potential of LLMs for supporting the creation of extended metaphors. Our work fills this gap by providing insight into the qualities science writers seek in extended metaphors, the interaction techniques (such as the ability to search, extend, and iterate) that a system should support, and the prompt engineering techniques we can use to generate desired extended metaphors. Below, we discuss various implications of this work and future directions.

7.1 Implication of LLM-based Pipeline

Our LLM-based pipeline consists of two techniques: prompt chaining and requirement concatenation. Prompt chaining helped break down the task of creating extended metaphors into manageable steps, improved output, reduced latency, and allowed users to get involved in each step of the extended metaphor creation. Requirements in the prompt provided a guardrail, restraining the LLM from, e.g., generating scientifically inaccurate metaphors. While the techniques themselves are not novel, the combination of these two techniques may be useful for tackling problems that require multiple steps and adherence to certain constraints. For example, designers often seek inspiration for product ideation; they search for products with set attributes (e.g., 'can comfortably carry with one hand'); the specific attributes here are the constraints in their search for products with such attributes (e.g., 'cup', 'pencil') [37]. The simplicity of our approach may allow it to be a widely applicable technique for content generation in other domains.

7.2 Implications Beyond Science Writing

Our work focused on the use of extended metaphors in science writing, which can facilitate the explanation and communication of complex scientific concepts to the general public. Scientific metaphors offer a great learning opportunity, helping readers to easily understand abstract science concepts [26] and remember information better [38]. Popular science writings in the form of science communication books and videos can reach an audience in the size of millions and even billions [78]. Beyond science writing, many other literary genres of writing, such as poetry, novel, and allegory, depend heavily on extended metaphors [28, 75, 81], for which the Metaphorian system can be useful for. As discussed in the previous section, our approach can be applied to other types of writing, such as argumentative writing or narrative writing, in which the writings need to conform to domain-specific constraints and practices. These identified constraints can be embedded within prompts to guide generative models to produce content that fulfills the specified requirements.

7.3 Improving LLMs for Real-Time Application

LLMs [16] can be adapted to a wide range of downstream tasks, but they sometimes entail an under-constrained problem [53]; since they are not trained for a specific task, they sometimes return too general outputs. That is because they are not given enough constraints to tailor their learning to the target tasks. According to the results of the exploratory LLM study, the model's improved performance via requirements concatenation implies that the under-constrained problem can be alleviated by including tasks' requirements as constraints in the prompt design, even for open-ended tasks like creative writing.

On the other hand, the LLM's inference time was impractically long to support real-time interactions in generating diverse ideas, reaching up to a minute for generating 20 vehicles. Meanwhile, the prompt chaining reduced the generation time to an acceptable level of fewer than five seconds by splitting the task into several sub-tasks (e.g., vehicle generation, mapping generation) and then gradually revealing the results of each stage according to users' selection (e.g., our system generates full-sentence metaphors only for the vehicles chosen by users). System designers should thus consider adopting prompt chaining for real-time applications powered by LLMs due to their time efficiency, as well as the likelihood of it returning the correct output.

7.4 Staying Up-to-Date with Language Models

The study was carried out in September 2022, prior to the release of advanced language models such as OpenAI's ChatGPT, GPT-4, and Google Bard. Therefore, the results of the study do not accurately reflect the capabilities of the state-of-the-art language models. However, an informal evaluation of ChatGPT shows that although it exhibits improved writing quality, the metaphors generated still suffer from the issues highlighted in the study and fall short of human-generated ones. With the rapid advancement and frequent

[3]The dataset is included in our Supplementary Materials.

updates of language models, such as ChatGPT being updated on a monthly basis, it becomes challenging to stay up-to-date with the latest models. As a result, rerunning the study was deemed impractical, as the results will very likely become outdated when presented to the audience. We encourage the audience to focus on our human-in-the-loop exploration approach, which can be applied to leverage large generative models in assisting tasks in other domains. In addition, despite these rapid advancements in LLMs, the value of our research remains intact. Specifically, the current language models struggle with maintaining internal structure and consistency while generating content [31]. In contrast, our prompt chaining technique facilitates the creation of extended metaphors while preserving their congruent structure.

7.5 Agency & Ownership in Co-Creation with AI

Previous research on support for creative writing has indicated that the homogeneity of machine-generated output may result in restricted diversity and produce similar writing styles [22, 59, 61]. However, Metaphorian addresses this issue through its MetaphorMap feature, which offers a range of options, and its MetaphorExtension and MetaphorBranch features, which enable iterative modifications. Our user study findings confirm the effectiveness of this approach, as all participants could generate unique and diverse extended metaphors for the same scientific concepts, indicating that the Metaphorian approach successfully prevents homogeneity in writing.

Although most participants agreed that the system provides an inspirational and immersive writing experience that supports the creation of original and understandable metaphors, their opinions on the ownership of the metaphors were split. Some participants felt a sense of agency as they could edit the suggested metaphor and decide what to include in the final piece. Others did not feel as if they were the author of the metaphors since the system suggested the metaphors in a fluent and polished way and they often integrated them into their writing as is. Those participants also expressed concerns about plagiarism when copying and pasting the suggested metaphors into their writing. Plagiarism is one of the most significant challenges posed by using generative models as writing support [1, 50, 79]. MetaphorMap currently gauges the novelty of each metaphor by leveraging the available metaphor dataset. However, future research can delve deeper by providing references and context of the existing metaphors in the dataset, enabling users to steer clear of plagiarism. Others, on the other hand, perceived the writing process as a collaborative process with the system, where they could receive the assistance and guidance they needed. Thus, the system designers can consider the diverse cognitive models of writers when designing the writing support tool powered by generative LLMs.

8 FUTURE WORK

We present several potential avenues for future research.

8.1 Studying Metaphorian in the Wild

One limitation of our study is the user study setting, namely the small sample size and the writing task not representative of their real-world experience. Our user study was conducted in a controlled setting, giving participants only 5-7 minutes to write a short paragraph, even though they likely write multiple paragraphs in a span of a few hours to a few days for their science articles. The fact that they are professional writers with many years of experience suggests that their assessments may not be far from what they reported, but since the amount of time they spent with the tool was limited, a longitudinal study would help reveal the full extent of Metaphorian's usefulness.

We recruited professional science writers for our user study because they can not only evaluate Metaphorian as potential users, but also provide insights into how readers might perceive the writing and extended metaphors generated by the tool. Given their extensive experience (average of 8 years) in publishing content for a general audience, they are well-suited to assess these aspects to a certain extent. However, given the limitations of this approach, it would be valuable to conduct a study with actual readers to explore questions such as whether they find articles written with Metaphorian (or LLM) more engaging than those written without it, and how they compare in terms of readability, accuracy, and writing time, among other factors

8.2 Enhancing Human in the Loop

While our prompt augmentation and chaining improved the quality of generated extended metaphors and user study demonstrated that Metaphorian successfully supports science writers with the creation of scientific metaphors, it is worth noting that LLM-generated metaphors still underperformed human-made metaphors in our exploratory study of LLM. This aligns with the results of our user study. While most participants found metaphors in Metaphorian novel and inspirational, some found them unoriginal. Further, when asked whether they have any suggestions, several participants suggested adding human-made metaphors into the database so that the metaphors generated by Metaphorian can be more novel. A related idea was embedding a mechanism that would allow Metaphorian to collect users' novel metaphors. For instance, the interface can have an input field for users to suggest metaphors if none of the generated metaphors are satisfactory.

On the other hand, users' control over the system and their activities is a key design consideration for human-AI collaboration [12] and creativity support tools [23]. Accordingly, future work can allow users to choose the level of AI involvement they wish to receive. For instance, a user can input both tenors and vehicles and ask AI to generate mappings that connect the tenors and input vehicles. They can also input partially complete extended metaphors and ask for assistance to complete the rest of the part.

8.3 Ensuring Scientific Accuracy in Metaphors

In scientific writing, scientific truthfulness is crucial as the use of metaphors can dilute the accuracy of scientific concepts by introducing secondary concepts (i.e., vehicles) to explain them. While this may not pose a significant issue for experienced science writers like Metaphorian's target user group, who are able to discern whether metaphors contain unscientific content, it is crucial that future writing support tools designed for novice writers or educational purposes take into account the potential pitfalls of metaphorical language and exercise greater caution. With advances in research, we expect future writing support tools to be able to produce more scientifically accurate content.

9 CONCLUSION

Metaphors are useful tools to science writers, as they convey complex scientific concepts in a relatable way. However, creating extended metaphors for science writing is challenging since it requires writers to ideate novel vehicles with coherent structures. We explored LLMs' ability in generating extended metaphors for science writing and found current LLMs do not produce extended metaphors that are up to professional writers' standards in terms of originality, fluency, or metaphoricity. We proposed leveraging prompt chaining and augmentation to break down the generation tasks and embed the structural requirement of extended metaphors within the prompts to improve LLM's ability in generating extended metaphors. Based on this LLM-based pipeline, we developed Metaphorian, a human-in-the-loop interactive system that supports the iterative search, extension, and revision of extended metaphors. The evaluation of Metaphorian using quantitative and qualitative measures demonstrated the effectiveness of the system for metaphor writing support, where participants found Metaphorian enabled them to flexibly explore and iterate extended metaphors while ensuring the desired coherence and increasing their confidence and inspiration in figurative writing.

REFERENCES

[1] [n. d.]. From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. https://doi.org/10.1145/3532106.3533506

[2] 2006. Topics. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Science/Topics

[3] 2012. Teach computer science. https://studio.code.org/courses?view=teacher

[4] 2015. Upwork. https://upwork.com/

[5] 2022. Communication. http://www.metamia.com/

[6] 2022. Copysmith. https://copysmith.ai/

[7] 2022. Elicit. https://elicit.org/

[8] 2022. Form submission: Topic arrangements of the NGSS. https://www.nextgenscience.org/overview-topics

[9] 2022. Jasper. https://www.jasper.ai/

[10] Keiga Abe, Kayo Sakamoto, and Masanori Nakagawa. 2006. A computational model of the metaphor generation process. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 28.

[11] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective interfaces for student-driven revision sessions for argumentative writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3411764.3445683

[12] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13. https://doi.org/doi/10.1145/3290605.3300233

[13] Ken Baake. 2012. *Metaphor and knowledge: The challenges of writing science*. SUNY Press.

[14] Saptarashmi Bandyopadhyay, Jason Xu, Neel Pawar, and David Touretzky. 2022. Interactive Visualizations of Word Embeddings for K-12 Students. In *EAAI-22: The 12th Symposium on Educational Advances in Artificial Intelligence*. https://doi.org/10.1609/aaai.v36i11.21548

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[16] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021). https://doi.org/10.48550/arXiv.2108.07258

[17] Hunter M Breland. 1996. Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science* 7, 2 (1996), 96–99.

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

[19] Robyn Carston. 2010. XIII—Metaphor: Ad hoc concepts, literal meaning and mental images. In *Proceedings of the Aristotelian society*, Vol. 110. Oxford University Press Oxford, UK, 295–321. https://doi.org/10.1111/j.1467-9264.2010.00288.x

[20] Susan Carter and Rod Pitcher. 2010. Extended metaphors for pedagogy: using sameness and difference. *Teaching in higher education* 15, 5 (2010), 579–589. https://doi.org/10.1080/13562517.2010.491904

[21] Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. *arXiv preprint arXiv:2103.06779* (2021). https://doi.org/10.48550/arXiv.2103.06779

[22] Liuqing Chen, Lingyun Sun, and Ji Han. 2023. A Comparison Study of Human and Machine-Generated Creativity. *Journal of Computing and Information Science in Engineering* 23, 5 (2023), 051012. https://doi.org/10.1115/1.4062232

[23] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25. https://doi.org/doi/10.1145/2617588

[24] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Designing Interactive Systems Conference 2021*. 1817–1833. https://doi.org/10.1145/3461778.3462050

[25] Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience* 26, 11 (2014), 2585–2595. https://doi.org/10.1162/jocn_a_00654

[26] Richard K Coll, Bev France, and Ian Taylor. 2005. The role of models/and analogies in science education: implications from research. *International Journal of Science Education* 27, 2 (2005), 183–198. https://doi.org/10.1080/0950069042000276712

[27] Charles R Cooper and Lee Odell. 1977. Evaluating writing: Describing, measuring, judging. (1977).

[28] Peter Crisp. 2008. Between extended metaphor and allegory: is blending enough? *Language and Literature* 17, 4 (2008), 291–308. https://doi.org/10.1177/0963947008095960

[29] Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 1173–1178. https://doi.org/10.18653/v1/D19-1109

[30] William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification* 1, 1 (1984), 7–24. https://doi.org/10.1007/BF01890115

[31] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics* 9 (2021), 1012–1031. https://doi.org/10.1162/tacl_a_00410

[32] Ian Fillis and Ruth Rentschler. 2005. *Creative marketing: An extended metaphor for marketing in a new age*. Springer.

[33] Andrea Gagliano, Emily Paul, Kyle Booten, and Marti A Hearst. 2016. Intersecting Word Vectors to Take Figurative Language to New Heights. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. 20–31.

[34] Katy Gero and Lydia Chilton. 2018. Challenges in finding metaphorical connections. In *Proceedings of the Workshop on Figurative Language Processing*. 1–6. https://doi.org/10.18653/v1/W18-0901

[35] 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. https://doi.org/doi/10.1145/3290605.3300526

[36] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*. 1002–1019. https://doi.org/doi/abs/10.1145/3532106.3533533

[37] Karni Gilon, Joel Chan, Felicia Y. Ng, Hila Liifshitz-Assaf, Aniket Kittur, and Dafna Shahaf. 2018. Analogy Mining for Specific Design Needs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 121, 11 pages. https://doi.org/10.1145/3173574.3173695

[38] Shawn M Glynn and Tomone Takahashi. 1998. Learning from analogy-enhanced science text. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 35, 10 (1998), 1129–1149. https://doi.org/10.1002/(SICI)1098-2736(199812)35:10<1129::AID-TEA5>3.0.CO;2-2

[39] Steven M Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N Horne, Michal Lahav, Robert Macdonald, Rain Breaw Michaels, Ajit Narayanan, et al. 2022. LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia. *arXiv preprint arXiv:2207.02308* (2022). https://doi.org/10.1145/3517428.3544819

[40] Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert. *Internet]. Available: https://maartengr. github. io/KeyBERT/index. html* (2020).

[41] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020). https://doi.org/10.48550/arXiv.2006.03654

[42] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331. https://doi.org/10.1177/1525822X04266540

[43] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438. https://doi.org/10.1162/tacl_a_00324

[44] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[45] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3502030

[46] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021). https://doi.org/10.1145/3560815

[47] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704* (2021). https://doi.org/10.48550/arXiv.2104.08704

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019). https://doi.org/10.48550/arXiv.1907.11692

[49] LuminosoInsight. [n. d.]. LuminosoInsight/exquisite-corpus: Put together a multilingual corpus from a variety of sources. used for wordfreq and word embeddings. https://github.com/LuminosoInsight/exquisite-corpus

[50] Meredith Ringel Morris, Carrie J Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. 2023. The design space of generative models. *arXiv preprint arXiv:2304.10547* (2023). https://doi.org/10.48550/arXiv.2304.10547

[51] Anita Naciscione. 2016. Extended metaphor in the web of discourse. *Mixing Metaphor. Amsterdam/Philadelphia: John Benjamins Publishing Company* (2016), 241–266.

[52] Kai Niebert, Sabine Marsch, and David F Treagust. 2012. Understanding needs embodiment: A theory-guided reanalysis of the role of metaphors and analogies in understanding science. *Science Education* 96, 5 (2012), 849–877. https://doi.org/10.1002/sce.21026

[53] Balder Onarheim and Michael Mose Biskjaer. 2015. Balancing constraints and the sweet spot as coming topics for creativity research. *Creativity in design: Understanding, capturing, supporting* 1 (2015), 1–18.

[54] Steve Oswald and Alain Rihs. 2014. Metaphor as argument: Rhetorical and epistemic advantages of extended metaphors. *Argumentation* 28, 2 (2014), 133–159. https://doi.org/10.1007/s10503-013-9304-0

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

[56] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019). https://doi.org/10.48550/arXiv.1908.10084

[57] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating creativity support tools in HCI research. In *Proceedings of the 2020 ACM designing interactive systems conference*. 457–476. https://doi.org/10.1145/3357236.3395474

[58] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7. https://doi.org/doi/10.1145/3411763.3451760

[59] Melissa Roemmele. 2021. Inspiration through observation: Demonstrating the influence of automatically generated text on creative writing. *arXiv preprint arXiv:2107.04007* (2021). https://doi.org/10.48550/arXiv.2107.04007

[60] Paula Rubio-Fernandez, Chris Cummins, and Ye Tian. 2016. Are single and extended metaphors processed differently? A test of two relevance-theoretic accounts. *Journal of Pragmatics* 94 (2016), 15–28.

[61] Murray Shanahan. 2022. Talking About Large Language Models. *arXiv preprint arXiv:2212.03551* (2022). https://doi.org/10.48550/arXiv.2212.03551

[62] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2. https://doi.org/10.5281/zenodo.1443582

[63] Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. Vol. 14. John Benjamins Publishing.

[64] Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych. 2020. Metaphoric paraphrase generation. *arXiv preprint arXiv:2002.12854* (2020). https://doi.org/10.48550/arXiv.2002.12854

[65] Sangho Suh, Jian Zhao, and Edith Law. 2022. CodeToon: Story Ideation, Auto Comic Generation, and Structure Mapping for Code-Driven Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16. https://doi.org/10.1145/3526113.3545617

[66] Jose M Tamayo. 1987. Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational and Psychological Measurement* 47, 4 (1987), 893–902.

[67] Asuka Terai and Masanori Nakagawa. 2010. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*. Springer, 142–147. https://escholarship.org/uc/item/5d96219g

[68] Paul H Thibodeau. 2016. Extended metaphors are the home runs of persuasion: Don't fumble the phrase. *Metaphor and Symbol* 31, 2 (2016), 53–72.

[69] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 248–258.

[70] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 680–690.

[71] Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: the autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573* (2020). https://doi.org/10.48550/arXiv.2010.10573

[72] Tony Veale. [n. d.]. Thesaurus Rex. http://ngrams.ucd.ie/therex3/. Accessed: 2022-07-01.

[73] Tony Veale and Yanfen Hao. 2007. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *AAAI*, Vol. 2007. 1471–1476.

[74] Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies* 9, 1 (2016), 1–160. https://doi.org/10.1162/COLI_r_00311

[75] Paul Werth. 1994. Extended metaphor—A text-world account. *Language and literature* 3, 2 (1994), 79–103.

[76] Diana Williams-Whitney, Jeffery Scott Mio, and Paul Whitney. 1992. Metaphor production in creative writing. *Journal of Psycholinguistic Research* 21, 6 (1992), 497–509. https://doi.org/10.1007/BF01067527

[77] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*. 1–22. https://doi.org/10.1145/3491102.3517582

[78] Haijun Xia, Hui Xin Ng, Zhutian Chen, and James Hollan. 2022. Millions and Billions of Views: Understanding Popular Science and Knowledge Communication on Video-Sharing Platforms. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 163–174. https://doi.org/10.1145/3491140.3528279

[79] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300415

[80] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[81] Shahriyar Yazdanjoo and Hadiseh Fallahpour. 2018. A study on the correlation between creative thinking of Iranian EFL learners and using metaphor in descriptive writing tasks. *International Journal of English Language and Literature Studies* 7, 2 (2018), 32–44. https://doi.org/10.18488/journal.23.2018.72.32.44

[82] Lixiu Yu, Aniket Kittur, and Robert E Kraut. 2014. Searching for analogical ideas with crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1225–1234. https://doi.org/doi/10.1145/2556288.2557378

[83] Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 861–871. https://doi.org/10.18653/v1/N19-1092

[84] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*. 841–852. https://doi.org/doi/abs/10.1145/3490099.3511105

[85] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 12697–12706.

A APPENDIX

Table 7: An example prompt design of prompt chaining in the 0-shot and 10-shot settings. The output of the LLM is used as a new input in the next run. In the 10-shot setting, professional writers’ quality examples are included as examples to guide the model. Note that “... <9 example vehicles> ...” and “... <9 example mappings> ...” are not part of the output. It indicates that 9 example vehicles or mappings were placed there. They are not included in the table due to space constraints but can be found in Supplementary Material.

Sub-task	0-shot	10-shot
Main Vehicle Generation	Write a metaphor for the {main tenor}. Metaphor: The {main tenor} is like	Write a metaphor for the {main tenor}. Metaphor: The cell is like a city. ··· <9 example vehicles> ··· Metaphor: The {main tenor} is like
Main Mapping Generation	Write reasons for the metaphor. Metaphor: {main tenor} is like the {main vehicle}. Reason:	Write a metaphor for the {main tenor}. Metaphor: The cell is like a city. ··· <9 example vehicles> ··· Metaphor: The {main tenor} is like
Sub-Vehicle Generation	Given the {main metaphor}, write sub-metaphors for each subconcept of the {main tenor}. Main metaphor: the {main tenor} is like the {main vehicle}. Subconcepts of the {main tenor}: {sub-tenors} Sub-metaphors:	Given the main metaphor, write sub-metaphors for each subconcept of the {main tenor}. Main metaphor: cell is like a city. Subconcepts of the cell: the membrane, the nucleus, the cytoplasm Sub-metaphors: The membrane is like the city walls. The nucleus is like the city center. The cytoplasm is like the city streets. ··· <9 example vehicles> ··· Main metaphor: the {main tenor} is like the {main vehicle}. Subconcepts of the {main tenor}: sub-tenors Sub-metaphors:
Sub-Mapping Generation	Given the main metaphor, reasons for main metaphor, and the sub-metaphors, write reasons for the sub-metaphors. Main metaphor: the {main tenor} is like the {main vehicle}. Reasons for main metaphor: {main mapping}. Sub-metaphors: {sub-vehicles} Reasons for sub-metaphor:	Given the main metaphor, reasons for main metaphor, and the sub-metaphors, write reasons for the sub-metaphors. Main metaphor: cell is like a city. Reasons for main metaphor: cell is like a city in that it has a complex system of cells that work together to perform specific functions. Sub-metaphors: The membrane is like the city walls. The nucleus is like the city center. The cytoplasm is like the city streets. Reasons for sub-metaphor: The membrane is like the city walls in that it protects the cell from outside invaders. The nucleus is like the city center in that it contains the cell’s DNA and controls the cell’s activities. The cytoplasm is like the city streets in that it contains the cell’s organelles and allows them to move around. ··· <9 example mappings> ··· Main metaphor: the main tenor is like the {main vehicle}. Reasons for main metaphor: {main mapping}. Sub-metaphors: {sub-vehicles} Reasons for sub-metaphor:

Table 8: The survey results of the user study (n=16), along with the questions, p-values, and mappings to the research questions. The questionnaire consists of six main sections: design goals, inspiration, creativity support index, writing experience, sense of agency, and cognitive demand. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

Category	Survey Question	Median (Std) Baseline	Median (Std) Metaphorian	statistic value (z-score)	p-value
RQ1 (DG1)	Did it allow you to explore the metaphors depending on the similarity between the scientific concept and the metaphor?	3.0 (2.1)	6.5 (0.8)	1.5	**<0.01** **
RQ1 (DG1)	Did it allow you to explore the metaphors depending on the properties of the scientific concept?	4.0 (2.0)	6.5 (1.1)	10.0	**<0.01** **
RQ1 (DG1)	Did it allow you to explore the metaphors depending on the theme of the metaphors?	3.0 (2.0)	7.0 (0.8)	1.0	**<0.05** *
RQ1 (DG2)	Did it help create original metaphors?	4.0 (1.8)	5.5 (1.5)	17.0	>0.05
RQ1 (DG2)	Did it help create understandable metaphors?	5.5 (1.7)	6.0 (1.3)	14.0	**<0.05** *
RQ2 (DG3)	Did it help you expand the extended metaphors?	3.0 (1.6)	7.0 (0.9)	0.0	**<0.0001** ****
RQ3 (DG4)	Did it help you modify the extended sub-metaphors?	2.0 (2.0)	7.0 (0.9)	0.0	**<0.001** ***
RQ4 (Inpiration)	The metaphor search results helped me get unstuck while writing.	4.0 (2.1)	7.0 (0.9)	11.5	**<0.01** **
RQ4 (Inpiration)	The metaphor search results inspire my writing.	4.5 (1.9)	6.0 (1.2)	11.0	**<0.05** *
RQ4 (Inpiration)	The metaphor search results helped me come up with new ideas.	4.5 (2.1)	7.0 (0.8)	9.0	**<0.01** **
RQ4 (CSI)	While I was doing the activity, the system "disappeared," and I was able to concentrate on the activity.	4.0 (1.7)	5.0 (1.8)	1.5	**<0.05** *
RQ4 (CSI)	I was very absorbed/engaged in this activity - I enjoyed it and would do it again.	5.0 (1.9)	7.0 (1.0)	6.0	**<0.01** **
RQ4 (CSI)	I was able to be very expressive and creative while doing the activity.	5.0 (1.9)	6.0 (1.3)	21.0	**<0.05** *
RQ4 (CSI)	What I was able to produce was worth the effort required to produce it.	4.0 (1.6)	6.0 (1.3)	14.0	**<0.05** *
RQ4 (CSI)	It was easy for me to explore many different options and ideas without a tedious, repetitive interaction.	3.5 (2.2)	7.0 (1.6)	15.0	**<0.05** *
RQ4 (CSI)	How successful were you in accomplishing what you were asked to do?	5.5 (1.7)	6.0 (1.4)	25.5	>0.05
RQ4 (Writing Exp)	I'm willing to use this system for metaphor creation for my article in practice.	4.5 (1.8)	7.0 (1.3)	15.0	**<0.01** **
RQ4 (Writing Exp)	I'm satisfied with the metaphors I created.	5.0 (1.6)	6.0 (1.6)	13.0	**<0.05** *
RQ4 (Writing Exp)	I feel confident that the metaphors I created are coherent with the writing context.	5.0 (1.7)	6.0 (1.3)	15.0	>0.05
RQ4 (Writing Exp)	I feel confident that the metaphors I created are original.	4.5 (2.1)	6.0 (1.5)	17.5	**<0.05** *
RQ4 (Writing Exp)	I feel confident that the metaphors I created are scientifically accurate.	5.0 (1.7)	6.0 (1.1)	16.0	>0.05
RQ5 (Sense of Agency)	I feel like I'm the author of the metaphors.	5.0 (2.2)	5.5 (2.4)	62.0	>0.05
RQ5 (Sense of Agency)	I feel like I'm in full control of creating metaphors.	4.5 (2.2)	5.5 (1.8)	24.5	>0.05
RQ5 (Sense of Agency)	I could understand why these metaphors were suggested to me.	4.0 (1.8)	7.0 (0.9)	0.0	**<0.01** **
RQ5 (Sense of Agency)	I feel like the metaphor is mine.	4.5 (2.1)	5.0 (2.2)	68.0	>0.05