Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity

Zijian Ding
College of Information Studies, University of Maryland
USA

Arvind Srinivasan
College of Information Studies, University of Maryland
USA

Stephen MacNeil
Computer and Information Sciences, Temple University
USA

Joel Chan
College of Information Studies, University of Maryland
USA

ABSTRACT

Cross-domain analogical reasoning is a core creative ability that can be challenging for humans. Recent work has shown some proofs-of-concept of Large language Models' (LLMs) ability to generate cross-domain analogies. However, the reliability and potential usefulness of this capacity for augmenting human creative work has received little systematic exploration. In this paper, we systematically explore LLMs capacity to augment cross-domain analogical reasoning. Across three studies, we found: 1) LLM-generated cross-domain analogies were frequently judged as helpful in the context of a problem reformulation task (median 4 out of 5 helpfulness rating), and frequently (~80% of cases) led to observable changes in problem formulations, and 2) there was an upper bound of ~25% of outputs being rated as potentially harmful, with a majority due to potentially upsetting content, rather than biased or toxic content. These results demonstrate the potential utility — and risks — of LLMs for augmenting cross-domain analogical creativity.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Large Language Models, Analogy, Creativity Support Tools

ACM Reference Format:
Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. In *Creativity and Cognition (C&C '23), June 19–21, 2023, Virtual Event, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3591196.3593516

1 INTRODUCTION

Large language models (LLMs) such as GPT-3 are attracting attention in the creativity research community for their potential to augment creative work by generating tailored design materials and prototypes. For example, researchers have explored the potential of LLMs to replicate or assist creative writing tasks such as metaphor generation [12, 38], science writing [39] and storytelling [3, 11, 22, 32, 75, 81, 86, 94]. This research thread parallels explorations in industry, such as the "AI Dungeon" startup's use of GPT-3 to assist with world-building in Dungeons and Dragons [87], prototype systems for AI-assisted content creation[1] and writing assistance[2], as well as other applications outside of creative writing, such as Q&A conversation[3] and code generation[4]. Here, we are particularly interested in probing the potential of LLMs to augment **cross-domain analogical creativity**.

Cross-domain analogical reasoning is the core cognitive ability to perceive and reason about deep structural similarities between situations that may differ on many surface details [34, 49, 50]; for example, using analogy, people can recognize a strong similarity between the solar system and the atom in terms of relational similarity (both involve a central mass — sun, nucleus — orbited by smaller bodies — planets, electrons), ignoring other surface dissimilarities such as their relative size and color. Cross-domain analogy is a frequent source of creative breakthroughs, enabling creators to develop powerful new concepts [16, 24, 35, 47, 50] or reformulations of their creative problems [4, 27, 45, 48, 73]. As an illustrative example, Dorst [27] reported a case study of how designers used an analogy to a music festival to (re)frame the problem of night violence in King's Cross from a crime prevention / reduction problem with a new entertainment perspective.

We are motivated by the potential of LLMs to augment creative analogy-making because cross-domain analogies can be hard for humans to retrieve in the creative process. Human retrieval is highly sensitive to surface similarities, favoring "near", or within-domain, analogs that share attributes of an object over "far", or structurally similar analogs from different domains that primarily share relations to the object [36, 37, 41, 50]. This can lead to creators failing to retrieve relevant analogs from other domains because they are fixated on surface features of their source problem [62, 63]. For example, people trying to solve Duncker's [31] radiation problem —- how to remove a cancerous tumor using radiation without damaging surrounding healthy tissue — are much more likely to retrieve analogs involving cancer or radiation than

This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

C&C '23, June 19–21, 2023, Virtual Event, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0180-1/23/06.
https://doi.org/10.1145/3591196.3593516

[1] https://www.regie.ai/, https://www.copy.ai/, https://writer.com/
[2] https://www.compose.ai/
[3] https://chat.openai.com/chat
[4] https://github.com/features/copilot

an analog of an army splitting up to attack a target. The ability of LLMs to dynamically generate elaborate text tuned to specific settings presents an opportunity to complement existing structured methods of supporting creative analogy-making [44], such as the WordTree method [61] or the TRIZ method [82], which are manual and effortful, and search-based systems that may require expensive, specialized pre-processing of databases of potential analogs [13, 33, 51, 54, 77, 78, 88].

Past work suggests that LLMs may be able to generate analogies that resemble human-generated analogies. For example, there are substantial systematic investigations of LLM performance on standard analogy completion word problems such as the SAT four-term analogy problems; an example of this task is to generate a term that answers a question like "man is to woman as king is to ...?" (where a correct answer is "queen") [5, 9, 90]. More recently, some researchers have shared proof-of-concept demonstrations of LLMs' ability to generate longer natural-language analogies [87], explanations of analogical mappings [5, 90], or analogy-inspired concepts for creative problems [90, 95, 96].

However, with some exceptions [96], prior work lacks systematic and direct investigation of how these analogies might be *useful* in the creative process. We want to answer questions like, can people use LLM-generated analogies, even if they may be "incorrect" (e.g., missing key relational mappings between a source problem and target analogy), to inspire problem reformulation or ideation? If so, to what extent does this happen? Can we predict in advance which LLM-generated analogies might turn out to be useful? Given concerns about potential bias/harm in LLMs [2, 8], how might LLM-generated analogies' potential usefulness for augmented creativity trade off with potential harmfulness or toxicity in LLM outputs?

These questions cannot be satisfactorily answered in a traditional NLP-oriented benchmark paradigm. For example it is common to compare outputs with a "correct" reference output, as in the analogy completion word problems; this is insufficient to capture the potential usefulness of analogical inspirations that may only emerge from usage, or run through the generating and iterating on "bad" ideas [17, 26, 46, 56]. Crowdsourced human evaluations of LLM outputs may also index only surface-level linguistic coherence vs. more substantive dimensions of quality without more specific (task-specific) instructions: for example, Clark et al [21] reported that crowd workers mostly relied on form vs. content heuristics to make their judgments about human-likeness of LLM-generated text; more specific instructions or task framings — such as contextualizing ratings within a creative task — may be necessary to move beyond surface judgments.

In this paper, we directly investigate the potential use of LLM-generated analogies in the creative task of design problem reformulation. First, following a *prompt-based learning* paradigm [64] (also called in-context learning in machine learning research [9]), we crafted a structured natural-language instruction prompt for generating cross-domain analogies (including, in some cases, one or more examples of cross-domain analogies). We used these prompts to generate 480 cross-domain analogies across six design problems, and systematically explored human judgments of predicted usefulness of LLM-generated analogies (i.e., not in the context of using analogies in a creative task), along with duplicate rate and semantic distance of analogies from the design problems (**prompt engineering explorations**; Section 3). Second, we collected user ratings of helpfulness of a subset of these analogies in the context of a problem reformulation task, along with content analysis of user reformulation behaviors with the analogies (**Study 1**; Section 4). Third and finally, we manually analyzed the nature and rates of potentially harmful outputs in the analogy generations (**Study 2**; Section 5).

Our primary findings were as follows:

(1) **Prompt engineering explorations:** We were able to construct one- or few-shot prompts that yielded cross-domain analogies where 70% of outputs were both unique and judged by researcher to be potentially useful for creative problem reformulation.
(2) **Study 1:** A majority of LLM-generated analogies were rated as helpful for individuals during a creative problem reformulation task, primarily spurring new considerations for the design problems, but also encouraging shifts in problem perspectives and redefining key elements in the original problem statements. Notably, there was no correlation between *a priori* judgments of potential usefulness from prompt engineering explorations and the use-time ratings of analogy helpfulness here.
(3) **Study 2:** There was some evidence of harmful or biased/toxic outputs in generated analogies (upper bound of 25% of outputs as screened by human raters), though the clear majority of potentially harmful outputs (~80%) were describing situations that could be conservatively considered upsetting under some circumstances (e.g., describing situations of poverty or difficult childbirth), rather than biased or toxic.

Overall, our findings suggest that LLM-generated analogies hold potential as a creativity support tool for cross-domain analogical problem reformulation, and extend previous demonstrations of LLMs' capabilities with analogical reasoning for creative settings. To facilitate further analyses by the community, we also share the code used to generate the datasets in prompt engineering explorations (along with the human judgment data), and raw and coded participant responses from Study 1. We hope that the rich descriptive data provided in this paper can help researchers of creativity support tools understand how to effectively leverage LLMs to augment cross-domain analogical creativity.

2 RELATED WORK

The field of natural language processing (NLP) has studied analogies due to their common use in language and their importance for understanding semantic relationships between words and phrases. For example, classic work in NLP focuses on the concept of word embeddings where words like king and queen are statistically likely to be used in the same contexts [70]. Careful analyses of pre-transformer architectures, such as word2vec [70], demonstrated that the surprisingly high accuracy (on the order of 60%) on analogy word problems (e.g., "man is to woman as king is to <ans:queen>") hid large variations in accuracy across subtypes of analogy word problems: for example, [19] observed consistently higher accuracy on syntactic analogies (e.g., based on morphological transformations

like "big is bigger as small is to <ans:smaller>") compared to semantic analogies based on causal relations [19]. Some researchers were nevertheless able to leverage these models as base layers or inputs to NLP pipelines that were able to do analogical retrieval in complex natural language datasets such as crowdsourced ideas [51], and research papers [13]. More recent transformer-based models, such as GPT-3, have shown improved performance on this task. For instance, the largest 175B parameter version of GPT-3 (codenamed *davinci*) achieved comparable performance to humans (~65% accuracy) on a set of SAT four-term analogy problems, which are considered more challenging than the analogy word problems studied in previous work [9]. There is also some evidence that more recent models that add alignment training procedures [76], such as GPT-3's text-davinci-003, can match or exceed human benchmark performance on these tasks [5, 25, 90].

More importantly for our current purposes, researchers and practitioners have begun to produce proof-of-concept demonstrations of LLMs' ability to generate more complex natural language analogies. For instance, Zhu et al [95] showed examples of using GPT-3's earlier davinci models to generate analogous design concepts when prompted with analogies between a source problem and real-world design. Similarly, Webb et al [90] replicated the classic analogical problem solving paradigm from Gick and Holyoak [40] with the text-davinci-003 version of GPT-3, showing that the model was able to generate a convergence solution for Duncker's radiation problem when prompted with the analogous generals story, as well as describing the analogical mapping between the problem and the analogy (though the model failed to generate a plausible analogous solution for a different physics-based problem, despite successfully describing the analogical mapping between the problem and its analogy). In a slightly different paradigm, but also with an aligned version of GPT, Bhavya et al [5] showed that ~60% of InstructGPT [76]-generated analogical explanations for scientific concepts were rated by crowd workers as containing a meaningful analogy, a rate comparable to a dataset of human-generated analogical explanations. On the industry/practitioner side, the prominent AI Dungeon startup published a blog post that described its experiments using GPT-3 to generate descriptions of fantasy worlds by analogy [87]. In contrast to the simpler analogy word problems, these more complex analogy-generation capabilities have received little systematic evaluation in the context of creative tasks. One exception is a recent study by Zhu et al [96], who followed up on their previous proof-of-concept [95] by integrating 10 analogically generated design concepts into a design team's brainstorming process, and obtained ratings of the novelty and feasibility of the concepts: the analogical concepts generally received high novelty scores but low feasibility scores.

Overall, existing research suggests that LLM-generated analogies might frequently include sufficient cross-domain analogical mappings to inspire creative problem reformulation and ideation. In this study, we seek to extend the predominantly informal, proof-of-concept demonstrations in prior work with systematic, direct evaluation of LLM-generated analogies in the context of a creative task.

3 PROMPT ENGINEERING EXPLORATIONS: ANALOGY GENERATION QUALITY AS A FUNCTION OF PROMPT-BASED LEARNING DESIGN

To develop analogies for this study, we followed a *prompt-based learning* paradigm [64] (also called in-context learning in machine learning research [9]): rather than fine-tuning GPT-3 for analogy generation (e.g., in a supervised learning paradigm), we "prompted" an LLM by providing a set of natural language instructions for a task of generating cross-domain analogies, and took the LLM's generated text completion as the output. We iteratively improved our prompt through prompt engineering [79], which involves crafting a prompt for a LLM, issuing the prompt, and evaluating the response from an LLM.

We supplemented this iterative process with a systematic evaluation of the best-performing prompt from the prompt engineering process, exploring zero-, one-, and few-shot learning variants of the prompt. The results from this systematic evaluation are reported in this section.

3.1 Task setup and prompt-based learning design

We conducted our experiments with OpenAI's GPT-3 API, with the text-davinci-002 model, with temperature = 1 for the largest output variety and token count = 400 to accommodate the output length. Because we are interested in exploring how LLM-generated analogies could be integrated into problem formulation, we experimented with a prompt design where we provided a structure of the input problem and the output analogy by dividing a design problem into four components—stakeholder, context, goal and obstacle led to higher quality outputs. This was inspired by previous work on problem formulation in design [68], and in our initial prompt engineering iterations, we informally observed that this enabled us to balance the distance (stakeholder and context are different) and usefulness (goal and obstacle are similar) of the generated analogy much better than an earlier design that prompted GPT-3 to generate an abstracted schema for a given problem statement, taking inspiration from previous work on schema-based analogical transfer [41, 93]. We also explored a range of zero-shot to few-shot prompt-based learning paradigms. The structure of the prompt given to GPT-3 in the experiments in this paper is shown in Figure 1. We constructed three example problem-analogy pairs to illustrate the core idea of a cross-domain analogy. Figure 2 shows the three examples we constructed.

We wrote six design problems across various topics (food insecurity, job security, entertainment, etc) as shown in Figure 3. Figure 3 also shows examples of prompt and GPT-3 generated output pairs with zero-shot, one-shot, and few-shot paradigms. We used GPT-3 to generate 20 analogies each for input problems, for each of three prompt programming paradigms: zero-shot - no example of input and output; one-shot - one example; and few-shot - three examples. To reduce the potential for overfitting on a specific example in the one-shot case, we conducted two runs of the one-shot paradigm, each with a different example (the `manager` and `Duncker` analogy examples). This process resulted in a total of 480 problem-analogy

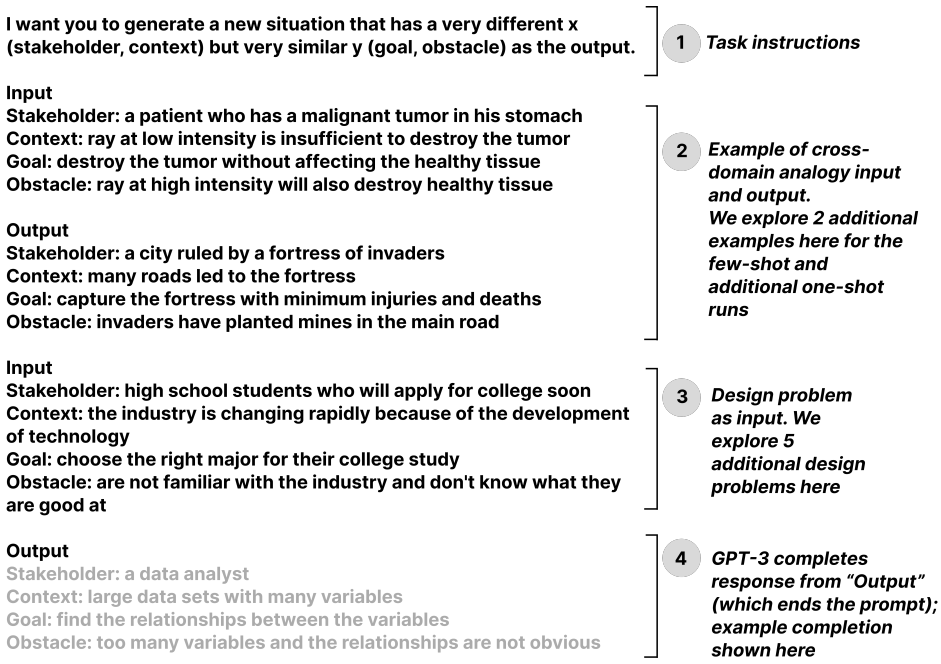

Figure 1: Our prompt-based learning prompts for generating cross-domain analogies.

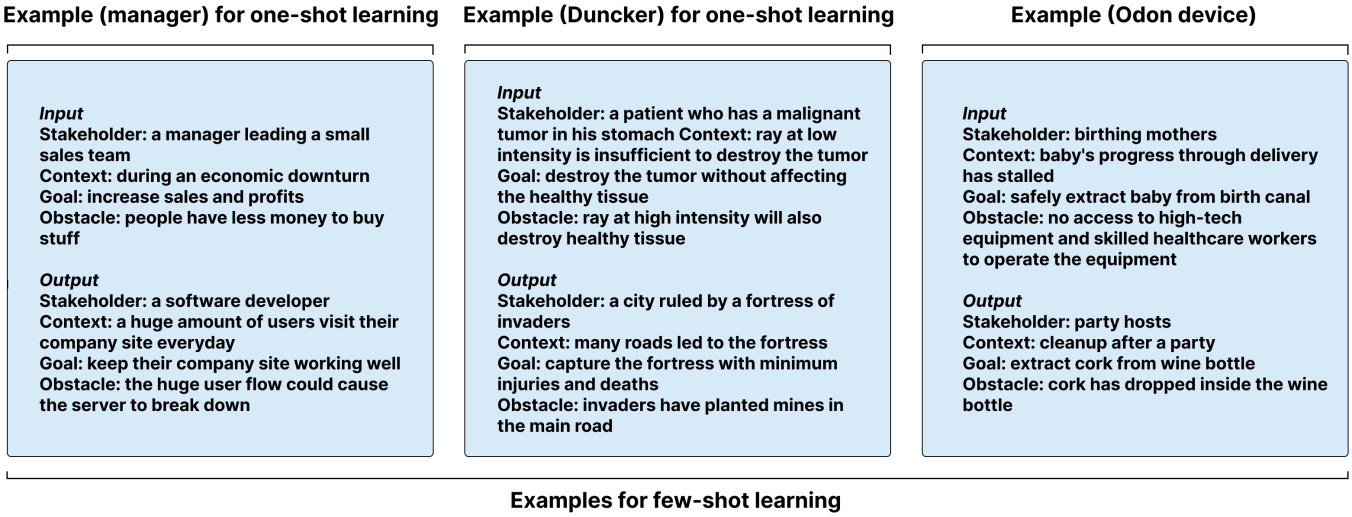

Figure 2: Problem-analogy pairs used as examples in prompts for our one-shot and few-shot learning paradigm runs.

pairs across four runs (one zero-shot, two one-shot, and one few-shot). The code for these prompts, along with the 480 generated analogies, is included in the Supplementary Material.

3.2 Measures

All 480 analogies generated with different methods were rated by a single PhD-level researcher with expertise in creativity support tools, example-based creativity, and problem reformulation, who was blind to condition. Analogies were rated for potential usefulness for creative reformulation. The primary criteria were 1) whether there was an analogical mapping between the generated analogy and the design problem, 2) whether the analogy was unique or a duplicate/repeat of any other generated analogy for the same problem, and 3) a "best guess" of whether the potential inferences from the analogy might be useful for inspiring new creative formulations of the problem. Note that the "potential usefulness" criterion would be more systematically and robustly investigated in Study 2; here, we use these heuristic judgments to inform our initial prompt engineering efforts, before systematically testing the outputs from the best-performing prompt design in the context of a creative task.

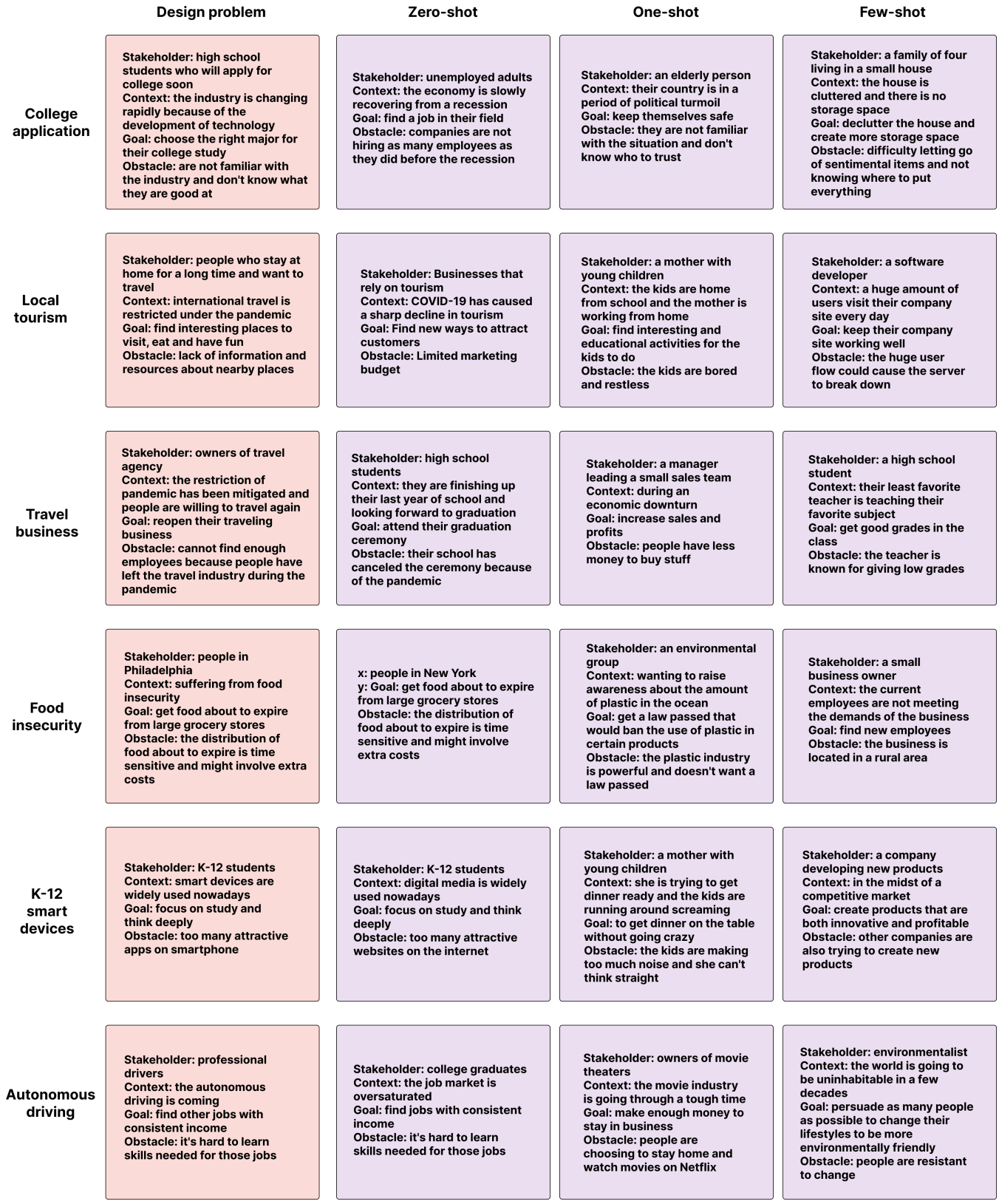

Figure 3: Six design problems used for Study 1 and corresponding examples of GPT-3 generated analogies with zero-shot, one-shot, and few-shot paradigms.

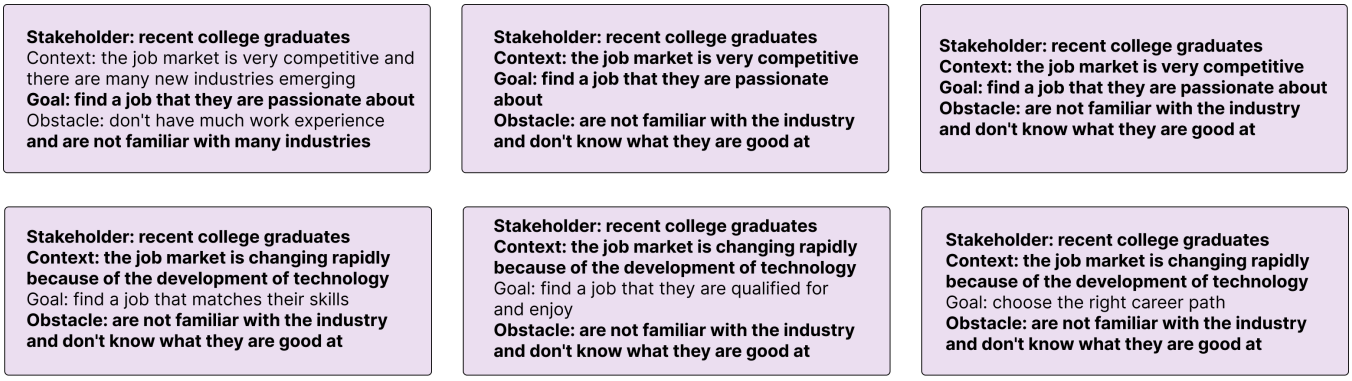

Figure 4: Examples of GPT-3 generated analogies that were judged as duplicates for the same design problem "choose the right major for their college study" with zero-shot learning. These duplicate analogies shared very similar concepts/aspects with substantial verbatim overlap in terms (bolded in the figure).

The researcher also measured the degree of duplication manually during screening, judging whether analogies were so similar to previously encountered analogies that it was redundant. To summarize, the rater screened the analogies and categorized them using three codes: 1) not potentially useful, 2) potentially useful and new, and 3) potentially useful but duplicated with a previous one. Examples of duplicate analogies are shown in Figure 4.

Because we are interested in the ability of LLMs to help with cross-domain analogical reformulation, we also measured the similarity of the generated analogies to the input problems, using semantic similarity calculated with SentenceTransformers[5].

3.3 Exploration results

As shown in Table 1, our key result was that one-shot outputs were judged to be consistently more potentially useful (~80% compared to ~40%) and dissimilar (<0.5 compared to >0.7) from the input problems compared to zero-shot outputs. Few-shot outputs were also substantially less similar to input problems (~0.3 compared to ~0.7) compared to zero-shot outputs, but the judged potential usefulness was slightly lower than one-shot outputs (~67% compared to ~80%.

While we found an advantage of the one-shot paradigm over the zero-shot paradigm in terms of potential usefulness and semantic distance, we are not confident that the one-shot paradigm would consistently produce better results than few-shot paradigms in general, given the limited number of examples tested and the strong prior from the literature that in-context learning performance improves with the number of examples [9, 91]. We *are* comfortable concluding for our study that few-shot learning (with k of at least 1) is likely to yield better results than zero-shot learning; practically, because the one-shot learning paradigm required fewer examples, reducing the effort and cost associated with prompt size, we systematically evaluated the outputs of the one-shot paradigm in the context of a creative task in Study 1.

[5]https://www.sbert.net/

4 STUDY 1: ANALOGY GENERATION QUALITY IN USAGE IN A CREATIVE PROBLEM REFORMULATION TASK

Having established an initial estimate of the rate at which GPT-3 is able to produce potentially useful cross-domain analogies (i.e., 70% unique and potentially useful cross-domain analogies in the best prompt design), we now directly study the degree to which these analogies could actually be useful in a creative task. To do this, we provided the analogy outputs from the best-performing one-shot learning paradigm (i.e., the prompt that included the Duncker example analogy) to people doing a creative ideation task, and asked them to try to use it to inform/inspire their creative ideation and evaluate the analogies' usefulness for that purpose.

4.1 Methods

4.1.1 Design and Materials. To account for potential variation across people in their reactions to the analogies, we sought to obtain two judgments of usefulness per analogy. We also aimed to minimize the anchoring of our results to any particular design problem (since there was some variation in the duplicate and potential judged usefulness rate across the design problems). Finally, we wanted to balance the time requirements per rater against minimizing the potential overhead of learning the task, since using analogies for problem reformulation may be different from a more typical NLP evaluation task like judging whether a given text is human-like (though still well within the range of everyday ability). Given these requirements, we designed our overall rating task to have two analogies for each rater for the same design problem, and approximately equal representation of analogies across the six design problems. The details of the rating task are explained in the next section.

To construct sets of two analogies for each rater, we first took the 92 of the 120 total analogies from the Duncker run of the one-shot learning paradigm in prompt engineering explorations that were judged to be both nonredundant and potentially useful, as well as the 17 analogy outputs that were judged to be nonredundant cross-domain analogies but unlikely to be useful. We ignored the

	Avg. semantic sim	Potential usefulness rate	Duplicate rate	Unique & potential usefulness rate
Zero-shot	0.72	0.68	0.24	0.43
One-shot (manager)	0.48	0.82	0.12	0.70
One-shot (Duncker) [for verification]	0.43	0.8	0.03	0.77
Few-shots	0.31	0.67	0.02	0.65

Table 1: Results of prompt engineering explorations: Average semantic similarities between original problems and GPT-3 generated analogical problems and human judgments of potential usefulness and duplication. In these runs, the one-shot learning paradigm runs produced the best combination (with an extra round with another example for verification).

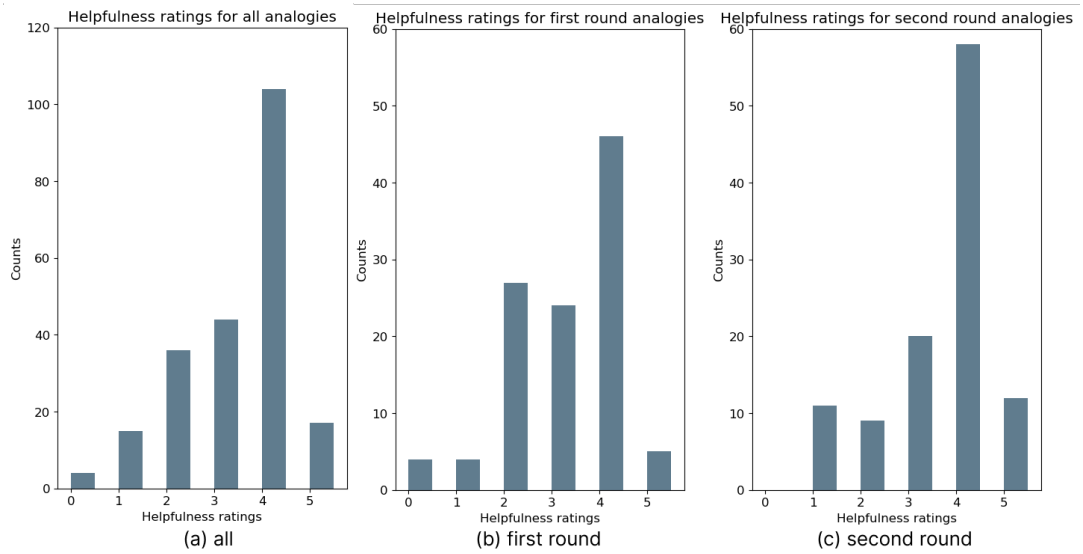

Figure 5: Histograms for the helpfulness ratings provided by participants for analogies in the context of the problem reformulation task in Study 1: a) all analogies, b) analogies in the first round, and c) analogies in the second round.

remaining 11 analogy outputs that were potentially useful but duplicates (N=4), highly similar to the original problem (N=4), and off-task (e.g., repeating the original problem, incorrect format; N=3). This resulted in 109 unique analogies across the six problems. To enable construction of pairs of analogies, we then randomly selected one of the analogies from a design problem that had an uneven number of usable analogies to drop from the sample. This resulted in a final set of 108 analogies. From these 108 analogies, we then constructed all possible pairwise combinations of analogies within each design problem, and sampled iteratively from these combinations to select pairs, such that each of the 108 analogies showed up in exactly 2 pairs. This resulted in 108 analogy pairs that were given to raters.

4.1.2 Task and Procedure. The overall rating task was designed to follow substeps of analogical problem reformulation, beginning with initial formulation of an incompletely specified problem, initial ideation, processing of an analogical stimulus in relation to the problem, and then reformulation and ideation over the reformulated problem. As an example, a designer given a problem of designing ways for people who want to find interesting places to visit and have fun but lack information and resources about nearby places might (re)formulate the problem by *adding* information about the stakeholder(s) (e.g., thinking about friends and family who might have good recommendations) or *rejecting/shifting* a potentially implicit assumption that they need to discover new places, and instead explore new goals around re-experiencing familiar places. This reformulation might be spurred by comparison to analogous problems, such as a parent trying to find ways to entertain their toddler during summer vacations (e.g., by considering how children can often repeatedly enjoy similar experiences or toys in a variety of ways).

Thus, the overall task was divided into three phases: an **initial formulation** phase, and two **analogical reformulation and ideation** phases, one for each analogy in the pair assigned to the participant. In the **initial formulation** phase, participants were first given a design problem and asked to construct an initial formulation of the problem through adding details to the problem components (stakeholder/context/goal/obstacle), in response to the question, "What characteristics of the stakeholder/context or other problem components do you think would be important to consider when trying to solve the problem?". Then, participants were asked to generate at least one solution to that problem with details added. In each of the two **analogical reformulation and**

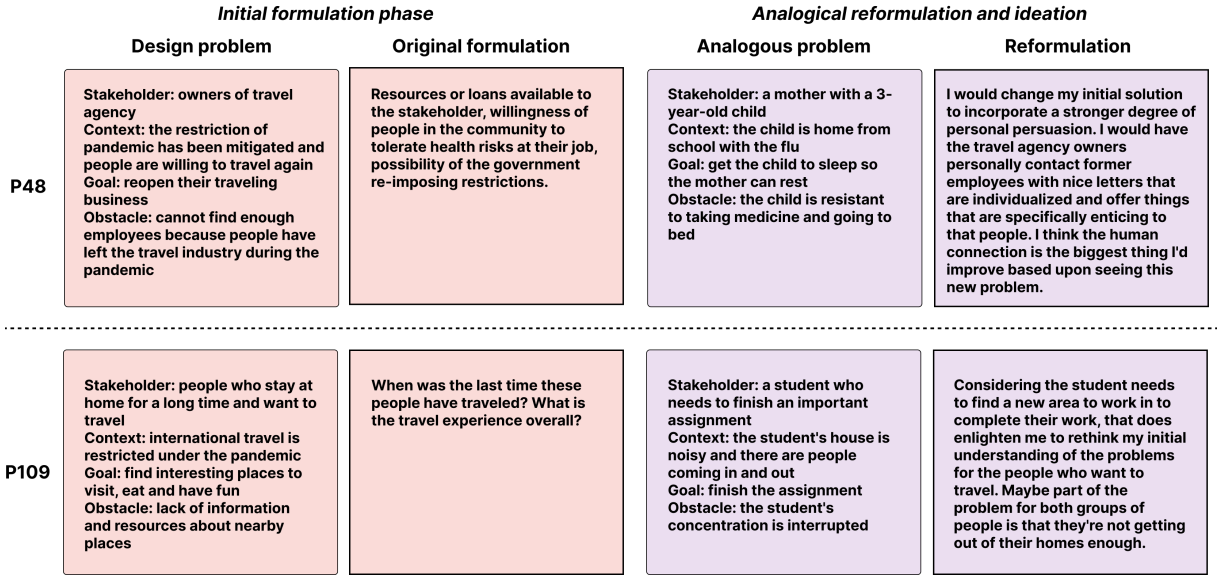

Figure 6: Examples of how GPT-3 generated analogies were used for reformulation by two participants (P48, top; and P109, bottom). We show the progression from their original formulation of the problem (from the initial formulation phase) to the reformulation of the problem in response to an analogy (from one of the analogical reformulation phases).

ideation phases, participants were presented with an analogy, and then answered two questions 1) "What interesting commonalities and differences do you notice between the problems (the original and analogical problems)?" 2) "Based on your comparison of the original problem with the new problem, can you think of ways to change your initial understanding of the original problem that might lead to new solutions? What might you add/change to your understanding of the problem components?" The first question was meant to facilitate analogical/case comparison, a known effective strategy for inducing analogical schemas for ideation [41, 58, 65]. The second question was meant to probe for potential *reformulation* of the problem based on the analogical mapping. After this, participants were asked to come up with a new solution to the original problem. To get a more realistic sense of the potential value of the analogies for ideation, we did not force participants to generate ideas in this phase from the analogies: participants were allowed to use Google to search for inspiration to generate ideas. After each task, participants were asked to rate how helpful they found it to compare their problem situation with the analogous one from 1 - "Not helpful at all" to 5 - "Very helpful" (no rating was 0). This process was then repeated for the second analogy in the set. After the second analogical reformulation and ideation phase, participants had a chance to give open-ended comments on the study. Appendix A shows screenshots of the task interface for the initial formulation and analogical reformulation and ideation phases of Study 1. We obtained institutional IRB approval for the project prior to Study 1.

4.1.3 Participants. We recruited participants from Prolific[6] as Prolific participants typically exhibit greater care in completing tasks compared to Amazon MTurk participants [74]. We limited all participants to U.S. residents. Each participant was paid $5.25 ($10.5 per hour) for their participation of 30 mins. A total of 123 participants enrolled the task, but 13 of them only finished the first round of analogical reformulation and ideation and did not complete the second round. After removing those participants, our final sample included 110 participants who finished the study. Due to our randomized procedure for assigning participants to analogy sets, removing incomplete responses also resulted in 22 analogies being used and evaluated by only one participant.

4.1.4 Systematic analysis of reformulation behaviors. To understand how participants used the analogies to reformulate the original problems, we systematically coded participants' responses to the reformulation question in the analogical reformulation and ideation phases:

> "Based on your comparison of the original problem with the new problem, can you think of ways to change your initial understanding of the original problem that might lead to new solutions? What might you add/change to your understanding of the problem components?"

We wanted our coding scheme to capture a core distinction in the creativity literature between formulation and reformulation. The former — formulation — entails adding details to to transform an ill-structured problem into a more structured problem that can make the search for solutions more tractable. This process is necessary because creative problems are not static "givens"; rather, these problems are ill-structured and malleable. In contrast to well-structured problems (e.g., the Tower of Hanoi, algebra), where the initial description of the problem sufficiently delineates the goal state and allowable transformations, in creative problem domains, such as innovative design or interdisciplinary research, the "problem statement" typically underspecifies what the desired solution

[6] https://www.prolific.co/

is. In information-processing terms, "the start state is incompletely specified, the goal state is specified to an even lesser extent, and the transformation function from the start to goal states is completely unspecified" [42]. Analogical transfer can be instrumental in this elaboration process, adding functional decompositions or subproblems to solve that might address the high-level problem [48]. The latter — reformulation — entails reshaping an existing structured problem, often by rejecting or relaxing constraints and assumptions, or even reformulating the goal itself. An innovative designer can choose to see the design problem differently, to relax certain constraints or redefine certain customer needs in other terms, thereby arriving at innovative solutions. Such problem-reframing has been characterized as key element of creative expertise in design [23, 83]. For example, Bucciarelli [10] observed in his ethnographic studies of engineering designers that a common strategy was to re-evaluate and sometimes alter formulations of the design problem. Goel and Pirolli [42] call this a "reversal of the transformation function", where, instead of reducing the difference between the current and goal states by means of a design "move" (e.g., a particular product configuration), a designer moves closer to the goal state by changing what the goal state is. In insight problem solving, too, a key mechanism for moving past design impasses, is to relax constraints and assumptions [57], or even change what details of the problem are attended to, which can be characterized as a search not for solutions, but for problem spaces, where solutions are more readily found [55]. Here, too, analogy can be used to support reformulation; For example, [27] described a case where Designers used an analogy to a music festival to (re)frame the problem of night violence in King's Cross from a crime prevention/reduction problem to an entertainment frame.

Thus, we initially employed two codes - "adding" and "shifting" - to capture notions of "formulation" (elaborating/adding), and "reformulation" (changing/rejecting/shifting). However, we discovered that the "shifting" code also involved instances of "adding," and disagreements arose regarding whether a reformulation constituted "adding," "shifting," or "adding" plus "shifting." This resulted in low inter-rater reliability. To address this, we decomposed the "shifting" code into two more concrete and reliable codes: "adding" and "rejecting", which were not mutually exclusive, and (if both are present) could still capture the notion of shifting. Our final coding scheme thus consisted of two independent codes: 1) **adding** (0/1), indicating any presence of *new* concepts or descriptions in the answer compared to the initial problem statement or formulation; and 2) **rejecting** (0/1), indicating direct statements from the participant that they are rejecting a goal/obstacle/context/ stakeholder from the problem or their initial formulation. A problem reformulation could be seen as a combination of adding and rejecting. Examples of adding and rejecting are provided below:

Example of adding
Original problem
Stakeholder: people who stay at home for a long time and want to travel
Context: international travel is restricted under the pandemic
Goal: find interesting places to visit, eat and have fun
Obstacle: lack of information and resources about nearby places
Analogy
Stakeholder: a software company
Context: developing a new mobile application
Goal: find out user requirements
Obstacle: most of the users are not familiar with the application's purpose
Participant's response to reformulation question
I would **add specific locations** to the problem because that would narrow down the scope of **how to get information to them**.

Example of rejecting
Original problem
Stakeholder: owners of travel agency
Context: the restriction of pandemic has been mitigated and people are willing to travel again
Goal: reopen their traveling business
Obstacle: cannot find enough employees because people have left the travel industry during the pandemic
Analogy
Stakeholder: a farmer
Context: the restriction of the use of pesticides has been mitigated
Goal: use pesticides to increase crop yield
Obstacle: the farmer cannot afford to buy pesticides
Participant's response to reformulation question
It's not that the travel agency can't find employees, it's that they can't afford to pay employees to work for them after being closed for so long, thus causing a feedback loop of: not enough employees -> less money ->cant afford to hire employees -> not enough employees.

Two graduate-level researchers with expertise in creativity support tools, example-based creativity, and problem reformulation, applied the coding scheme to the participants' reformulation responses. After five rounds of coding, discussing (and resolving) disagreements, and refining the coding scheme between two expert coders, the inter-rater reliability reached Cohen's Kappa=1 (perfect agreement) for adding codes and Cohen's Kappa=0.847 (near perfect agreement) for rejecting codes respectively for 50 problem-analogy pairs [59]. After developing the substantial inter-rater reliability, one coder coded the remaining 170 reformulation responses.

4.2 Quantitative Results

All 220 analogy-response pairs from all completed participants (N=110) are included in the Supplementary Material. We first report a range of quantitative analyses of LLM-generated analogies' quality at a whole sample level, before describing some illustrative case studies of how the analogies were processed and used to contextualize the high level results.

4.2.1 Median 4 out of 5 helpfulness rating for analogy, with increased rating on the second analogy. Since analogies received unequal numbers of ratings, we report the median as an estimate of the central tendency of ratings across analogies. The median rated helpfulness of analogies was 4 out of 5 (see Table 2 and Figure 5). Interestingly,

	Count	Median of helpfulness rating
Adding	185 (84.09%)	4
Rejecting	23 (10.45%)	4
Adding + Rejecting	22 (10%)	4
Total	220	4

Table 2: Systematic coding results (adding / rejecting / adding + rejecting) of analogical reformulations in Study 1.

we noticed that ratings were higher for the analogies that were rated second in the analogical reformulation phase ($median = 4$) compared to the analogies that were rated first ($median = 3$). A Wilcoxon test (within subjects, two groups) estimated that this increase in median ratings from the first to second analogy was statistically significant ($statistic = 561.5, p - value = 0.002$). We also compared the median of helpfulness ratings for analogies judged by two or more participants and the median of helpfulness ratings for all analogies to check whether there was a difference between only one judgment vs. multiple judgments: both medians were 4, suggesting that we obtained similarly reliable estimates from single vs. multiple-judgments for each analogy

4.2.2 Frequent analogical reformulations, primarily by adding information, positively correlated with helpfulness ratings. The results of our systematic coding of the reformulation responses are reported in Table 2, showing 185/220 (84.09%) instances of adding, and 23/220 (10.45%) instances of rejecting. There was high overlap between instances of adding and rejecting: specifically, there were 22/220 (10%) instances of both adding and rejecting (shifting).

Since the add/reject/any change codings were binary and the helpfulness ratings were continuous, we calculated point-biserial correlations between them. We observed a statistically significant positive correlation between reformulation responses coded as having an addition and helpfulness ratings for the corresponding problem-analogy pair ($point - biserial = 0.2559, p = 0.0001$) and between any change codings (add or reject) and helpfulness ratings ($point - biserial = 0.2669, p = 0.00006$). However, we observed no statistically significant correlation between reject codings and helpfulness ratings ($point - biserial = 0.0094, p = 0.8891$).

4.2.3 No correlation between a priori usable ratings and in-use helpfulness ratings. To explore the correspondence between the *a priori* judgments of potential usefulness from prompt engineering explorations (outside the context of a creative task) with in-use ratings of usefulness of analogies, we calculated the correlation between the potential usefulness ratings from prompt engineering explorations and the helpfulness ratings here. Since the potential usefulness ratings were binary and the helpfulness ratings were continuous, we calculated a point-biserial correlations between them. We observed no statistically significant correlation between *a priori* usable ratings and helpfulness ratings: $point - biserial = -0.0266, p = 0.6939$. However, since we only utilized one researcher's ratings to predict the usefulness ratings provided by another set of participants, we cannot discount the possibility that the lack of correlation might be due to alternative explanations. For instance, it could be that the specific rater was unable to predict usefulness without actually using the analogies, or that the difficulty of predicting usefulness for someone else contributed to the lack of correlation.

4.3 Qualitative descriptions of how people used LLM-generated analogies to reformulate problems

To complement our systematic quantitative analyses of participants' reformulation behaviors in 4.1.4, we illustrate how participants used analogies with two more cases shown in Figure 6. In the first example of the "local tourism" problem, P48 wanted to gain more information of the context such as external resources and constrains ("resources or loans", "willingness of people", etc). P48 described additional contexts that were not included in the original design problem, such as "resources or loans that were available" and "willingness of people... to tolerate health risks." After seeing the analogical problem that a "child is resistant to taking medicine and going to bed", P48 highlighted the role of persuasion to solve the original problem and incorporated personal persuasion into the solution. P48 did not explicitly reformulate the original problem, but instead began ideating a solution, which is a common behavior for novice [66, 67]. But the solution suggested that P48 noticed a potential core obstacle "a lack of personal persuasion" within the original problem that employers "cannot find enough employees because people have left the travel industry during the pandemic". This is interesting because "the child is resistant to taking medicine and going to bed" can be seen as an interesting and plausible analogy to "previous employees of the travel industry who are resistant to return". In the second example, the original problem formulation of P109 was around the stakeholder. After seeing the analogy, P109 noticed the high-level similarity between "lack of entertainment information for people who stay at home" and "noisy environment of a student's house" and then explicitly reformulate the root problem as "they're not getting out of their homes enough". Those two examples demonstrated that the generated analogies were able to facilitate participants to deepen their understanding of the original problem and reformulate the root issues, which could inform better solutions.

Complementing these example cases are some notable open-ended comments from participants about how the analogies impacted their formulation and ideation processes. For example, P40, P29, and P109 commented on how the analogies opened their mind:

> Comparing the two problems **opened new mental doors**, only one of which led to a practical, workable solution. But the fact that those mental doors were completely invisible to me before seeing the analogous problem suggests that looking at related but not identical problems can be extremely helpful. (P40)

> I think reading from analogies can definitely **open your mind up** to new ideas and improve critical thinking skills. I didn't even realize at first that it was

> "people" in Philadelphia, until the next slide. I totally overlooked it, but it's interesting how so many things correlate, even if they're very different. (P29)
>
> This was extremely interesting for me! I really like that I was able to compare problem situations with the analogies here and expand my original thinking. This allowed me to identify new, creative solutions to the original problem that I would not have thought of without the expanded analogies! (P109)

Similarly, P79 noted:

> These **juxtapositions** create very higher order thinking involving public policy expertise as well as scientific and community knowledge. Very interesting! These creative solutions are necessary to improve social welfare and environmental problem solving. Thank you! (P79)

Another participant's comment also illuminated the lack of correlation between pre-judged potential usefulness and helpfulness in the context of the creative task:

> Overall, I found the analogies to be a bit random and not the best fit, but they did make me think of the problem from a slightly different angle, which was interesting." (P68).

Overall, these qualitative descriptions and open-ended comments complement the quantitative results above by adding more concrete details to illustrate *how* the LLM-generated analogies were useful in the creative problem formulation task.

5 STUDY 2: POTENTIAL HARMFULNESS AND TOXICITY OF GENERATED ANALOGIES

In this final section, we systematically investigate the degree to which GPT-generated analogies might be harmful or toxic. We do this because we are interested in real-world usage of a method like this, and want to facilitate weighing of cost-benefit considerations and design of mitigations for potential harms, especially given concerns about potential harms from large language model usage [2].

Our analysis here is anchored on the potential use case of giving LLM-generated analogies to designers as an aid to problem formulation. Inspired by previous analyses of the ethics of LLMs based not just on direct harm to the user but also participation in harmful and oppressive systems of power [2, 6], we consider the following primary categories of harm: 1) perpetuating hegemonic systems of oppression via uncritical incorporation of harmful stereotypes and biases in LLM outputs, 2) unknowingly framing problems in a way that leads to solutions that violate laws, and 3) experiencing psychological distress from upsetting or abusive content.

Analyzing the bias and stereotype-based categories of harm typically involves consideration of specific social categories at risk of harm from biases, such as gender [84], or disability [52]. As Bender et al [2] note, which particular social categories are salient is culture-bound. Thus, we note here that we anchored our conceptualization of protected social categories from a US-centric perspective due to our positionality as US-based researchers, and considered the dimensions of disability, race/ethnicity, gender, sexual orientation and age. To analyze potential illegal suggestions, we considered the context of US laws, again due to our positionality. Finally, we intentionally drew on raters from a variety of demographic backgrounds of race/ethnicity and gender, to allow for more expansive detection of content that may be considered psychologically distressing. We note all these to emphasize the situatedness of our results, and make no claims as to their generality to other contexts, and share our raw outputs to assist with extensions of this work that consider this question from other positionalities.

From a measurement perspective, harmfulness audits have often been done using automated measures such as the Perspective API[7] toxicity classifier [2]. However, the sensitivity and accuracy of these measures is uncertain, and there is evidence that they may produce false positives with mere mentions of terms associated with marginalized identities, such as race or disability [52, 53]. Thus, we chose to manually review all outputs, and estimate ranges in rates of harmfulness or toxicity, as well as qualitatively describe patterns of outputs on these dimensions.

5.1 Methods

Three members of the research team (1 White male junior faculty, 1 Asian immigrant male junior faculty, and 1 Asian immigrant male PhD student), and two additional research assistants (1 Asian female undergraduate student, 1 North African female PhD student) reviewed all 108 analogies used in Study 1. We report the general demographics of the screeners to facilitate understanding of how our results here on the potential harmfulness of the outputs may be shaped by our positionality, recognizing that other researchers or users in different positionalities may reach different conclusions about the LLM outputs.

Using the three-fold conceptualization of harm described above (bias, illegality, and potential psychological distress), each screener independently went through each analogy, and flagged it as potentially harmful or not. Screeners also added descriptive comments about potential harm, where appropriate. We use the *harmful/ not harmful* flags to estimate rates of harmfulness, and the descriptive comments to qualitatively describe the nature of potential harms.

5.2 Results

There was wide variation in the number of flagged analogies across screeners, ranging from 1 to 27 out of 108 total analogies: specifically, the number of flagged analogies across the five screeners was, in ascending order: 1, 2, 2, 10, and 27. Using the most expansive approach of counting any flag by any screener increased the upper bound rate to 28/108 (26%).

Qualitatively, the majority of the analogies that were flagged across screeners were for potentially upsetting content, rather than biased/abusive language or illegal behavior. Figure 7 shows the 10 analogies that were flagged by at least 2 of the 5 raters as potentially harmful, along with the associated screener notes. There was one clear instance of biased depiction of people in a small village in Africa with multiple stereotypes of Africa as poverty-ridden and a "wild safari". There was also one instance of illegal behavior described (smuggling liquor in the 1920's). The remaining 8 analogies were flagged for potentially upsetting content, such as difficult

[7] https://www.perspectiveapi.com/

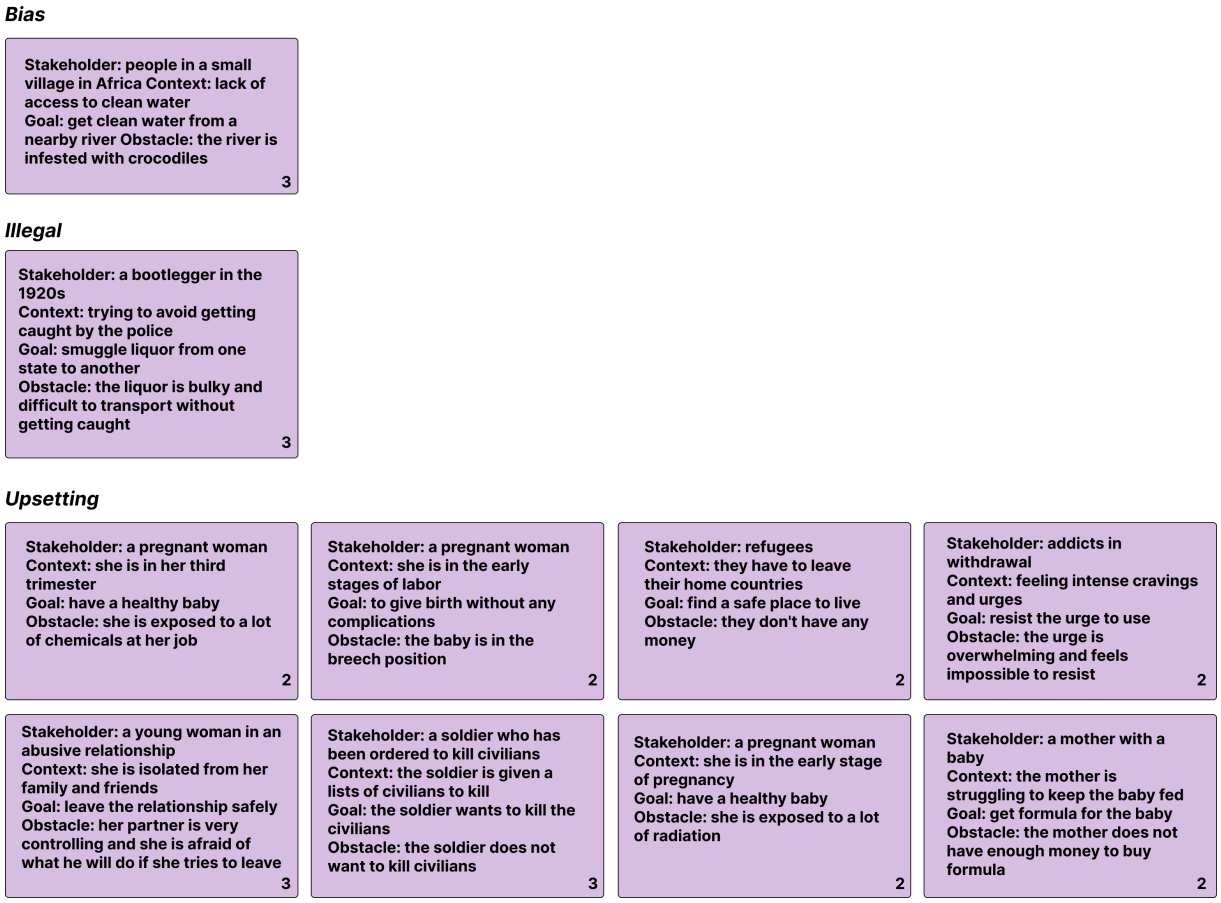

Figure 7: Analogies that were flagged by at least 2 of the 5 raters as potentially harmful under three categories: biased, illegal, and upsetting. The subscript for each example showed how many people flagged the example as harmful.

birthing/pregnancy, refugees fleeing their country, trying to escape an abusive relationship, substance addiction, and war. In the subset of screening results from the screener who flagged the most analogies, we also observed a similar pattern where predominantly potentially upsetting content accounted for most of the potentially harmful analogies.

6 DISCUSSION

6.1 Summary and interpretation of findings

In this paper, we directly investigated the potential of LLMs to generate analogies that might be useful for creativity support, specifically in the task context of augmenting cross-domain analogical reformulation. Through carefully crafted prompts, we were able to use prompt-based learning to generate cross-domain analogies with GPT-3 that were semantically distant from source problems and judged to be potentially useful at a high rate (~70-80% of outputs; prompt engineering explorations). We complemented this prompt engineering exploration with a systematic analysis of how people used LLM-generated analogies in the context of a creative problem reformulation task, and found that they were able to leverage the analogies to reformulate design problems (~80% of cases), rated the analogies overall as helpful for reformulating the problems (average ~3-4 Likert rating out of 5), and described the analogies as augmenting/broadening their thinking process (Study 1). Investigating potential tradeoffs against toxic/bias/harmful outputs, we found an upper bound of ~25% of outputs being potentially harmful in some way, with a clear majority (~80%) of such cases being judged as such due to descriptions of potentially upsetting situations, rather than biased or toxic descriptions of specific social categories.

We infer from these analyses that LLMs, with carefully crafted prompts, can frequently generate analogies that can augment cross-domain analogical reformulation. The cost/benefit analysis of deploying these analogies in creativity support tools may vary by situation given the potential for upsetting content. However, note that our screeners were emphasizing minimizing false negatives, and many problematic situations may inherently contain content that may be upsetting to some (since they, by definition, describe situations that people want to transform to "better" states [27, 71, 80]).

More broadly, our results add to prior work on the analogical reasoning capacities of LLMs, extending from simpler analogy word

problems [5, 9, 90] to more complex cross-domain analogical outputs, and extending prior proofs-of-concept [5, 90, 95] with systematic testing of analogical outputs in the context of creative tasks. Our prompting methodology with structured inputs also has the potential to be extended to other domains with repeating structural components, such as scientific reasoning (e.g., theory, evidence, method) or legal reasoning (e.g., precedent, defendant, issue). Additionally, our observed lack of correlation between a priori judgments of potential usefulness from prompt engineering explorations and helpfulness ratings and reformulation behaviors in the context of an actual task underscore the importance of going beyond simple benchmark-based measures of output "quality" [60], or even human judgments of quality separate from usage in a creative task [21] (although, as we note in the results section for Section 4, we cannot rule out alternative explanations such as lack of ability for this specific rater, or the possibility of successfully predicting usefulness for oneself, but not others). The significant increase in helpfulness ratings from the first to second analogy in the analogical reformulation phases in Study 1 also suggest that some degree of "settling into" a creative analogy-making task is needed to get more accurate measures of the quality of LLM-generated analogies; this effect is reminiscent of prior results showing the induction of a "relational mindset" from initial processing of cross-domain analogies that improved subsequent recognition and processing of cross-domain analogies [43, 89].

6.2 Limitations and next steps

6.2.1 Beyond the one-shot prompt-based learning paradigm. In this paper, we only sampled from one run of the one-shot paradigm with one example for Study 1's in-use evaluation due to the cost of human evaluation. This made sense given that prompt engineering explorations showed an advantage of the one-shot paradigm over the zero-shot paradigm in terms of uniqueness and judgments of potential usefulness, and comparable uniqueness and potential usefulness to the few-shot paradigm. However, while we are fairly confident that the one-shot paradigm produces better results than the zero-shot paradigm, we are not confident that one-shot paradigms would be consistently better than few-shot paradigms, given the small number of examples we tested, and the substantial literature on how in-context learning improves with the number of examples [9, 91]. A potential direction for future work is to expand on different instantiations of few-shot learning paradigms. For example, future work might explore the effect of providing uniform or diverse examples on the usefulness of generated analogies. It may also be fruitful to more systematically explore variations of prompt programming, such as more specific, varied, or longer / more complex prompts, as well as investigating the extent to which a more traditional fine-tuning approach may yield substantially better results. That said, from a practical standpoint, one-shot learning paradigms have the benefit of requiring less examples, which reduces prompt engineering effort and costs associated with prompt size.

6.2.2 Improving the depth of insight in LLM-generated analogies. While the rate of generating cross-domain analogies was relatively high, it is unclear to what extent these analogies could spark deep insights or highly creative conceptual leaps. Note that the rate of *rejecting*/shifting elements of the problem in participants' reformulation (10.45%) was fairly low. This relative lack of rejecting/shifting may reflect the particulars of the task and participants in Study 1. But we also wonder if, as others have suggested [90], the insightfulness of LLM-generated analogies may be bounded by the degree to which the model has access to rich domain knowledge, such as commonsense or physical knowledge. In this sense, LLMs' capacity for analogy generation observed here may be more a function of what Mahowald et al [69] call "formal linguistic competence" (knowledge of rules and patterns of a given language), as opposed to deeper "functional linguistic competence" (being able to understand and use language in real-world tasks, such as formal reasoning, situation modeling, and social reasoning. This general idea is consistent with the informal observations of the screener for prompt engineering explorations, who noticed that many of the analogies seemed to be quite diverse, but didn't describe deep details about any particular domain. We wonder what the LLM-generated analogies might inspire if they were generated from a combination of LLMs and more structured/specialized knowledge bases (e.g., modeling common sense, domain ontologies, or physical world modeling); a recent proof-of-concept of something like this is a prototype using ChatGPT as a natural language interface, drawing from the much more structured and curated computational knowledgebase of Wolfram Alpha [92].

6.2.3 From static to iterative LLM-assisted cross-domain analogical reasoning. Our experiments were conducted with a static set of pregenerated examples for the zero- and few-shot prompts. We wonder how LLM-powered analogy generation might be integrated into the more iterative nature of design and creative cognition [28]. For instance, rather than using a fixed design problem as a source for LLM-generated analogies, could we integrate users' reformulations into subsequent prompts? Might the information of current user path enable LLMs to provide more personalized creative support for taking a deep dive or a "creative leap" [18, 72, 85]? We are curious how future work might fruitfully build on explorations by [96], who used LLMs not just to generate analogies, but also generate explanations for potential mappings in an analogy, and generate potential concept ideas. We can imagine design patterns where an initial round of LLM-generated analogies helps to stimulate memory retrieval of diverse domains — to assist with fixation [62, 63] and surface similarity bias in analogical retrieval [36, 37, 41, 50] — and then subsequent integration of LLM-assisted "deep dives" via explanations of potential mappings, generating variations of ideas within a theme, and so on.

Another interesting direction may be to leverage LLM-generated analogies to improve the process of iteration from feedback [14, 15, 29]. Studies of analogies in creative work show that they are useful for more than concept generation and problem reformulation: analogies are frequently used to assist with explaining unfamiliar concepts [1, 20, 30]. For example, the deployable space array technology is more explainable with an origami-folding analogy [7]. We wonder how AI-generated analogies could be used by creative workers not only to inspire themselves, but also to assist them in conveying design concepts to others by connecting to their domain knowledge.

ACKNOWLEDGMENTS

This research was supported in part by ONR N000142012506.

REFERENCES

[1] C. R. Bearman, L. J. Ball, and T. C. Ormerod. 2002. An Exploration of Real-World Analogical Problem Solving in Novices. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Shlomo Berkovsky, Yoshinori Hijikata, Jun Rekimoto, Margaret Burnett, Mark Billinghurst, and Aaron Quigley. 2018. How Novelists Use Generative Language Models: An Exploratory User Study. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan.

[4] S. Bhatta, A. Goel, and S. Prabhakar. 1994. Innovation in Analogical Design: A Model-Based Approach. In *Artificial Intelligence in Design '94*, John S. Gero and Fay Sudweeks (Eds.). Springer Netherlands, Dordrecht, 57–74. https://doi.org/10.1007/978-94-011-0928-4_4 00000.

[5] Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. http://arxiv.org/abs/2210.04186 arXiv:2210.04186 [cs].

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. https://doi.org/10.48550/arXiv.2005.14050 arXiv:2005.14050 [cs].

[7] Diana Bolanos, Collin Ynchausti, and Spencer P. Magleby. 2022. A Preliminary Approach to Select an Origami Source Pattern for Deployable Space Arrays. In *Volume 7: 46th Mechanisms and Robotics Conference (MR)*. American Society of Mechanical Engineers, St. Louis, Missouri, USA, V007T07A070. https://doi.org/10.1115/DETC2022-90033

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.07258 00021 arXiv: 2108.07258.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (June 2020). http://arxiv.org/abs/2005.14165 00030 arXiv: 2005.14165.

[10] Louis L. Bucciarelli. 1994. *Designing engineers*. MIT Press, Cambridge, MA.

[11] Chen Cao. 2023. Scaffolding CS1 Courses with a Large Language Model-Powered Intelligent Tutoring System. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 229–232. https://doi.org/10.1145/3581754.3584111

[12] Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding. http://arxiv.org/abs/2103.06779 arXiv:2103.06779 [cs].

[13] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of ACM Human-Computer Interaction: CSCW* 2, CSCW (2018), 31.

[14] Joel Chan, Steven C. Dang, and Steven P. Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*. https://doi.org/10.1145/2818048.2820023

[15] Joel Chan, Steven P. Dow, and Christian D. Schunn. 2015. Do The Best Design Ideas (Really) Come From Conceptually Distant Sources Of Inspiration? *Design Studies* 36 (2015), 31–58. https://doi.org/10.1016/j.destud.2014.08.001 00105.

[16] Joel Chan, Katherine Fu, Christian. D. Schunn, Jonathan Cagan, Kristin L. Wood, and Kenneth Kotovsky. 2011. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design* 133 (2011), 081004. https://doi.org/10.1115/1.4004396 00304.

[17] Joel Chan and Christian D. Schunn. 2015. The importance of iteration in creative conceptual combination. *Cognition* 145 (Dec. 2015), 104–115. https://doi.org/10.1016/j.cognition.2015.08.008

[18] Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin T. Solovey, Krzysztof Z. Gajos, and Steven P. Dow. 2017. Semantically Far Inspirations Considered Harmful?: Accounting for Cognitive States in Collaborative Ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. ACM, New York, NY, USA, 93–105. https://doi.org/10.1145/3059454.3059455

[19] Dawn Chen, Joshua C. Peterson, and Thomas L. Griffiths. 2017. Evaluating vector-space models of analogy. *arXiv:1705.04416 [cs]* (May 2017). http://arxiv.org/abs/1705.04416 arXiv: 1705.04416.

[20] B. T. Christensen and C. D. Schunn. 2007. The relationship of analogical distance to analogical function and preinventive structure: the case of engineering design. *Memory & Cognition* 35, 1 (2007), 29–38. https://doi.org/10.3758/BF03195939

[21] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *arXiv:2107.00061 [cs]* (July 2021). http://arxiv.org/abs/2107.00061 00008 arXiv: 2107.00061.

[22] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 329–340. https://doi.org/10.1145/3172944.3172983

[23] Nigel Cross and Anita Clayburn Cross. 1998. Expertise in engineering design. *Research in Engineering Design* 10, 3 (Sept. 1998), 141–149. https://doi.org/10.1007/BF01607156

[24] Darren W. Dahl and Page Moreau. 2002. The Influence and value of analogical thinking during new product ideation. *Journal of Marketing Research* 39, 1 (2002), 47–60. https://doi.org/10.1509/jmkr.39.1.47.18930 00714.

[25] Zijian Ding and Joel Chan. 2023. Mapping the Design Space of Interactions in Human-AI Text Co-creation Tasks. *arXiv e-prints* (2023), arXiv–2303.

[26] Alan Dix, Corina Sas, P Gomes da Silva, L McKnight, T Ormerod, and M Twidale. 2006. Why bad ideas are a good idea. In *HCIEd'06*.

[27] Kees Dorst. 2015. *Frame Innovation: Create New Thinking by Design*. The MIT Press. 00435.

[28] Kees Dorst and Nigel Cross. 2001. Creativity in the design process: co-evolution of problem–solution. *Design studies* 22, 5 (2001), 425–437. https://doi.org/10.1016/S0142-694X(01)00009-6 02587.

[29] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2010. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17, 4 (Dec. 2010), 18:1–18:24. https://doi.org/10.1145/1879831.1879836

[30] K. N. Dunbar. 1997. How scientists think: On-line creativity and conceptual change in science. In *Creative thought: An investigation of conceptual structures and processes*, T. B. Ward, S. M. Smith, and J. Vaid (Eds.). Washington D.C., 461–493.

[31] Karl Duncker. 1945. On problem-solving. *Psychological Monographs* 58, 5 (1945), i–113. https://doi.org/10.1037/h0093599

[32] Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural Generation of Interactive Stories using Language Models. In *International Conference on the Foundations of Digital Games*. ACM, Bugibba Malta, 1–8. https://doi.org/10.1145/3402942.3409599

[33] Katherine Fu, Joel Chan, Jonathan Cagan, Kenneth Kotovsky, Christian Schunn, and Kristin Wood. 2013. The Meaning of Near and Far: The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output. *Journal of Mechanical Design* 135, 2 (2013), 021007. https://doi.org/10.1115/1.4023158

[34] D. Gentner. 1983. Structure Mapping: A Theoretical Framework for Analogy. *Cognitive Science* 7 (1983), 155–170.

[35] Dedre Gentner. 2002. Analogy in Scientific Discovery: The Case of Johannes Kepler. In *Model-Based Reasoning*, Lorenzo Magnani and Nancy J. Nersessian (Eds.). Springer US, Boston, MA, 21–39. https://doi.org/10.1007/978-1-4615-0605-8_2

[36] Dedre Gentner and Russell Landers. 1985. Analogical reminding: A good match is hard to find. In *Proceedings of the international conference on systems, man and*

cybernetics.

[37] D. Gentner, M. J. Rattermann, and K. D. Forbus. 1993. The roles of similarity in transfer: separating retrievability from inferential soundness. *Cognitive Psychology* 25, 4 (1993), 524–75.

[38] 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 (CHI '19)*. ACM, New York, NY, USA, 296:1–296:12. https://doi.org/10.1145/3290605.3300526 event-place: Glasgow, Scotland Uk.

[39] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 1002–1019. https://doi.org/10.1145/3532106.3533533

[40] Mary L. Gick and Keith J. Holyoak. 1980. Analogical problem solving. *Cognitive Psychology* 12, 3 (1980), 306–355. 03379.

[41] M. L. Gick and K. J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6 04045.

[42] Vinod Goel and Peter Pirolli. 1992. The structure of design problem spaces. *Cognitive Science* 16 (1992), 395–429. https://doi.org/10.1207/s15516709cog1603_3

[43] Micah B. Goldwater and Anja Jamrozik. 2019. Can a relational mindset boost analogical retrieval? *Cognitive Research: Principles and Implications* 4, 1 (Dec. 2019), 47. https://doi.org/10.1186/s41235-019-0198-8

[44] W. J. J. Gordon. 1961. *Synectics: The Development of Creative Capacity*. New York, NY. 00000.

[45] Kazjon Grace, John Gero, and Rob Saunders. 2015. Interpretation-driven mapping: A framework for conducting search and rerepresentation in parallel for computational analogy in design. *AI EDAM* 29, 2 (May 2015), 185–201. https://doi.org/10.1017/S0890060415000062 Publisher: Cambridge University Press.

[46] Howard E. Gruber and Paul H. Barrett. 1974. *Darwin on man: A psychological study of scientific creativity*. E. P. Dutton, New York, NY, England. Pages: xxv, 495.

[47] Andrew Hargadon and Robert I. Sutton. 1997. Technology Brokering and Innovation in a Product Development Firm. *Administrative Science Quarterly* 42, 4 (1997), 716–749. https://doi.org/10.2307/2393655 03534 Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University].

[48] M. E. Helms, S. Vattam, and A. K. Goel. 2008. Compound analogical design, or how to make a surfboard disappear. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. 00013.

[49] D. R. Hofstadter and The Fluid Analogies Research Group. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York, NY.

[50] Keith J. Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT Press, Cambridge, MA. 02398.

[51] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating Innovation Through Analogy Mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 235–243. https://doi.org/10.1145/3097983.3098038

[52] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

[53] Jigsaw. 2018. Unintended Bias and Identity Terms. https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

[54] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Transactions on Computer-Human Interaction* (March 2022). https://doi.org/10.1145/3530013 Just Accepted.

[55] Craig A. Kaplan and Herbert A. Simon. 1990. In search of insight. *Cognitive Psychology* 22, 3 (1990), 374–419. https://doi.org/10.1016/0010-0285(90)90008-R

[56] Madeline K. Kneeland, Melissa A. Schilling, and Barak S. Aharonson. 2020. Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation. *Organization Science* 31, 3 (April 2020), 535–557. https://doi.org/10.1287/orsc.2019.1328 00011 Publisher: INFORMS.

[57] G. Knoblich, S. Ohlsson, H. Haider, and D. Rhenius. 1999. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25, 6 (1999), 1534–1555. https://doi.org/10.1037/0278-7393.25.6.1534 00691.

[58] Kenneth J. Kurtz, Olga Boukrina, and Dedre Gentner. 2013. Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39, 4 (July 2013), 1303–1310. https://doi.org/10.1037/a0031847

[59] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[60] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating Human-Language Model Interaction. http://arxiv.org/abs/2212.09746 arXiv:2212.09746 [cs] version: 2.

[61] J. Linsey, A. B. Markman, and K. L. Wood. 2008. WordTrees: A method for design-by-analogy. In *Proceedings of the 2008 ASEE Annual Conference*.

[62] J. S. Linsey, J. T. Murphy, A. B. Markman, K. L. Wood, and T. Kurtoglu. 2006. Representing analogies: Increasing the probability of innovation. In *Proceedings of IDETC/CIE 2006 ASME 2006 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*.

[63] J. S. Linsey, K. L. Wood, and A. B. Markman. 2008. Modality and representation in analogy. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 22, 02 (2008), 85–100. https://doi.org/10.1017/S0890060408000061

[64] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (Sept. 2023), 1–35. https://doi.org/10.1145/3560815

[65] J. Loewenstein. 2010. How One's Hook Is Baited Matters for Catching an Analogy. In *Psychology of Learning and Motivation*, B. H. Ross (Ed.). Vol. 53. San Diego, CA, 149–182.

[66] Stephen MacNeil, Sarah Abdellahi, Mary Lou Maher, Jin Goog Kim, Mohammad Mahzoon, and Kazjon Grace. 2019. Designing with and for the Crowd: A Cognitive Study of Design Processes in NatureNet. In *Design Computing and Cognition '18*, John S. Gero (Ed.). Springer International Publishing, Cham, 55–73. https://doi.org/10.1007/978-3-030-05363-5_4

[67] Stephen MacNeil, Zijian Ding, Ashley Boone, Anthony Bryce Grubbs, and Steven P. Dow. 2021. Finding Place in a Design Space: Challenges for Supporting Community Design Efforts at Scale. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–30. https://doi.org/10.1145/3449246

[68] Stephen MacNeil, Zijian Ding, Kexin Quan, Thomas j Parashos, Yajie Sun, and Steven P. Dow. 2021. Framing Creative Work: Helping Novices Frame Better Problems through Interactive Scaffolding. In *Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3450741.3465261

[69] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. https://doi.org/10.48550/arXiv.2301.06627 arXiv:2301.06627 [cs].

[70] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (Jan. 2013). http://arxiv.org/abs/1301.3781 00000 arXiv: 1301.3781.

[71] A. Newell and H. A. Simon. 1972. *Human problem solving*. Englewood Cliffs, NJ.

[72] Lior Noy, Yuval Hart, Natalie Andrew, Omer Ramote, Avi Mayo, and Uri Alon. 2012. A quantitative study of creative leaps, Mary Maher, Kristian Hammond, Alison Pease, {Rafael Pérez, Dan Ventura, and Geraint Wiggins (Eds.). 72–76. http://computationalcreativity.net/iccc2012/wp-content/uploads/2012/05/072-Noy.pdf

[73] T. Okada, S. Yokochi, K. Ishibashi, and K. Ueda. 2009. Analogical modification in the creation of contemporary art. *Cognitive Systems Research* 10, 3 (2009), 189–203. 00032.

[74] Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on Paid Crowdsourcing Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376677

[75] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–10. https://doi.org/10.1145/3411763.3450391

[76] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. http://arxiv.org/abs/2203.02155 arXiv:2203.02155 [cs].

[77] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P Dow. 2021. The" Active Search" Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 325–329.

[78] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P Dow. 2021. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[79] 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*. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3451760 00018.

[80] HWJ W. J. Rittel and MM M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences* 4 (1973), 155–169.

[81] Melissa Roemmele. 2016. Writing Stories with Help from Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (March 2016). https://doi.org/10.1609/aaai.v30i1.9810

[82] S. D. Savransky. 2000. *Engineering of creativity: Introduction to the TRIZ methodology of inventive problem solving.* Boca Raton, FL.

[83] Donald A. Schön. 1983. *The reflective practitioner: How professionals think in action.* Basic books.

[84] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339

[85] Pao Siangliulue, Joel Chan, Krzysztof Gajos, and Steven P. Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the ACM Conference on Creativity and Cognition.* https://doi.org/10.1145/2757226.2757230

[86] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction* (Feb. 2022), 3511599. https://doi.org/10.1145/3511599

[87] Latitude Team. 2020. World Creation by Analogy. https://aidungeon.medium.com/world-creation-by-analogy-f26e3791d35f 00000.

[88] Swaroop Vattam, Bryan Wiltgen, Michael Helms, Ashok K. Goel, and Jeannette Yen. 2011. DANE: Fostering Creativity in and through Biologically Inspired Design. In *Design Creativity 2010*, Toshiharu Taura and Yukari Nagai (Eds.). Springer London, 115–122. http://link.springer.com/chapter/10.1007/978-0-85729-224-7_16

[89] Michael S. Vendetti, Aaron Wu, and Keith J. Holyoak. 2014. Far-Out Thinking: Generating Solutions to Distant Analogies Promotes Relational Thinking. *Psychological Science* 25, 4 (April 2014), 928–933. https://doi.org/10.1177/0956797613518079 00066.

[90] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2022. Emergent Analogical Reasoning in Large Language Models. http://arxiv.org/abs/2212.09196 arXiv:2212.09196 [cs].

[91] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. http://arxiv.org/abs/2109.07684 arXiv:2109.07684 [cs].

[92] Stephen Wolfram. 2023. Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT. https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/

[93] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Distributed Analogical Idea Generation: Inventing with Crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).* ACM, New York, NY, USA, 1245–1254. https://doi.org/10.1145/2556288.2557371

[94] 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.* ACM, Helsinki Finland, 841–852. https://doi.org/10.1145/3490099.3511105

[95] Q. Zhu and J. Luo. 2022. Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society* 2 (May 2022), 1825–1834. https://doi.org/10.1017/pds.2022.185 Publisher: Cambridge University Press.

[96] Qihao Zhu, Xinyu Zhang, and Jianxi Luo. 2023. Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. *Journal of Mechanical Design* 145, 4 (Jan. 2023). https://doi.org/10.1115/1.4056598

APPENDIX

A SCREENSHOTS OF USER INTERFACE FOR STUDY 2

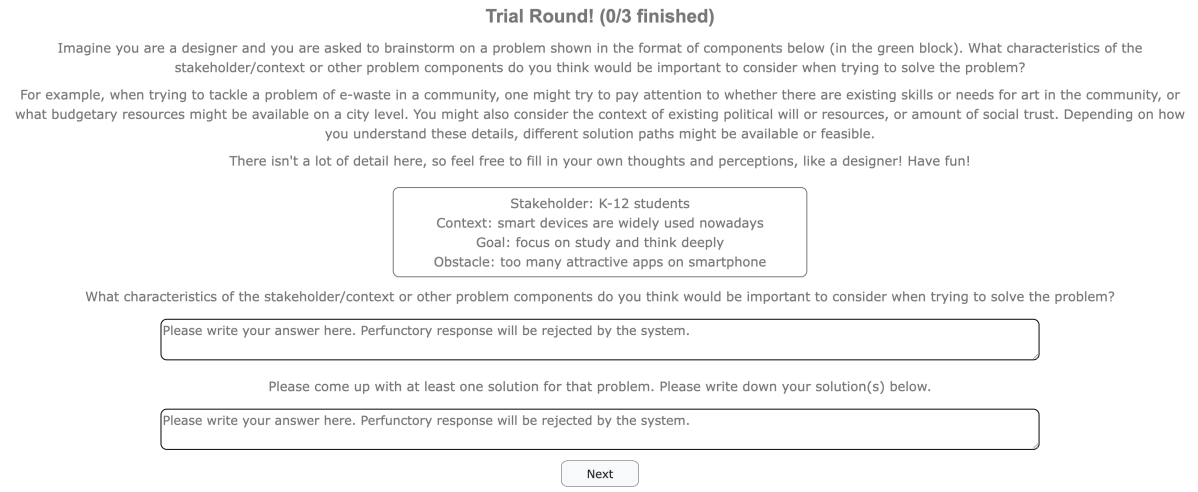

Figure A1: Screenshot of the initial formulation phase task interface and instructions.

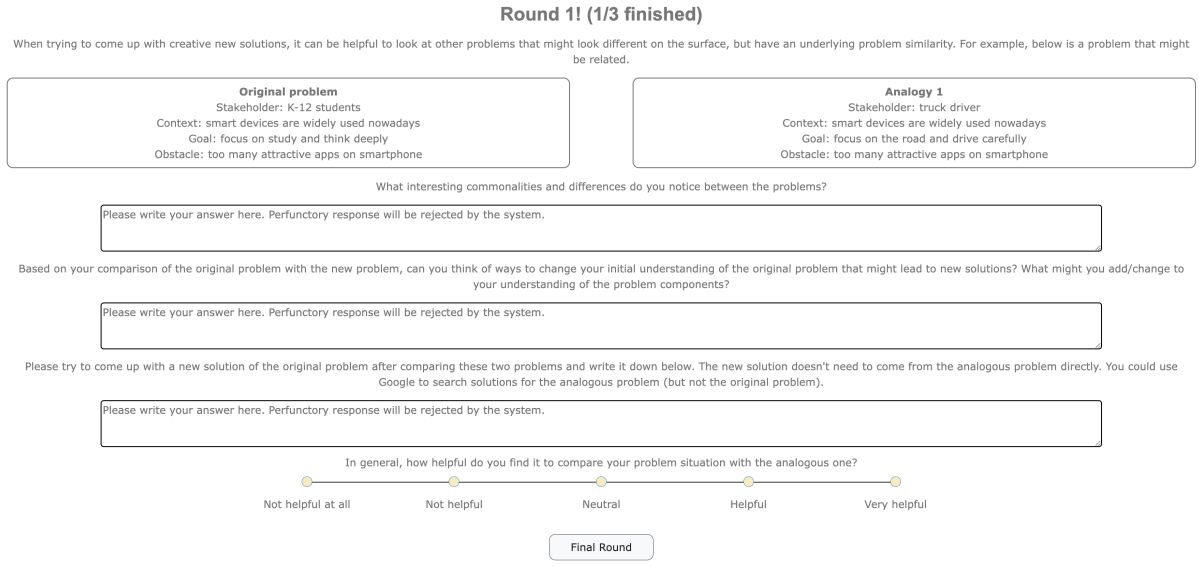

Figure A2: Screenshot of the analogical reformulation and ideation phase task interface and instructions.