Augmenting Scientific Creativity with an Analogical Search Engine

HYEONSU B. KANG, Carnegie Mellon University, USA
XIN QIAN, University of Maryland, College Park, USA
TOM HOPE, Allen Institute for AI and The University of Washington, USA
DAFNA SHAHAF, Hebrew University of Jerusalem, Israel
JOEL CHAN, University of Maryland, College Park, USA
ANIKET KITTUR, Carnegie Mellon University, USA

Analogies have been central to creative problem-solving throughout the history of science and technology. As the number of scientific papers continues to increase exponentially, there is a growing opportunity for finding diverse solutions to existing problems. However, realizing this potential requires the development of a means for searching through a large corpus that goes beyond surface matches and simple keywords. Here we contribute the first end-to-end system for analogical search on scientific papers and evaluate its effectiveness with scientists' own problems. Using a human-in-the-loop AI system as a probe we find that our system facilitates creative ideation, and that ideation success is mediated by an intermediate level of matching on the problem abstraction (i.e., high versus low). We also demonstrate a fully automated AI search engine that achieves a similar accuracy with the human-in-the-loop system. We conclude with design implications for enabling automated analogical inspiration engines to accelerate scientific innovation.

ACM Reference Format:
Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Trans. Comput.-Hum. Interact.* 1, 1, Article 1 (January 2022), 36 pages. https://doi.org/10.1145/3530013

1 INTRODUCTION

Analogical reasoning has been central to creative problem solving throughout the history of science and technology [32, 43, 50, 54, 60, 86]. Many important scientific discoveries were driven by analogies: the Greek philosopher Chrysippus made a connection between observable water waves and sound waves; an analogy between bacteria and slot machines helped Salvador Luria advance the theory of bacterial mutation; a pioneering chemist Joseph Priestly suggested charges attract or repel each other with an inverse square force by an analogy to gravity.

Today the potential for finding analogies to accelerate innovation in science and engineering is greater than ever before. As of 2009 fifty million scientific papers had been published, and the number continues to grow at an exceedingly fast rate [12, 28, 68, 85]. These papers represent a

Authors' addresses: Hyeonsu B. Kang, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, hyeonsuk@cs.cmu.edu; Xin Qian, xinq@umd.edu, University of Maryland, College Park, College Park, MD, USA, 20742; Tom Hope, tomh@allenai.org, Allen Institute for AI and The University of Washington, Seattle, WA, USA; Dafna Shahaf, dshahaf@cs.huji.ac.il, Hebrew University of Jerusalem, Israel; Joel Chan, joelchan@umd.edu, University of Maryland, College Park, College Park, MD, USA, 20742; Aniket Kittur, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA, nkittur@cs.cmu.edu.

1073-0516/2022/1-ART1
https://doi.org/10.1145/3530013

potential treasure trove for finding inspirations from distant domains and generating creative solutions to challenging problems.

However, searching analogical inspirations in a large corpus of papers remains a longstanding challenge [34, 44, 83, 99]. Previous systems for retrieving analogies have largely focused on modeling analogical relations in non-scientific domains and/or in limited scopes (e.g., structure-mapping [36–38, 42, 106], multiconstraint-based [33, 59, 65], connectionist [57], rule-based reasoning [3, 15, 16, 111] systems), and the prohibitive costs of creating highly structured representations prevented hand-crafted systems (e.g., DANE [65, 110]) from having a broad coverage of topics and being deployed for realistic use. Conversely, scalable computational approaches such as keyword or citation based search engines have been limited by a dependence on surface or domain similarity. Such search engines aim to maximize similarity to the query which is useful when trying to know what has been done on the problem in the target domain but less useful when trying to find inspiration outside that domain (for example, for Salvador Luria's queries: "how do bacteria mutate?" or "why are bacterial mutation rates so inconsistent?", similarity maximizing search engines may have found Luria and Delbrück's earlier work on E.coli [81] but may have failed to recognize more distant sources of inspiration such as slot machines as relevant).

Recently a novel idea for analogical search was introduced [61]. In this idea what would otherwise be a complex analogical relation between products is pared down to just two components: purpose (*what problem does it solve?*) and mechanism (*how does it solve that problem?*). Once many such purpose and mechanism pairs are identified, products that solve a similar problem to the query but using diverse mechanisms are searched to help broaden the searcher's perspective on the problem and boost their creativity for coming up with novel mechanism ideas. Anecdotal evidence suggests that this approach may also be applicable to the domain of scientific research. For example, while building lighter and more compact solar panel arrays has been a longstanding challenge for NASA scientists, recognizing how the ancient art form of origami may be applied to create folding structures led to an innovation to use compliant mechanisms to build not just compact but also self-deployable solar arrays [27, 89, 119] (diagrammatically shown in fig. 1). The first remaining challenge of analogical search in the scholarly domain is how we might represent scientific articles as purpose and mechanism pairs at scale and search for those that solve similar purposes using different mechanisms. Recent advances in natural language processing have demonstrated that neural networks that use pre-trained embeddings to encode input text can offer a promising technique to address it. Pre-trained embeddings are real-valued vectors that represent tokens (*Tokenization* means breaking a piece of text into smaller units; *Tokens* can be words, characters, sub-words, or n-grams.), in a high-dimensional space (e.g., typically dimensions of a few dozens to a few thousands) and are shown to capture rich, multi-faceted semantic relations between words [8, 100]. Leveraging them, neural networks may be trained to identify purposes and mechanisms from text [61, 62] to enable search-by-analogy (i.e. different mechanisms used for similar purposes). Once candidate papers are retrieved, searchers may use them to come up with novel classes of mechanisms or apply them directly to their own research problems to improve upon the current state. Prior studies in product ideation showed that users of analogical search systems could engage with the results to

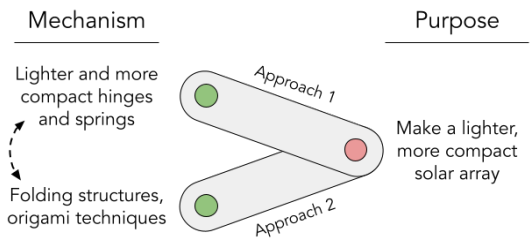

Fig. 1. A diagram of two different yet analogical approaches (dashed arrow) for building lighter and more compact solar arrays, and their representations in purposes and mechanisms.

engender more novel and relevant ideas [21, 48, 74]. Here, we study the remaining open questions as to whether such findings also generalize to the scientific domains of innovation and how they may differ.

In this paper we present a functioning prototype of an analogical search engine for scientific articles at scale and investigate how such a system can help users explore and adapt distant inspirations. In doing so our system moves beyond manually curated approaches that have limited data (e.g., crowdsourced annotations in [21] with ~2000 papers) and machine learning approaches that have been limited to simple product descriptions [48, 61, 62]. Using the prototypical system, we explore how it enables scientists to interactively search for inspirations for their personalized research problems in a large (~1.7M) paper corpus. We investigate whether scientists can recognize mapping of analogical relations between the results returned from our analogical search engine and their query problems, and use them to come up with novel ideas. The scale of our corpus allows us to probe realistic issues including noise, error, and scale as well as how scientists react to a search engine that does not aim to provide only the most similar results to their query.

In order to accomplish these goals we describe how we address several technical issues in the design of an interactive-speed analogical search engine, ranging from developing a machine learning model for extracting purposes and mechanisms in scientific text at a token level granularity, the pipeline for constructing a *similarity space* of purpose embeddings, and enabling these embeddings to be queried at interactive speeds by end users through a search interface. We construct the similarity space by putting semantically related purpose embeddings in close indices from each other such that related purposes can be searched at scale.

In addition to the technical challenges there are several important questions around the design of analogical search engines that we explore here. A core conceptual difference that distinguishes analogical search engines from other kinds is that the analogs they find for a search query need to maintain some kind of distance from the query, rather than simply maximizing the similarity with it. However, only certain kinds of distance may support generative ideation while others have a detrimental effect. Another question remains as to how much distance is appropriate when it comes to finding analogical inspirations in other domains. While landmark studies of analogical innovation suggest that highly distant domains can provide particularly novel or transformative innovations [46, 47, 55], recent work suggests the question may be more nuanced and that intermediate levels of distance may be fruitful for finding ideas that are close enough to be relevant but sufficiently distant to be unfamiliar and spur creative adaptation [22, 39, 49]. Using a concrete example from one of our participants who studied ways to facilitate heat transfer in semiconductors, a keyword search engine might find commonly used mechanisms appropriate for direct application (e.g., tweaking the composition of the material) while an analogical search engine might find similar problems in more distant domains which suggest mechanisms that inspire creative adaptation (e.g., nanoscale fins that absorb heat and convert it to mechanical energy). Though more distant conceptual combinations may not always lead to immediately feasible or useful ideas, they may result in outsized value after being iterated on [9, 23, 75].

In the following sections we explore the technical and design challenges for an analogical search engine and how users interact with such a system. First, we describe the development of a human-in-the-loop search engine prototype, in which most elements of the system are functional but human screeners are used to remove obvious noise from the end results in order to maximize our ability to probe how users interact with potentially useful analogical inspirations. Using this prototype we characterize how researchers searching for inspirations for their own problems gain the most benefit from papers that partially match their problem (i.e., match at a high level purpose but mismatch at a lower level specifications of the purpose), and that the benefits are driven not by direct application of the ideas in the paper but by creative adaptation of those ideas to their

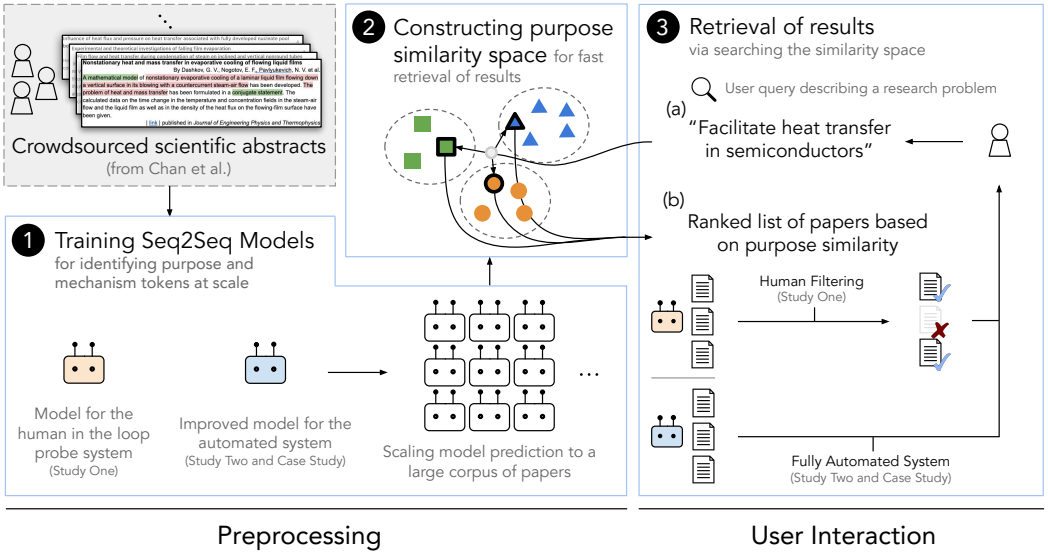

Fig. 2. Components of our system design that address the three core challenges. ① Purpose and mechanism tokens are extracted from paper abstracts at scale. We develop sequence-to-sequence classifiers to classify tokens into purpose, mechanism, or neither, going beyond previous approaches that worked on sentences or relied on crowds. ② We embed the extracted purpose texts using a pre-trained language model (Google's Universal Sentence Encoder (USE) [20]) and train a tree-based index of vectors to place high semantic similarity vectors in close neighborhoods for efficient lookup. ③ When the user query arrives at the system, it is first embedded with USE. This query embedding is then used to lookup the pre-computed tree indices for high similarity purpose embeddings. Paper abstracts for the corresponding purpose embeddings are retrieved from Google Datastore. In the first system, additional human filtering is performed to remove obviously irrelevant results that may have been included due to model errors. Finally, a set of papers with similar purposes to the query but different mechanisms are returned to the users for ideation.

target domain. Subsequently we describe improvements to the system to enable a fully automated, interactive-speed prototype and case studies with researchers using the system in a realistic way involving reformulation of their queries and self-driven attention to the results. We synthesize the findings of the two studies into design implications for next-generation analogical search engines.

Through extensive in-depth evaluations using an ideation think-aloud protocol [35, 107] with PhD-level researchers working on their own problems, we evaluate the degree to which inspirations spark creative adaptation ideas in a realistic way on scientists' own research problems. Unlike previous work which has often used undergraduate students in the classroom or lab [110], and often evaluated systems on pre-determined problems [40], this study design provides our evaluation with a high degree of external validity and allows us to deeply understand the ways in which encountering our results can engender new ideas. Our final, automated search engine demonstrates how the human-in-the-loop filtering can be removed while achieving a similar accuracy. We conclude with the benefits, design challenges, and opportunities for future analogical search engines from case studies with several researchers. To encourage innovation in this domain, we release our corpus of purpose and mechanism embeddings[1].

2 SYSTEM DESIGN

[1]https://github.com/hyeonsuukang/augmenting_tochi22

The design of our analogical search engine for scientific papers involves three main system requirements. First, a computational pipeline for automatically identifying purposes (*what problems does it solve?*) and mechanisms (*how does it solve those problems*) at scale (e.g., millions of papers), in a token-level granularity from scientific abstracts. Second, an efficient retrieval algorithm for incorporating the identified purpose and mechanism texts into the system to enable search-by-analogy (i.e. paper abstracts that contain similar purposes to a query problem but different mechanisms). Third, end-user interactivity for querying problems of interest (e.g., "transfer heat in semiconductors," "grow plants using nanoparticle fertilizers"). We describe the system design in detail in the following subsections.

Kind (# of papers)	Avg. length	# of PP	# of MN
Train (2021)	196	65261	120586
Validation (50)	170	1510	1988

Table 1. Summary statistics of the training and validation datasets: the number of purpose (PP) and mechanism (MN) tokens, the number and avg. token length of paper abstracts.

Domain	CS	Eng	BioMed	B & Eng	Total
Count	675K	568K	336K	145K	1.7M

Table 2. Corpus used in the deployed search engine and its topical distribution: Computer Science (CS), Engineering (Eng), Biomedicine (BioMed), and Business and Engineering (B & Eng).

2.1 Stage One. Training Seq2Seq models for identifying purpose and mechanism tokens

2.1.1 Overview of Modeling. In the first stage of the system, purpose and mechanism tokens are identified from paper abstracts (fig. 2, ①). Research paper abstracts often include descriptions of the most important purpose or *the core problem addressed in a paper* and the proposed mechanism or *the approach taken to address the problem*, making them good candidates for identification and extraction of tokens corresponding to them. For example, for a similar problem of facilitating heat transfer, Paper A may propose an approach that modifies the structure of the material used at the interface between crystalline silicon (semiconductor material) and the substrate, while Paper B may propose a more distant mechanism (due to the mismatch on scale) of fin-based heat sinks commonly used for electronic devices. The goal of this first stage is to automatically identify and extract tokens that correspond to the similar purpose (e.g., 'facilitate heat transfer') as well as the mechanisms (e.g., 'modifying the structure of the material used at the interface between crystalline silicon' vs. 'fin-based heat sinks') from the abstract A and B.

One relevant automated approach for identifying purposes and mechanisms from scientific abstracts is DISA [63], which formulates the task as supervised sentence classification. However, we found that many key sentences in abstracts include both purpose and mechanism, breaking the assumptions of a sentence-level classifier (e.g., "In this paper, [*a wavelet transforms based method*] for [*filtering noise from images*] is presented."). To overcome this limitation we follow [62] and frame purpose and mechanism identification as a sequence-to-sequence (Seq2Seq) learning task [5, 101] and develop deep neural networks with inductive biases capable of learning token-level patterns in the training dataset. Our dataset consists of crowdsourced annotations from Chan et al. (the dataset is constructed via application of [21] to a larger corpus of around 2000 paper abstracts largely in computer science domains) (table. 1). We train the models to classify input features (tokens or spans of tokens) as either purpose (PP), mechanism (MN) or neither.

We train two deep neural networks (Model 1 and 2), achieving increasing accuracy of classification. The first model is based on a Bi-directional LSTM (BiLSTM) architecture for sequence tagging [56, 64], in which the forward (the beginning of the sequence to the end) and the backward passes

condition each token position in the text with its left and right context, respectively. A main source of improvement of Model 2 over Model 1 is the ability to more selectively attend to informative tokens in a sentence rather than treating each token in a sequence as independent of each other (as a hypothetical example, an extremely effective model based on this approach may assign more weights to the tokens 'selectively attend to informative tokens', as they represent the core mechanism described in the previous sentence) and to leverage the regularities of co-occurrence with surrounding words through the self-attention mechanism [109].

2.1.2 Seq2Seq Model Implementation Details. We implement the BiLSTM architecture of Model 1 in PyTorch [87]. We use pre-trained GloVe [88] word embeddings with 300 dimensions, consistent with prior work [11, 78, 88] to represent each token in the sequence as 300-dimensional input vectors for the model. We train the model with a cross entropy loss objective for per-token classification in the three (PP, MN, Neither) token classes.

For Model 2, we adapt the SpanRel [67] architecture and implement it on AllenNLP [41]. We implement a self attention mechanism that tunes weights for the core word in each span as well as the boundary words that distinguish the context of use, consistent with [79]. We use the pre-trained ELMo 5.5B [90] embeddings for token representation following the near state-of-the-art performance reported in [67] on the scientific Wet Lab Protocol dataset. We train the model using a similar procedure as Model 1. We leave detailed training parameters for Model 1 and 2 to the Appendix.

2.1.3 Introducing Human-in-the-loop Filtering for Model 1. The final classification performance (F1-scores) of Model 1 on the validation set is 0.509 (Purpose), 0.497 (Mechanism), and 0.801 (neither). We found that the limited accuracy contributed to how the system retrieves irrelevant search results. Because reactions to obviously irrelevant results are not useful, we added a human-in-the-loop [31] filtering stage. The filtering proceeded as follows: members from the research team inputted problem queries received from study participants into the system. Once the model produced matches, they went over from the top of the sorted list and removed only those that are irrelevant to the problem context. They continued filtering until at least 30 papers with reasonable purpose similarity were collected. After Winsorizing at top and bottom 10% [116], the human filterers reviewed 45 papers per query (SD: 27.6, min: 6, max: 138) for 5 queries (SD: 2.4, min:2, max: 9) to collect 33 (SD: 3.5, min: 30, max: 40) purpose-similar papers (about 12/45 = 26% error rate). In Study 1 we show that the limited retrieval accuracy of Model 1 is sufficient for use as a probe with this additional human-in-the-loop filtering. In Study 2 and case studies, we demonstrate how this filtering can be removed with Model 2 while achieving a similar accuracy.

2.1.4 Scaling Model Inference. In order to have sufficient coverage to return diverse results, we collected an initial corpus of 2.8 million research papers from Springer Nature[2]. After deduplication (based on Digital Object Identifier using BigQuery[3]) and filtering only papers with at least 50 characters in the abstract we were left with 1.7 million papers in four subjects (Table 2). We stored the resulting corpus in Google Cloud storage buckets[4]. To scale the classification of the Seq2Seq models we used the Apache Beam API[5] on Google Cloud Dataflow[6] to parallelize the operation.

[2]https://dev.springernature.com/
[3]https://cloud.google.com/bigquery
[4]https://cloud.google.com/storage
[5]https://beam.apache.org/
[6]https://cloud.google.com/dataflow/

2.2 Stage Two. Constructing a purpose similarity space

2.2.1 Overview. In the second stage, the identified purpose texts are incorporated into the system to enable search-by-analogy of papers that solve similar problems using different mechanisms, at an interactive speed (fig. 2, ②). Relevant previous approaches include Hope et al. [61] which first clusters similar purposes (through k-means with pruning) and subsequently samples within each cluster of similar purposes to maximize the diversity of mechanisms (via a GMM approximation algorithm [92]), or [62] which employs similarity metrics to balance the *similarity* to a purpose query and the *distance* to a mechanism query (and vice versa). In contrast, from pilot tests in our corpus we discovered that even close purpose matches of scientific papers already had high variance in terms of the mechanisms they propose. We hypothesize that this may be the case due to the enormous span of possible research topics and the relative sparseness of their coverage in our corpus, and/or due to the emphasis on novelty in scientific research that discourages future papers which might contribute relatively small variations to an existing mechanism. We leave exploration of these hypotheses for future work and simplify our sampling of the scientific papers to the one based solely on the similarity of purpose, sufficient for ensuring diversity.

In order to support fast retrieval (e.g., sub-second response time) of papers with similar purposes at scale (e.g., millions of papers), we pre-train Spotify's Annoy[7] indices of nearest neighboring purposes. Annoy trains a neural network to assign an embedding vector corresponding to a purpose an index in the high-dimensional space that brings it close to other indices of purpose vectors that have similar meaning (see §2.2.3 for details of the metric used for the similarity of meaning). Annoy uses random projection and tree-building (see [1, 2]) to create read-only, file-based indices. Because it decouples creation of the static index files from lookup, it enables efficient and flexible search by utilizing many parallel processes to quickly load and map indices into memory.

2.2.2 Interactive Speed. Additionally Annoy minimizes its memory footprint in the process. This efficiency, critical for real-time applications such as ours, was further validated during our test of the end-to-end latency on the Web, with the average response taking 2.4s (SD = 0.56s)[8]. The level of latency we observed was sufficiently low to enable interactive search by end users (both human-in-the-loop filterers in Study One and researcher participants in case studies).

2.2.3 Implementation Details. To construct the similarity space, we first encode the purpose texts into high-dimensional embedding vectors which then can be used to compute pairwise semantic similarity. Here, the choice of an encoding algorithm depends on three main constraints. First, the pairwise similarity, when computed, should correlate well with the human-judged semantic similarity between the purposes. Second, similarity calculation between varying lengths of texts should be possible because extracted purposes can differ in length. Third, computationally efficient methods are preferred for scaling. To meet these requirements, we chose Universal Sentence Encoder (USE)[9] to encode purposes into fixed 512-dimensional vectors. Universal Sentence Encoder trains a transformer architecture [109] on a large corpus of both unsupervised (e.g., Wikipedia) and supervised (e.g., Stanford Natural Language Inference dataset [13]) data to produce a neural network that can encode text into vectors that meaningfully correlate with human judgment (e.g., evaluated on the semantic textual similarity benchmark [19]). USE can handle texts of varying lengths (e.g., from short phrases to sentences to paragraphs), and with high efficiency [20], thereby making it suitable for our system.

[7] https://github.com/spotify/annoy

[8] We tested with 20 topically varied search queries that have not previously been entered to the engine to test the latency end-users experience and to exclude the effect of caching from it.

[9] https://tfhub.dev/google/universal-sentence-encoder-large/5

We pre-compute pairwise similarity of the purpose embeddings and store the indices in neighborhoods of high similarity for fast retrieval of similar purposes. As mentioned before, we train the Annoy indices on Google Cloud AI Platform[10]. We use 1 - the Euclidean distance of normalized vectors (i.e., given two vectors $\mathbf{u}$ and $\mathbf{v}$, $\text{distance}(\mathbf{u}, \mathbf{v}) = \sqrt{(2\,(1 - \cos\,(\mathbf{u}, \mathbf{v})))}$) as a similarity metric (using a Euclidean distance based metric for nearest neighbor clustering shows good performance, see [4] for a related discussion on the impact of the distance metric on the retrieval performance). We set the hyper-parameter k specifying the number of trees in the forest to 100 (larger k's result in more accurate results but also decreases performance; see [2] for further details). Empirically, 100 seemed to strike a good balance between the precision-performance trade-off, thus we did not experiment with this parameter further.

2.3 Stage Three. Retrieving the results

In the last stage, the front-end interface interacts with end users and receives problem queries. These queries are then relayed to the back-end for retrieval of papers that solve similar problems using different mechanisms. The retrieved papers are presented on the front-end for users to review (fig. 2, ③). When a user query is received, the back-end first encodes it using the same encoding algorithm used as the construction method of the purpose similarity space (i.e. Universal Sentence Encoder). Using this query embedding, the back-end searches the pre-trained similarity space for papers with similar purposes. The papers with high purpose similarity are then returned to and displayed on the front-end. We describe the actual interfaces used in the studies in the corresponding design sections (§3.2.4, §3.2.5).

Together the design of our system enabled what is to our knowledge the first functioning prototype of an interactive analogical search engine for scientific papers at scale. In the following sections we report on how such a search engine can help researchers find analogical papers that facilitate creative ideation.

3 STUDY 1: CREATIVE ADAPTATION WITH A HUMAN-IN-THE-LOOP ANALOGICAL SEARCH ENGINE

In Study 1 we set out to establish the viability of an analogical search engine using a human-in-the-loop probe in the domain of scholarly recommendations. We investigate whether analogical search returns a distinct and novel set of papers compared to keyword search results, and capture participants' reaction to each result in a randomized order, blind to condition. To deeply understand the process of ideation using analogical papers we ask participants to come up with new ideas for their own research projects after reviewing each paper. Using this data we code ideation outcomes in depth to explore the various ways in which analogical distance can shape ideation outcomes, such as inspiring direct transfer of solutions, or sparking adaptation of ideas into novel combinations.

3.1 Coding ideation outcomes

We are interested in studying whether an analogical search engine provides distinctive and complementary value to other commonly used search approaches that rely on surface similarity. In particular, our focus is on the inspirational value rather than the immediate relevance of search results or the direct usefulness of solutions. The highest value of creative inspiration often comes from creatively adapting ideas to reformulate a problem and recognizing new bridges to previously unknown domains that open up entirely new spaces of ideas. For example, recognizing a connection from the ancient art form of origami to fold intricate structures with paper and building a

[10]https://cloud.google.com/ai-platform

sufficiently compact, deployable solar panel arrays and radiation shields led NASA to hire origami experts [27, 89, 119].

Our approach to measuring ideation outcome is through the use of a quaternary variable categorizing the types of ideation. To capture the inspirational value of analogical search and move beyond the measurements focused on the immediate relevance or the direct usefulness we distinguish the Creative Adaptation and Direct Application types of ideation. In our studies these two types corresponded to think-alouds that resulted in novel ideas whereas the rest (Background and None) corresponded to think-alouds in which no new ideas were produced.

• **Creative Adaptation:** Novel mechanism ideas that involve substantial adaptation of the information provided in the paper. These ideas are typically associated with a higher uncertainty of success due to the less familiar nature of the domains involved.

• **Direct Application:** More directly applicable ideas that involve less adaptation than Creative Adaptation. These ideas are typically associated with a lower uncertainty of success because researchers are more familiar with the domains.

• **Background:** The information provided in the paper is good for background reading (e.g., to learn about other domains).

• **None:** Did not result in new ideas nor was useful for background reading.

Creative Adaptation ideas generally involved a substantial amount of adaptation, while Direct Application ideas were closer to the source domain and more directly applicable. For example, using the data from one of our participants, applying the techniques for manipulating thermal conductance at solid-solid interfaces was considered a direct application idea for P1 (fig. 3, left) because he was familiar with the concept of controlling the interfacial thermal conductivity given the relevant approaches he developed in his current and past research projects. Thus the connections to the source problem were directly recognizable. On the other hand, creating a fin-based wall structure for heat transfer was an example of creative adaptation idea (fig. 3, right) because of its novelty and the participant's unfamiliarity in relevant domains. The unfamiliarity and uncertainty was generally more associated with analogs for creative adaptation than direct application. On the other hand, the unfamiliarity also sometimes acted as a barrier to participants' openness and subsequent ideation. Though challenging, in order to recognize novel connections to the source problem the participants may need to suspend their early rejection of a seemingly foreign idea and its surface-level mismatches and engage in deeper processing which could lead to re-imagination and re-formulation of the research problem at hand. To code the Creative Adaptation and Direct Application types of ideation outcomes, the coders took into consideration different linguistic and contextual aspects of the descriptions of the ideas and their think-aloud process (details in §3.2.3).

3.2 Design of the study

3.2.1 Participants. We recruited eight graduate (four women) researchers in the fields of sciences and engineering via email advertisement at a private R1 U.S. institution. Four were senior PhD students (3rd year or above and one recently defended their thesis) and the rest was 2nd year or below. Disciplinary backgrounds of the participants included: Mechanical (3), Biomedical (2), Environmental (1), Civil (1), and Chemical Engineering (1). Once a participant signed up for the study, we asked them to describe their research problems and send the research team search queries they use to look for inspirations on popular search engines such as Google Scholar[11]. Members of the research team screened papers with relevant purposes using these queries on the filtering interface (fig. 4, left). Despite our efforts to collect papers over diverse topical areas, the search engine did not contain enough papers for two of the participants who work on relatively novel

[11]https://scholar.google.com/

fields (e.g., "machine learning methods of 3D bioprinting"). These participants were interviewed on their current practices for reviewing prior works and coming up with new ideas for research and were not included in the subsequent analyses.

3.2.2 Study Procedure and Keyword-search Control. The rest of the participants were then invited to in-person interviews. To ensure that participants would be exposed to a sufficiently diverse set of analogical mechanisms and to maximize our power to observe the ideation process, we generated a list of top 30 results from the analogical search engine using the search queries provided by the study participants. As a control condition we also included top 15 results from a keyword-based search engine using the standard Okapi BM25 algorithm [82] ($k_1 = 1.2, b = 0.75$) using the same search queries as the analogical search engine. The order of results in the list was randomized and participants were blind to condition. To account for the difference in the quantity of exposure in the analysis, we normalized the ideation outcomes by the number of results returned in each condition. Using this list we employed a think-aloud protocol [80, 108] in which participants were presented with the title, abstract, and other metadata of papers and asked to think aloud as they read through them with the goal of generating ideas useful for their research using our Web-based interface (fig. 4, right). Although time consuming, this approach allowed us to capture rich data on participants' thought process and how those processes changed and evolved as participants considered how a paper might relate to their own research problems. In addition, we asked the participants to make a judgment on the novelty of each paper on a 3-point Likert-scale. After participants finished reviewing the 45 papers, we interviewed them about their overall thoughts on the results' relevance and novelty and whether there were any surprising or unique results. Each interview lasted about one and a half hours and the participants were compensated $15/hr for their participation.

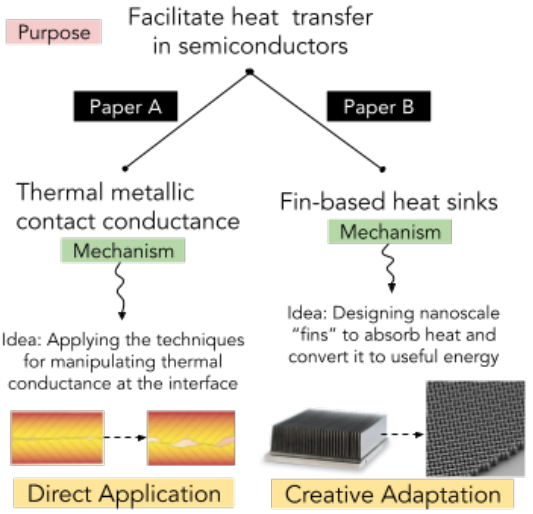

Fig. 3. Example papers for the purpose of facilitating heat transfer heat in semiconductors. (Top) A Direct Application paper involves directly applicable ideas and techniques for manipulating the interface material and structure to control thermal conductance. (Bottom) A Creative Adaptation example involves transferring a distant idea (fin-based design for heat sinks) and creatively adapting it into the target problem context (designing nano-scale fins that could absorb heat and convert it to useful energy). Figure credits: contact configurations and interface resistance from [117], fin-based heat sink from [104], nano-fins from [94].

3.2.3 Data and Coding. In total, our data consisted of 267 paper recommendations for six participants and their Likert-scale questionnaire responses measuring the content novelty, after removing 3 within-condition duplicates (these papers included cosmetic changes such as different capitalization in the title or abstract). One participant ran out of time towards the end of the interview and only provided novelty measures for the last 17 paper recommendations in the randomized list. Thus, 250 transcripts of participants' think-aloud ideation after reading each paper were used for analyzing ideation outcomes. To code the distance between the Creative Adaptation and Direct Application types of ideation outcomes, the coders took into consideration (1) the verbs used to

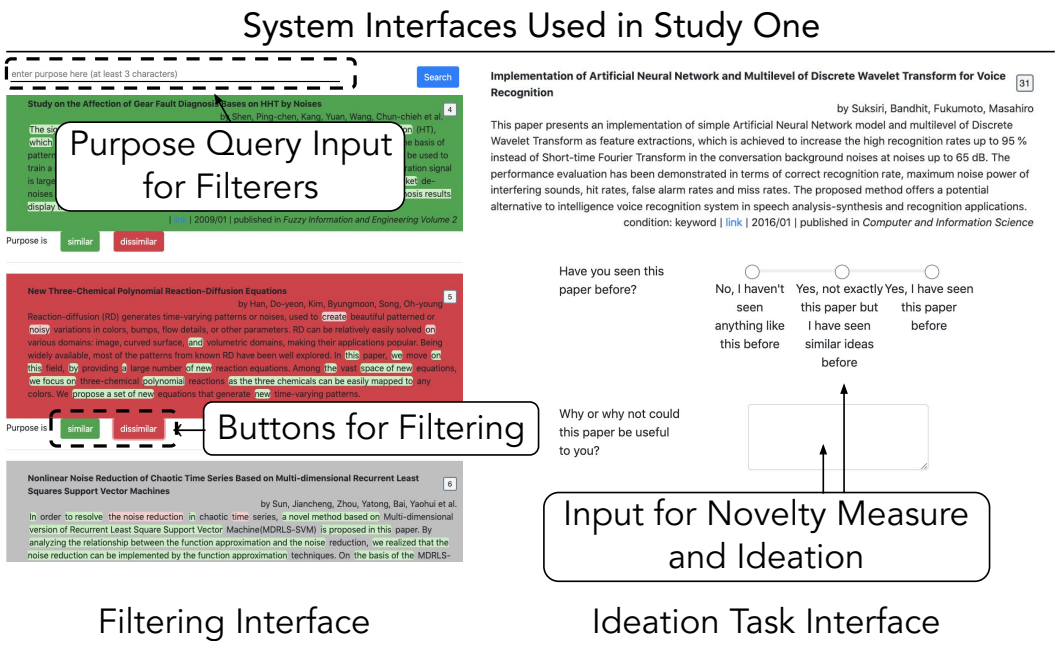

Fig. 4. The front-end interfaces. (Left) Human reviewers used this filtering interface to input search queries received from the participants and remove papers with obviously irrelevant purposes. To assist the reviewers' filtering process, model predicted purpose (e.g., *the noise reduction* and *time*, highlighted in red at the bottom of the filtering interface) and mechanism (highlighted in green) tokens were also provided along with the title and the abstract text. The background color turned green when the "Similar" button is clicked and red when the "Dissimilar" button is clicked. (Right) The ideation task interface was populated with a list of human filtered papers for review by the participants in Study 1 (the order of papers was randomized).

describe the ideas (e.g., 'design', 'develop', or 'invent' were generally associated more with distant ideas compared to 'apply', 'use', 'adopt'; see Table. 3); (2) the context of ideas such as participants' expression of unfamiliarity or uncertainty of the domain involved (e.g., "I'm not really sure" vs. "I'm familiar with this domain"); and (3) participants' perceived immediacy of the idea's applicability (i.e., ideas perceived by participants as more immediately applicable were associated with direct application but not creative adaptation ideas). Two of the authors coded a fraction of the data together (13/250, 5.2%) and then independently coded the rest blind-to-condition, using the four ideation outcomes types described in §3.1 and with the following protocol: The coders first judged the existence of an idea. If there was, then its type was further distinguished between Creative Adaptation and Direct Application using the linguistic and contextual descriptions described above (e.g., Creative Adaptation ideas were more frequently associated with the 'design' words, higher unfamiliarity and uncertainty of the domains, and less immediate applicability, compared to Direct Application ideas). In case there was no concrete idea in the data, coders further distinguished between the Background vs None cases.

The agreement between coders was significant, with Cohen's $\kappa = 0.89$ (near perfect agreement) for the four categories of ideation outcome. Given the high level of agreement between the coders, any disagreements were resolved via discussion on a case-by-case basis.

3.2.4 Apparatus 1: the human-in-the-loop filtering interface. In Study 1, members of the research team first received search queries from study participants and reviewed the model-produced purpose matches to filter irrelevant papers using a filtering interface (fig. 4, left). This additional step was introduced to ensure that papers with obviously dissimilar purposes are not returned to study participants. Reviewers determined whether each paper contained a clearly irrelevant purpose in which case it was removed by clicking the *Dissimilar* button at the bottom of the paper. On the other hand when the *Similar* button was clicked it turned the background of the paper green in the interface and increased the number of the papers collected so far. Reviewers continued the screening process until at least 30 papers with reasonable purpose similarity were collected.

3.2.5 Apparatus 2: the ideation task interface. The filtered papers were then displayed as a randomized list of papers to study participants (fig. 4, right). In addition to the content and metadata of papers (e.g., authors, publication date, venue, etc.), each paper was presented with a Likert-scale question for measuring content novelty and a text input for ideation.

3.2.6 Limitations. To reduce potential biases, our coders were blind to experimental conditions and relied on participants' statements of ideas' novelty and usefulness (e.g., "I've never seen something like this before," "this is not a domain I would've searched if I used Google Scholar"), and achieved a high inter-rater reliability. We believe coders had a reasonable understanding of how participants arrived at specific ideas from descriptions of their current and past research topics, think-alouds, and end-of-experiment discussions. Despite this, we also acknowledge the limitations of this approach and discuss how future research may improve upon it (see §7.2.1).

3.2.7 On reporting the results. We report the result of our studies below. To denote statistical significance we use the following notations: $^{*}(\alpha = 0.05)$, $^{**}(\alpha = 0.01)$, $^{***}(\alpha = 0.001)$, $^{****}(\alpha = 0.0001)$. Alpha levels were adjusted when appropriate in post-hoc analyses using Bonferroni correction.

3.3 Result

Finding novel papers for creative ideas. Our key measure of success is how paper recommendations from the analogy search engine (hereinafter *analogy papers*) help scientists generate creative ideas for their own research problems. To this end, we investigate a) whether analogy papers are novel and complementary to the papers found from the keyword-search baseline (hereinafter *keyword papers*) and b) whether analogy papers resulted in more creative adaptation ideas than direct application ideas in ideation.

3.3.1 Analogy papers differed from keyword papers and were judged more novel. The viability of our approach is based on the assumption that the analogy search pipeline returns a different distribution of results than a keyword-based baseline. This assumption appeared to hold true: the keyword-search and analogy-search conditions resulted in almost completely disjoint sets of paper recommendations. Out of the total 267 papers, the overlap between analogy and keyword papers was only one. Analogy papers appeared to represent a complementary set of results users would be unlikely to encounter through keyword-based search.

To further examine this assumption we had participants rate the novelty of the results by asking them "*have you seen this paper before?*" on a 3-point Likert scale response options of 1: "*Yes, I have seen this paper before*", 2: "*Yes, not exactly this paper but I have seen similar ideas before*", and 3: "*No, I have not seen anything like this before*". Participants found papers recommended in the analogy condition to contain significantly more novel ideas (2.7, SD: 0.48) compared to the keyword condition (2.3, SD: 0.55) (Welch's two-tailed t-test, $t = -5.53, p = 1.33 \times 10^{-7}$) (fig. 5, left).

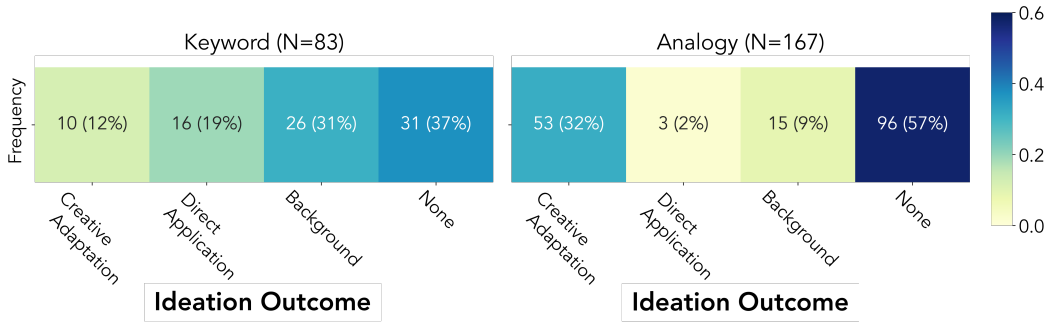

Fig. 6. Frequency of the ideation outcome types by condition. Darker colors represent higher rates. Creative adaptation is 5.3 times more frequent among analogy papers (53 in Analogy vs. 10 in Keyword), while most of direct application is from keyword papers (3 in Analogy vs. 16 in Keyword). The distributions differed significantly (chi-squared test, $\chi^2(3) = 52.12, p < 1.0 \times 10^{-10}$ overall and $\chi^2(1) = 28.41, p = 9.84 \times 10^{-8}$ for the contrast between the rates of creative adaptation and direct application ideas).

Participants thought the "variance in results is much higher than using other search engines" (P5) and "there're a lot of bordering domains... which can be useful if I want to get ideas in them" (P4).

This difference was also reflected in the content of papers, with keyword papers having significantly more overlapping terms with participant-provided query terms (4.1, SD: 1.74) than analogy papers (1.6, SD: 1.42) (Welch's two-tailed t-test, $t(145.27) = 11.70, p = 1.10 \times 10^{-22}$) (fig. 5, right)[12]. More occurrences of familiar query terms in keyword papers' titles and abstracts may have led participants to perceive them as more familiar.

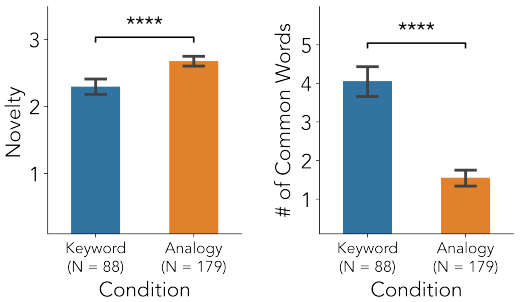

Fig. 5. (Left) Participants judged analogy papers significantly more novel. The mean response to the question *"Have you seen this paper before?"* was significantly higher in Analogy: 2.7 (SD: 0.48) than in Keyword: 2.3 (SD: 0.55). (Right) There were significantly more overlapping words between search query terms provided by participants and the title and abstract text of papers: Keyword: 4.1 (SD: 1.74) vs. Analogy: 1.6 (SD: 1.42).

3.3.2 Analogy papers resulted in more creative adaptation ideas than direct application ideas. We found that the distribution of ideation outcome types differed significantly between analogy and keyword papers ($\chi^2(3) = 52.12, p < 1.0 \times 10^{-10}$). Participants came up with more creative adaptation ideas (N = 53; 32% of total) over direct application ideas (N = 3; 2%) using analogy papers. In contrast, keyword papers resulted in more direct application ideas (N = 16; 19%) than creative adaptation ideas (N = 10; 12%) (fig. 6). The difference between creative adaptation and direct application was significant ($\chi^2(1) = 28.41, p = 9.84 \times 10^{-8}$).

To illustrate more concretely the divergent patterns of ideation leading to Creative Adaptation and Direct Application ideas, we describe vignettes from three participants (table 3). While Direct

[12]We measured the term overlap between participants' queries and the content of papers (title and abstract). To preprocess text, we used NLTK [10] to tokenize papers' content, remove stopwords, digits, and symbols, and lemmatize adjectives, verbs, and adverbs. Finally, using the processed tokens we constructed a set of unique terms for each paper and the query which was then compared to find overlapping terms.

PID	Research Problem	Type	**Paper Title → New Idea** (paraphrased)
1	Improve nanoscale heat transfer in semiconductor material	Direct Application	*Experimental investigation of thermal contact conductance for nominally flat metallic contact* → Apply the techniques in the paper to manipulate thermal conductance at the solid-solid interface
		Creative Adaptation	*Investigation on periodically developed heat transfer in a specially enhanced channel* → Design nanoscale "fins" to absorb heat and convert it to mechanical energy
2	Grow plants better by optimizing entry of nanoparticle fertilizers into the plant	Direct Application	*Nanoinformatics: Predicting Toxicity Using Computational Modeling* → Apply the computational modeling from the paper for predicting toxicity of candidate nanoparticles
		Creative Adaptation	*Identification of Plant Using Leaf Image Analysis* → Invent a hyperspectral 3D imaging mechanism for plants that optically senses, traces, and images plant cells in 3-dimensional structures
3	Enhance the evaporation efficiency of thin liquid films in heat pipes and thermosyphones	Direct Application	*Thin film evaporation effect on heat transport capability in a grooved heat pipe* → Adopt the techniques in the paper for manipulating the solid interface's surface properties to balance the film thickness and disjoining pressure
		Creative Adaptation	*Alkaline treatment kinetics of calcium phosphate by piezoelectric quartz crystal impedance* → Design novel liquid film materials for manipulating hydrophobicity to change disjoining pressure

Table 3. Examples of Direct Application and Creative Adaptation types for three participants (PID). Each participant's research problem is described in the Problem column. While the topics of research problems vary, Creative Adaptation ideas are more distant in terms of content compared to the source problem than Direct Application ideas are, and may be characterized by the use of different sets of verbs ({*design*, *invent*} in Creative Adaptation ideas versus {*apply*, *adopt*} in Direct Application ideas).

Application ideas represented close-knit techniques and mechanisms directly useful for the source problem (described with verbs such as *apply* and *adopt*), Creative Adaptation type ideas were more distant from the source problem and could be characterized with the use of different verbs associated with significant adaptation (*design* and *invent*). For example, P1's research focused on the methods for improving nanoscale heat transfer in semiconductor materials. Previously he developed mechanisms for manipulating the thermal conductivity at solid-solid interfaces, specifically by adjusting the semiconductor wall structures. Thus, a paper reporting experimental

results of manipulating thermal conductance on planar metallic contact points was deemed a directly useful paper that might contain helpful techniques. On the other hand an analogy paper which dealt with the heat transfer phenomenon at a macroscale, using fin-based heat sink designs for electronic devices, gave him a new inspiration: to adapt fins for nanoscale heat transfer in semiconductors to not only transfer heat but also convert it into a useful form of mechanical energy. Despite the mismatch on scale ([macroscale] ↮ [microscale]), challenging the assumption of the typical size of a fin-based design engendered an idea to creatively adapt it to convert heat into energy through an array of tiny fins, rather than merely dissipating it into space as in the original formulation of the problem. P1 also found another analogy paper focused on thermal resistance at a liquid-solid interface useful for future ideation because despite its surface dissimilarities, there was a potential mapping that may open up a new space of ideas (e.g., [liquid] ↮ [polymer substrate], [solid] ↮ [germanium], yet the pairwise relation [liquid:solid] ↔ [polymer substrate:germanium] may be analogous and interesting): "This is liquid... but it's about liquid-solid interface which can be useful... because for the substrate that sits on top of silicon or germanium you use polymers which have liquid-like properties" (P1).

In the case of P2, a paper focused on computational methods for toxicity prediction was deemed directly helpful because "if certain nanomaterials are toxic to certain microorganisms that eat plants or kill them but safe for the plant, we can target these organisms using the nanomaterials as pesticide. Another way this can be helpful is in predicting the chance of toxicity of the nanoparticles in our fertilizers" (P2). Whereas an analogy paper that uses image analysis for plant identification reminded her of "hyperspectral imaging in plants, like a CT scan for plants. So making a hyperspectral 3D model using something like this... to optically sense and trace plant cells (such that the entry of fertilizer nanoparticles into plant cells can be monitored, a sub-problem of P2's research problem) would be pretty cool."

As a third example, P6's research focused on recording and simulating electrical activity using microelectrode arrays. To him, an analogy paper about printing sensors for electrocardiogram (ECG) recording seemed to present an interesting idea despite its mismatch in terms of scale ([nanoscale] ↮ [macroscale]) and manufacturing mechanism ([fabrication] ↮ [printing]), because the pairwise relation between [nanoscale:fabrication] ↔ [macroscale:printing] engendered a reflection on the relative advantages of different methods and future research directions): "Interesting idea! Instead of nanoscale fabrication, printing can be a good alternative for example for rapid prototyping. But I think the resolution won't be enough (for use) in nanoscale... works for this particular paper's goal, but an idea for future research is whether we can leverage the benefit of both worlds – rapid printing and precision of nanoscale fabrication" (P6).

3.3.3 The level of purpose-match had different effects on the ideation outcome. Suggested in these examples is a certain kind of distance the ideas in analogy papers maintain in order to spur creative adaptation. We hypothesize that some amount of difference in purpose facilitates creative adaptation. This process may involve a curvilinear relationship between the degree of purpose mismatch and the resulting ideation outcome, with too much or too little deviation leading to a little-to-no benefit or even an adverse effect on the ideation outcome, a pattern that is consistent with the findings in the literature of creativity and learning outcomes (e.g., Csikszentmihalyi's optimal difficulty [25]). For this analysis, we coded each paper based on three levels of purpose-match to the source problem:

- **Full:** Both high- and low-level purposes match
- **Part:** Only the high-level abstract purpose matches. Explicit descriptions of the high-level purpose exist in either title and abstract of the paper. At the same time, certain low-level aspects of the participant's research problem are mismatched as evidenced by relevant comments from the participant

Purpose-Match	PID	Participant Comment
Full	2	"It's a little bit old (from 2010) but I have read papers from that era. I love this... because the paper mentions everything else and especially one word which is 'disjoining pressure' – if I were to publish my current project that's going to be the core topic."
Part	1	"Though I'm not familiar with GFRP-GFRP... but I can see that they're referring to glass fiber reinforced plastic, so this is something not crystalized material... learning about this kind of materials is interesting."
None	3	"I don't know what a lot of words mean. I don't typically work with animals cells."

Table 4. Examples of different purpose-match types. Purpose-Match shows the level of purpose-match between a recommended paper and each participant's research problem (see table 3 for descriptions of research problems). Fully matching purposes are those that match at both high- (more abstract) and low-levels (specific details). Partial matches only match at the high-level abstraction and differ in details. The Participant Comment column shows relevant excerpts from the participant.

• **None:** Neither high- nor low-level purposes match

Examples of these types of purpose-match are provided in Table 4. High-level match can be considered as a first-order criterion of purpose match and low-level match as a second-order criterion: If the paper does not have overlapping terms in terms of its purpose with the user query cast at a high level (e.g., transfer heat, grow plants) then the low-level match does not matter, but if the paper's purpose matches at the high level, its low-level alignment (e.g., specific aspects of the purpose, such as its scale or materialistic phase) will additionally determine full (i.e., aligned in both high- and low-level aspects of the purpose) vs partial match (i.e., aligned only in the high-level but not low-level aspects of the purpose). Therefore, the coding procedure was symmetrical to the procedure described for coding four types of ideation outcome, with the high-level purpose match deciding between {Full, Part} and None match types, while the low-level purpose further distinguishing between Full vs. Partial match. Following this procedure, two independent coders achieved an inter-rater reliability Cohen's $\kappa = 0.72$ (substantial agreement) and disagreements were resolved with case-by-case discussion.

We used the MEDIATION package[13] [105] to conduct a mediation analysis between the condition, the kind of purpose-match, and the binary Creative Adaptation ideation outcome. The analysis showed that the effect of condition (Keyword vs. Analogy) on the binary outcome of creative adaptation was mediated by the degree of purpose-match, but not by the novelty of content, suggesting that the difference between full vs. partial matching on purpose is much more significant than the variance in the content novelty. We come back to this result in the discussion (§7.2.3). Table 5 presents the result of the mediation analyses. The regression coefficient between creative adaptation and condition was significant as was the regression coefficient between the degree of purpose match and creative adaptation. The indirect effect was $(-.42) \times (.21) = -.09$. We tested the significance of this indirect effect using a bootstrapping procedure [91] ($p < 2 \times 10^{-16}$), by

[13]https://cran.r-project.org/web/packages/mediation/index.html

Mediator	Effect of Condition on Mediator (a)	Unique Effect of Mediator (b)	Indirect Effect ($a \times b$)	CI 95%	
				Lower	Upper
Purpose-match	-0.42^{****} (.08)	0.21^{****} (.05)	-0.09^{****}	−0.14	−0.05
Novelty	0.40^{****} (.07)	−0.06 (.05)	−0.02	−0.07	0.02
Pid	−0.02 (.22)	0.03^{*} (.02)	−0.001	−0.02	0.02

Table 5. Regression table of three mediation analyses using *Purpose-match*, *Novelty* and *Pid* (Participant ID) as mediators between Condition and the binary Creative Adaptation outcome variable. Purpose-match was the only significant mediator between Condition and Creative Adaptation (indirect effect=-.09, significant using a bootstrapping method [91] with 1000 iterations, $p < 2 \times 10^{-16}$).

computing the unstandardized indirect effects for each of 1000 boostrapped samples as well as the 95% confidence interval (CI)[14].

Partial purpose matches in both keyword and analogy papers led to creative adaptation, but the rate was significantly higher with analogy papers. As expected, the ratio of direct application decreased from the keyword papers that fully match in purpose (Keyword Full, 68%) to the keyword papers that partially match in purpose (Keyword Part, 6%) (fig. 8). At the same time, the rate of creative adaptation increased from the keyword papers that fully match in purpose (Keyword Full, 0%) to the keyword papers that partially match in purpose (Keyword Part, 21%). However, the rate of creative adaptation differed significantly between the keyword and analogy papers, with the rate more than doubling among the analogy papers over keyword papers (Analogy Part 47% vs. Keyword Part 21%). Homing in on the partial matches, these papers led to creative adaptation ideas significantly more often in analogy search (47%) than keyword search (21%) (Welch's two-tailed t-test, $t(112.22) = -3.40, p = 9.0 \times 10^{-4}$, fig. 7, left). While the partial purpose mismatch was highly associated with creative adaptation ideas, it could be a double-edged sword. Among the analogy papers, 38% of the partial mismatches resulted in no useful ideation outcome as opposed to the 47% that resulted in creative adaptation (fig. 8, Analogy Part). Therefore, **knowing what mismatches are beneficial to creative adaptation** has important implications for facilitating generative misalignment for ideation.

4 STUDY 2: ENABLING A FULLY AUTOMATED ANALOGICAL SEARCH ENGINE

4.1 Motivation and structure of the study

The findings of Study 1 suggest potential benefits of an analogical search engine for scientific research, but a core limitation of interactivity due to the human-in-the-loop system design prevented its use as a more realistic probe for understanding researchers' natural interaction with analogical results. Specifically, the results of Study 1 are limited by the lack of participants' ability to reformulate search queries and the study design that involved returning only a fixed number of papers that blended both keyword and analogy papers in a randomized order. These factors significantly deviate from realistic usage scenarios of a deployed analogical search engine and prevent us from observing

[14] Alternatively, it is possible that the mediating effect of the degree of purpose-match on the likelihood of creative adaptation outcome is moderated by novelty. However, the result of our analysis showed that this was unlikely: The effect was insignificant using the boostrapping method -.04, ($p = 0.12$, 95% CI = $[-.09, .01]$).

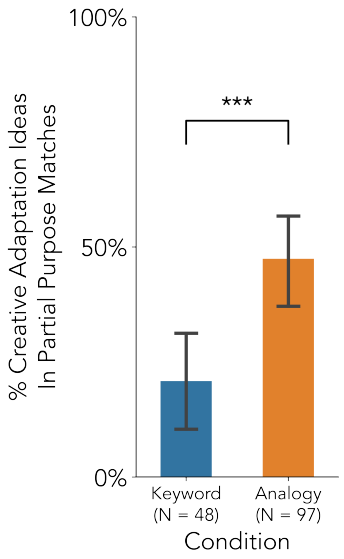

Fig. 7. Proportion of creative adaptation ideas among the partial purpose-match papers. Creative Adaptation was significantly more frequent among the analogy papers (47%) than keyword papers (21%) (Welch's two-tailed t-test, $p = 9.0 \times 10^{-4}$.

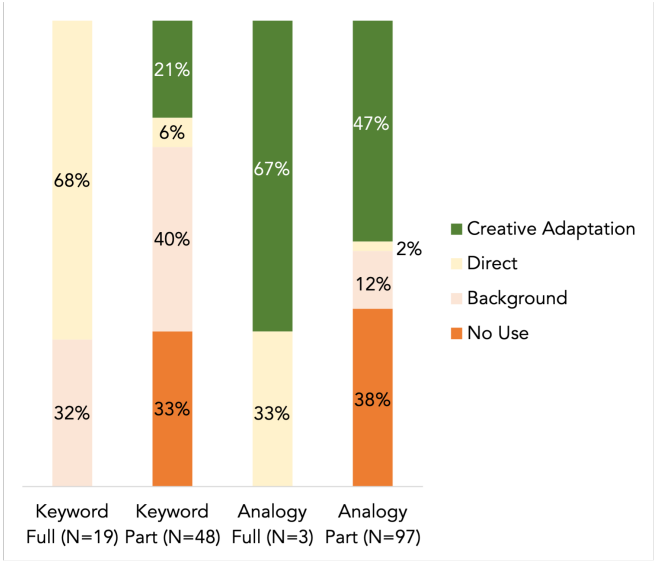

Fig. 8. The rate of ideation outcome types in full and partial purpose matches. Among the keyword papers as the purpose mismatch increases, the rate of creative adaptation also increases from 0% to 21% (middle). However, this rate is significantly higher among the analogy papers (47%) than the keyword papers (21%). Note that while purpose mismatches led to more creative adaptation among analogy papers, a large portion of them also resulted in no ideation outcome (38%).

the full scope of user interaction. In order to move beyond these limitations, first we need a fully automated pipeline that removes the need for human-in-the-loop filtering, thus allowing us to enable query reformulation and interaction with corresponding search results. To achieve this, we improved the model accuracy on extracting purposes and mechanisms from paper abstracts by training a more sophisticated neural network that leverages more nuanced linguistic patterns. Specifically, we implemented an attention mechanism within a span-based sequence-to-sequence model (Model 2) such that it could learn words that frequently co-occur to describe coherent purposes or mechanisms in paper abstracts, and as a result, learning more informative words for our purpose (see Appendix for details of implementation). Through evaluating the system backed by this improved pipeline, we demonstrate how it can remove the human-in-the-loop while maintaining similar levels of accuracy. In the following sections, we report the evaluation results that show 1) an improved token-level prediction accuracy using the span-based Model 2; 2) rankings of the results aligning well with human-judgment of purpose-match from Study 1; and 3) top ranked results of the system maintaining a similar rate of partial purpose matches relative to that of the human-in-the-loop system from Study 1.

The interactivity enabled by the automated analogical search pipeline further allows us to observe its use in more realistic scenarios. To probe how researchers would interact with an analogical search engine and what challenges they might face in the process, we ran case studies with six researchers (§5). From these studies, we uncover potential challenges (§5) and synthesize design implications for future analogical search engines (§6).

4.2 Result

Model	Embedding (finetuned)	All	PP	MN
1. Model 2 [67]	ELMo (N)	**0.65**	0.65	0.64
2. BiLSTM	ELMo (N)	0.63	0.67	0.59
3. BiLSTM	SciBERT (N)	0.62	0.69	0.55
4. BiLSTM-CRF [90]	ELMo (N)	0.58	0.59	0.57
5. BiLSTM	GloVe (Y)	0.55	0.56	0.53
6. Model 1	GloVe (N)	0.50	0.51	0.50

Table 6. F1 scores of different models, sorted by the overall F1 score of Purpose (PP) and Mechanism (MN) detection. The span-based Model 2 gave the best Overall F1 score (blue). In comparison, the average agreement (%) between two experts' and crowdworkers' annotations was 0.68 (PP) and 0.72 (MN) [21]. We used AllenNLP [41] to implement the baseline models 1 – 5.

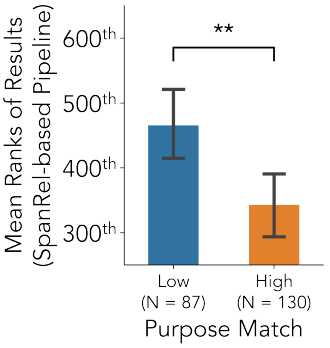

Fig. 9. Mean ranks of human-judged high and low purpose match papers from the span-based pipeline. Low matches were ranked significantly lower (the rank number was higher), on average at 465^{th} (SD: 261.92) than high matches at 343^{th} (SD: 279.48).

4.2.1 Improved token-level prediction of a span-based model. First we compared the span-based Model 2 with five other baselines to evaluate the token-level classification performance (Table 7). Model 2's overall F1 score was the highest at 0.65 (Purpose; PP: 0.65, Mechanism; MN: 0.64, an 0.14- and 0.14-absolute-point increase from Model 1, respectively) on the validation set which represents an overall 0.15-absolute-point increase from Model 1 used for the initial human-in-the-loop analogical search engine.

4.2.2 Pipeline with a span-based model reflected human judgment for ranking the results. The improved token-level prediction performance materialized as an increase in the pipeline's ability to judge the degree of purpose match. For this evaluation, we first recorded every query provided by Study 1 participants that human-in-the-loop filterers used to search and filter the relevant papers. Then, we simulated the search condition of the filterers for the automated pipeline by providing it input as the exact queries they used. We capped the number of top search results sufficiently large at 1000 for each query. From these top 1000 results, we selected papers that also appeared in the human-in-the-loop system and collected the corresponding human-vetted judgments of high or low purpose-match. For each of these papers, we also collected its corresponding rank positions on the new (automated) pipeline's list of results.

We compared the mean ranks of papers that are judged by human filterers as high purpose match to those of low purpose matches. The result showed that the new pipeline indeed was able to distinguish between the two groups of papers; low purpose matches (i.e. papers that were deemed not relevant and subsequently filtered by the judges in Study 1) were placed at significantly lower positions on the list than high purpose matches (i.e. unfiltered papers in Study 1). The mean rank for low purpose matches was 465 while for high purpose matches it was 343 (fig. 9). This difference was significant ($t(192.49) = 3.29, p = 0.0012$. Welch's two-tailed t-test.).

4.2.3 Different model performance on finding papers that fully or partially match on purpose. **Data and coding.** In addition to the overall rankings reflecting human-vetted judgments we also found

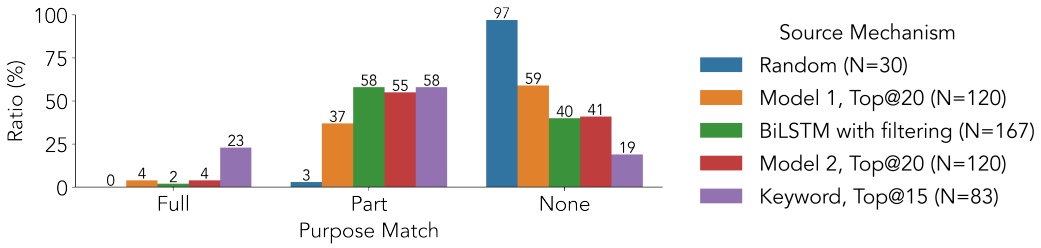

Fig. 10. Distribution of Full, Part, and None purpose matches among the five sourcing mechanisms: *BiLSTM with filtering* represents the human-in-the-loop system (Study 1); *Model 1* represents a system based on the BiLSTM model alone, without human-in-the-loop filtering; *Model 2* represents the fully automated system; *Random* represents randomly sampled papers; *Keyword* represents keyword-based search (Control in Study 1). *Model 2* and *BiLSTM with filtering* showed a similar distribution of purpose matches, and more partial purpose matches than *BiLSTM* alone. Random showed mostly no matches. The *Keyword* condition resulted in the highest number of fully matched papers and the lowest number of no matches, suggesting that keyword-based search may be an effective mechanism for direct search tasks, but potentially less effective for inspirational/exploratory search tasks.

that the proportion of partial purpose matches was significant among the top-ranked results. We sourced top 20 results for each participant's research problem with the automated system (Model 2) using the exact queries and order used by the human-in-the-loop filterers in Study 1. We compared this to four other approaches: 1) the human-in-the-loop system in Study 1 (*BiLSTM with filtering*), 2) a BiLSTM-based system excluding the human-in-the-loop from 1 (*BiLSTM*), 3) randomly sampled papers (Random), and 4) a keyword-based search results, which was used as control in Study 1 (*Keyword*). There were no overlapping papers between Model 2 and other conditions except for the Keyword condition which had 1 overlapping paper. To code the degree of purpose match, we blended the results of Model 2, biLSTM, and Random conditions. Two of the authors coded a fraction of the data together blind-to-condition (7.4%, $N = 20/270$) following the same procedure used in Study 1. Then they independently coded the rest blind-to-condition achieving an inter-rater agreement of $\kappa = 0.80$ (substantial agreement). We resolved any disagreement through discussion on an individual case basis.

Result. We found that the Model 2-based system achieved a parity with the human-in-the-loop system (Study 1) for finding purpose matches (fig. 10), with more than half of the system's top 20 results judged to be partial purpose matches. In contrast, when human-in-the-loop filtering was removed from the BiLSTM-based system, the frequency of partial purpose matches decreased from 58% to 37% while the frequency of no matches increased from 40% to 59%. Random sampling resulted in mostly irrelevant results, with no alignment on purpose with the source problem. An interesting point of comparison is between the keyword-based and Model 2-based search results. Keyword search mostly outperformed Model 2-based system by finding full purpose matches at a much higher rate (23% in keyword search vs. 4% in the Model 2-based system), with similar rates of partial purpose matches (58% vs. 55%), and significantly less no purpose matches (19% vs. 41%). On average the purpose match score was the highest in keyword-search followed by the Model 2-based and the human-in-the-loop systems (fig. 11). Combined with the results of Study 1, this suggests the complementary value of analogical search: The higher rate of full-matches in keyword-search may be good when searchers know what they are looking for, such as in direct search tasks and foraging from familiar sources of ideas. Nonetheless, because analogy papers were both deemed significantly more novel by the scientists and had little-to-no overlap with keyword-search papers,

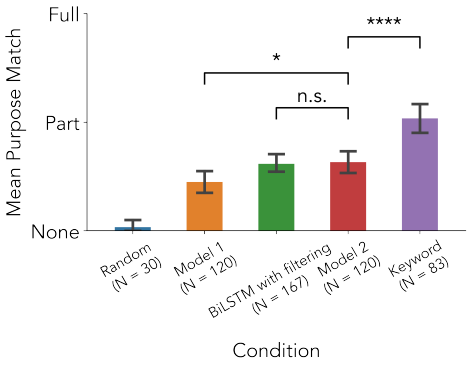

Fig. 11. The distribution of mean purpose match scores over different conditions (mappings: None $\mapsto$ 0, Part $\mapsto$ 1, and Full $\mapsto$ 2). The mean purpose-match score of the system backed by Model 2 (0.63, SD: 0.56) is significantly higher than that of the system used in Study 1 without the human-in-the-loop (BiLSTM, $\mu = 0.45$, SD: 0.58) (Welch's two-tailed t-test, $t(237.87) = 2.49, p = 0.0135$), similar to that of the system with the human-in-the-loop (BiLSTM with filtering, $\mu = 0.62$, SD: 0.52) ($t(244.65) = 0.25, p = 0.80$), and significantly lower than that of the keyword-based search (Keyword, $\mu = 1.04$, SD:0.65) ($t(159.38) = -4.57, p = 0$).

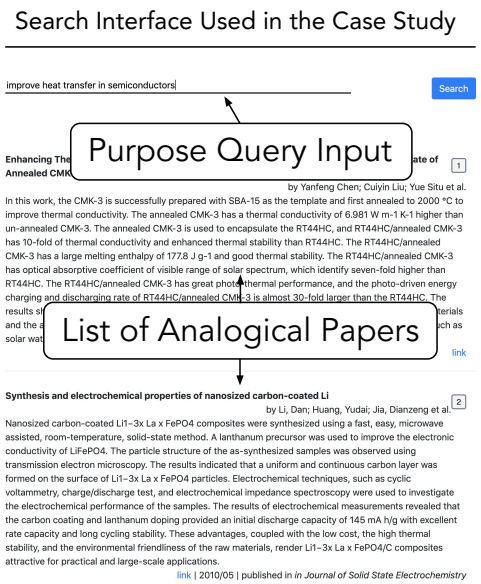

Fig. 12. The search interface used for case studies featured an input for query reformulation which participants used to iteratively reformulate their queries.

they augmented keyword-based search results with a complementary set of papers that introduce useful mistmatches in their purposes. This set of papers may open up new domains of ideas that scientists may not have been aware of, and encourage creative adaptation.

5 CASE STUDIES WITH RESEARCHERS

To further understand what potential interaction challenges prevent future analogical search engines from reaching their full potential, we ran case studies with 6 participants. To this end, we developed a frontend interface that includes a text input for reformulating purpose queries (fig. 12, right). This frontend interfaced with our automated, Model 2-based backend to display a ranked list of analogical results for a given purpose query. Leveraging the fully automated search engine, we also removed the structure of Study 1 that asked participants to engage with each result they encountered, thus allowing us to observe which results researchers more naturally attend to and engage with. In sum, the design of our case studies differ from Study 1 in three aspects: 1) participants interacted with only the analogical search results ranked in the order of purpose similarity, without blended keyword-based search results; 2) participants reviewed search results returned for their queries and reformulated the queries when needed; and 3) participants looked for papers that interest them and may serve as sources of inspiration for their research problems at their own pace, without being explicitly asked to engage with each result they encounter.

The primary goal of our case studies was to identify generalizable challenges that analogical search engines may face in interactive use, thus providing us insights on how future engines may be designed and improved. Specifically, we were interested in the challenges related to 1) how

PID	Participants' Description of Research Problem
1	Improve heat pipe evaporation
2	Computer simulations for fluids in nanoscale and uncovering their heat-transfer properties
3	Developing a model to identify complex steps in Nuclear Power Plant (NPP) operation, and understanding what task features and structures cause the complexity and how this influences the operators' performance
4	Designing simulators for training bridge inspectors
5	Developing algorithms and extensible frameworks for detecting personal protective equipment (PPE) in construction sites to improve the safety of construction workers
6	Convergence rates of optimization algorithms under multiple initial starting positions

Table 7. Case study participants' descriptions of own research problems

researchers recognize relevance of analogical search results and 2) how researchers formulate and reformulate purpose search queries while interacting with analogical search results.

5.1 Participants and Design

Participants were asked to formulate purpose queries for their own research problems and interact with the results to find interesting papers. If a paper gave them a new idea relevant to their research project, they were asked to write a short project proposal in a shared Google Doc and explain how the paper helped them to come up with the idea. Interviews were conducted via Zoom and lasted for roughly an hour. Participants were paid $20 in compensation. One participant was an assistant professor in mechanical engineering at a public R1 U.S. university and five were PhD researchers in the fields of sciences and engineering at a private R1 U.S. university. Two were senior PhD students (3rd year or above) and the rest were 2nd year or below. Disciplinary backgrounds of the participants included Chemical (2), Civil (3), and Mechanical Engineering (1). We note that one participant previously took part in Study 1, whose research focus was the same in terms of its general domain. However, the participant's ideas and the specific papers of interest that led to them did not have overlap between the two studies. Table 7 describes participants' research problems.

Apparatus: Search interface. The improved performance of Model 2 backed the fully automated pipeline without human filtering. The search interface interacting with this back-end included a text input for reformulating purpose search queries as well as a list view of search results that showed a sorted list of papers with similar purposes (fig. 12).

5.2 Result

5.2.1 Overall impressions. Overall participants described their experience with the analogical search engine in a positive light (e.g., "helps me think at a broad topic or a big picture level" – P2; "find some very interesting and useful ideas, the design is also very simple, good when focusing on key areas of research" – P5; and "very interested now what the future of this engine would look like" – P3), but a deeper look suggested that the success of ideation depended on how well searchers were able to engage with analogical results that deviate from their expectations: "It's surprising that the engine recommends examples like these" – P3; "If I input the same search queries on Google Scholar it'd not normally return these things... this search engine works in a different way" – P1.

*5.2.2 "Not the kind of paper I'd look for **but...**": The challenge of early rejections.* Unlike similarity-maximizing search engines, the diversity in analogical search results can lead to premature rejection of alternative mechanism ideas. One of the factors contributing to premature rejection of alternatives

may be the tendency for adherence to a set of existing ideas or concepts, as studied in the literature of design fixation (e.g., [66]). In our study, the participants found the variety of domains featured in search results confusing, and it sometimes prevented them from engaging with the ideas therein. For example, P3, whose research studies ways to manage or reduce task complexity for nuclear power plant operators, expected to see results similar to Google Scholar which are typically in the domains of operational and managerial sciences, but was surprised by unfamiliar domains represented in search results: "These (*distributed networked systems design* or *path planning for automated robots*) are not the kinds of fields that I normally read in, if I found them elsewhere I would've probably thought they're irrelevant and skipped" (P3). Ranging from unfamiliar terms (P1, P4, P5) to unfamiliar categories of approaches (e.g., "Not sure what 'Gauss-Newton approach for solving constrained optimization' is" – P6), or high-level research directions (e.g., "this is different from my research direction, people who work on this direction might find it interesting, though" – P1), participants saw the diversity of results as a challenge for engagement. P1 pointed out a perceived gap between the expectation of least effort and the cognitive processing required when engaging with analogical ideas and adapting them:

> *"As I understand it, I think this search engine is trying to present papers from related but different fields to let people make connections. But people expect less friction. (The result is) something interesting but I can't directly write it into a project proposal... I think it would be challenging to make people get interested in investing time to read the papers in depth to come up with connections. I wonder what would happen if this was hosted just as an online website (instead of the study context)"* – P1

On the other hand, analogs that did get examined more deeply could ultimately lead to creative adaptation. For example, P3 mapped task scheduling among computer processes to task assignment among the nuclear power plant operators, and came up with an idea to adapt algorithmic scheduling used in real-time distributed systems to a scheduling mechanism that could be useful in her research context. Represented symbolically this process was akin to ideating what might best fill in the '?' in the relational structure [scheduling algorithm:processes in distributed systems] ↔ [?:nuclear power plant operators]: "I think the algorithms proposed in this paper could be useful for calculating the operator task execution time, the power plant system's response time, and the time margin between the execution time and the system response time... so that the next task assignment can factor in these margins and things related to workers' well-being like rest and time required between switching tasks" (P3).

Participants seemed to recognize a small number of core relations as kernel for creative adaptation. In the example of P3, *scheduling processes* in the distributed systems paper piqued her interest and led her to connect them with similar concepts in the literature she was already familiar with: "You need to make that connection... I saw parallels between (distributed systems domain) concepts like [scheduling] and [tasks] and [scheduling tasks for the operators]" (P3). Similarly, P5 recognized a similarity between [monitoring people's performance] in fitness training and [monitoring whether construction workers are wearing personal protective equipment] in construction sites. He then adapted the idea of tracking heat emission in the fitness context to his own: "I like the idea of [monitoring heat emissions] in fitness training... maybe I can also detect heat emissions from construction workers to see if they are wearing the safety vests or masks while also monitoring the site conditions and worker efficiency. It also gives me an idea to monitor the CO_2 emissions from workers so as to improve the robustness of detection" (P5). In this case, *monitoring* and the *physical nature* of the activities involved helped P5 see the connection useful for creatively adapting the source idea.

5.2.3 "I don't know what to type in": The challenge of query (re-)formulation. Another challenge participants faced was that they were not used to formulating their search queries in terms of high level purposes of their research. On average participants entered 5.2 queries (Min: 1, Max: 18, SD: 5.87), 87% (27) of which were in the form of a single noun phrase (e.g., "heat pipe evaporation," – P1, "task complexity" – P3, "theoretical optimization convergence for non-convex functions" – P6) or a comma-separated set of multiple noun phrases (e.g., "heat transfer, nanoscale, fluid" – P2) that represented specific aspects related to research purposes rather than the core purposes themselves. For example, the purpose of 'heat pipe evaporation' may be to transfer heat, and the purpose of searching for 'theoretical optimization convergence for...' may be to detect when optimization converges or diverges, or to effectively sample unknown (non-convex) distributions.

One of the reasons why participants formulated search queries in this way may be wrongly assuming that the search engine used keyword matching to find results. For example, extensive prior experience with search engines that highlight matching keywords in abstracts (e.g., Google Scholar) in response to users' search queries can reinforce such assumptions among the users. In addition, participants' domain knowledge useful for judging which of the returned papers are relevant may have led them to notice a set of keywords the inclusion of which strongly signifies the relevance of a paper. In contrast, the analogical search results often seemed to not feature such directly similar terms and this contributed to the difficulty of judging whether a result is relevant and how: "I find these papers not very related to my search query at first. It'd be better if you can use some graph or some pictures to indicate how these papers can relate to my keywords" (P5); "I'd not consider... (because) they are totally different, right? They look irrelevant... until I think about it I can realize that it's useful. But if you give me the paper, at first I don't realize that" (P3).

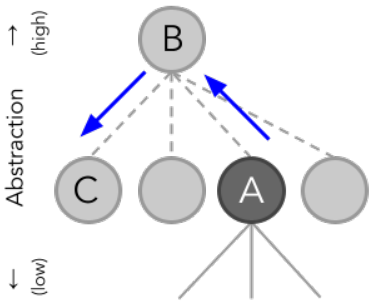

Fig. 13. Diagram showing different abstraction levels of purposes and their relations. Node Ⓐ corresponds to a more specific query than its higher-level representation, denoted as Ⓑ. Similarly, node Ⓒ represents a more specific purpose representation of Ⓐ, accessible via the Ⓐ $\xrightarrow[\text{abstraction}]{}$ Ⓑ $\xrightarrow[\text{specification}]{}$ Ⓒ path.

While it may not feel as compelling or natural to participants, formulating and abstracting queries at a high level may lead to searching more distant results that are analogous at a higher level. For example, by querying "detect personal protective equipment" instead of "personal protective equipment construction," P5 found novel mechanisms of detection, such as general image segmentation algorithms or an approach to monitoring heat in the context of fitness training not specific to construction sites and personal protective equipment but nonetheless useful for creative adaptation. Querying "scheduling tasks" instead of "task complexity" for P3 resulted in finding scheduling algorithms in distributed computer systems that otherwise P3 would not have encountered, while "assigning tasks" led to novel auction mechanisms which made her think about a system in which each power plant operator can bid for a task as opposed to being assigned one. Schematically, fig. 13 shows how formulating queries at a higher level of abstraction than specifying the problem context in full details (Ⓐ → Ⓑ) may lead to discovering novel mechanisms that are relevant at the high level of abstraction, and in more distant ways from the original problem formulation (Ⓑ → Ⓒ).

6 DESIGN IMPLICATIONS

From both the case studies' and Study 1's participants' reflection on the challenges of interacting with analogical search results, common themes emerged. Here we present three design implications

for future analogical search systems synthesized from these results. We use subscripts to denote which study participants participated in when appropriate.

6.1 Support purpose representation at different levels of abstraction

Analogical search engines should support users to formulate their purpose queries at different levels of abstraction. Additionally the search engine may prompt users to consider abstracting or specifying their purpose queries in the first place, and explain how it might help bring new insights into their problems. As seen in the case studies (Section 5.2.3), scientists recognized their purpose queries may be represented at multiple levels, but prior experiences with similarity maximizing search engines may also have anchored them around pre-existing rigid formulation of purposes. Prompting users to consider their research problems at multiple levels may work against this rigidity, and providing candidate suggestions at varying levels may further reduce the cognitive demand. Moving up on the hierarchy to abstract purpose queries may be possible through removing parts of the query words that correspond to specific constraints, or by replacing them with more general descriptions. For example two participants of Study 1 had an identical purpose representation at a high level ("facilitate heat transfer") despite the differences in materialistic phases targeted in each purpose: solid material and semiconductors for $P1_{Study\ 1}$ and liquid thin films for $P3_{Study\ 1}$.

Furthermore, we also observed that looking for only the exact match of a purpose can lead to missed opportunities. For example, although "fins represent a different idea for transferring the heat" and "they (fins) don't match in terms of the scale – macro, not nano," it nevertheless made $P1_{Study\ 1}$ wonder "what if we could design nanoscale wall structures that act like fins that convert heat to mechanical energy?". A corrollary to this observation is that sometimes the superpositions of misalignment with just the right amount can lead to interesting results. For $P4_{Study\ 1}$, a paper presenting experimental techniques for piezoelectric properties was interesting despite its misalignment such as [*simulation*-based] (source) ⇿ [*experimental*] (analog) and [*dielectric properties*] (source) ⇿ [*piezoelectric properties*] (analog): "Though it's an experimental study, it's very close in terms of the material and phenomenon so likely to be helpful. Because we might be able to pick up some trends like, if we increased the temperature, the dielectric response gets stronger or weaker, inferred from the experimental piezoelectric responses, which can then be used to corroborate simulation results or help configure its parameters" ($P4_{Study\ 1}$). However, too much deviation seemed detrimental to its potential for inspiration: "[Molecular dynamic simulation] is the same tool, but (this paper studies) [thermal] (not [dielectric]) properties on [polymer composites]... [polymer composites] are harder to model" ($P4_{Study\ 1}$). In sum, analogical search engines should support not only the capability to 'narrow it down' with specific constraints, but also ways to relax them to broaden the search space when suitable, thus making feasible the sweet spot between too little (i.e. similarity maximization and trivial matches) and too much deviation (i.e. critical misalignment and unusable analogs).

6.2 Support iterative steering from critical misalignment and towards generative misalignment

Analogical search engines should recognize that important constraints may be discovered by users only after seeing misaligned analogs, and support this discovery process by presenting effective examples of misalignment to users. Analogs that deviate on some aspects of the source problem but preserve important relations may be particularly conducive to analogical inspiration that opens up not just individual solutions, but entirely new domains of solutions. However at the same time scientists also found it challenging to know how to come up with effective search queries because combinations of misalignment can sometimes lead to an unintended intersection of domains: "I feel like I'm tricking the machine because [thin film] is often used with [solids], and the term

[pressure] also appears a lot in [manufacturing]... so combining them gives a subset of papers concerned with heat transfer in solid materials and in manufacturing" ($P3_{Study\ 1}$); "on Google Scholar also, I get a lot of polymer strings and get (irrelevant) results like *we use an [electric] device to study [vibration and stress] of [polymers]*... the machine is picking up [electric] and [properties] such as vibration and stress in the context of studying polymers but what I really want is [electric properties] of [polymers] *not* [electronic devices] to study the [mechanical properties] of [polymers]" ($P4_{Study\ 1}$). Nonetheless, seeing misaligned analogs can be an effective way of reasoning about salient constraints and reflecting on hidden assumptions. For example, while evaluating papers about designing microelectrode arrays, $P6_{Study\ 1}$ said: *"Now I think about this (result), I assumed a lot of things when typing that search query... though impedance and topology are my main focus in microelectrode arrays, the coating, size, interface between a cell membrane and electrodes/sensors, biocompatibility, softness of electrodes, fabrication process, material of the platform: silicon or polymer or graphene, form factor: attaching electrodes to a shank-like structure or a broom-like structure, degree of invasiveness, are all part of the possible areas of research and it makes sense that they showed up – there is no way the machine would have known that from my query."* This excerpt illustrates how knowing what the necessary specifications are and which constraints need to be abstracted to cast a wide-enough net to catch interesting ideas appeared to be a difficult task for scientists, especially when they had to recall important attributes rather than simply recognize them from examples of misalignment. Prior work in cognitive sciences also show how dissimilarity associated with various factors in analogical mappings [45] can pressure working memory [112], increase cognitive load [102], and increases response time taken to produce correct mappings for analogy problems [71]. Therefore, analogical search engines should help to reduce the cognitive effort required in the process, for example by proactively retrieving results that are 'usefully' misaligned such that searchers can better recognize (as opposed to having to recall) salient constraints and refine their problem representation. This process is deeply exploratory [93, 115, 118] in nature, and suggest the importance of both providing end-users a sense of progress over time [103] as well as adequate feedback mechanisms for the machine to adjust according to the changing end-user search intent [72, 95, 96]. For example, while the machine may 'correctly' recognize a significant anaogical relevance at a higher level of purpose representation and recommend *macro*-scale mechanisms to a scientist who studies *nano*-scale phenomena ($P1_{Study\ 1}$) or solid and semiconductor-based cooling mechanisms to a scientist in liquid and evaporative cooling systems ($P3_{Study\ 1}$), these analogs may be critically misaligned on the specific constraints of the problem (i.e. the scale or materialistic phase) and thus considered by end-users as useless and even harmful.

6.3 Support reflection and explanation of analogical relevance

Throughout the process of analogical search, human-AI coordination is critical for success, and an important aspect is how deeply the human users can reflect on the retrieved analogs [53] and recognize how different notions of relevance may exist for their own problem context, despite potential dissimilarity on the surface. Looking at previous examples of the tools and techniques developed for targeted reflection support may be useful to this end. For example, ImageCascade [76] provides intelligent support such as automatically generated mood-boards and semantic labels for groups of images to help designers communicate their design intent to others. Another system, Card Mapper, visualizes relative co-occurrences of design concepts using proximity in the design space [26]. Similarly representing the space of analogical ideas using spatial encoding of similarity between two analogs, or designing information that supports getting a sense of the space of search results — e.g., semantic category labels similar to ImageCascade's or the distribution of the domains that analogs are pulled from — may be an avenue for fruitful future research. The explanation of relevance is also important especially when there is a risk of early rejection (§5.2.2). Using examples

from the case studies, one approach to explaining relevance might be to surface a small number of core common features between an analog and a problem query. Such common features were considered useful by scientists for making analogical connections, and they could creatively adapt them for their own research problem context. When common features are not directly retrieved, generation of more elaborate explanations may be required. We refer to [6, 14, 70, 98] for those interested in future design considerations of automatically generated recommendation explanation. Further complementing the direct explanation of relevance approach, techniques such as prompting or reminding the searchers of previously rejected or overlooked ideas may also trigger deeper reflection and delay premature rejection of the ideas based solely on their surface dissimilarity. Participants from both studies commented that the critical first step towards analogical inspiration may be raising similarly enough attention and interest above the initial 'hump' of cognitive demand. Gentle reminders (e.g., "Ask me later if this would be interesting and also provide a list of items" – $P1_{Case\ Studies}$) or resurfacing previously rejected papers in light of new information ($P1_{Case\ Studies}$, $P3_{Case\ Studies}$) may help with users cross this barrier.

7 DISCUSSION

7.1 Summary of contribution

With the exponential growth of research output and the deepening specialization within different fields, encouraging analogical inspiration for scientific innovation that connects distant domains becomes ever more challenging. Our human-in-the-loop and fully automated analogical search engines represent an approach for supporting such analogical inspirations for challenging scientific problems. We have demonstrated in Study 1 that our human-in-the-loop system finds novel results that participants would be unlikely to encounter from keyword-based search, and that these results lead to high levels of creative adaptation. Through a mediation analysis we also showed that this success was driven by the analogical search engine's ability to find *partial* purpose matches (e.g., matching at the high-level purpose but differs at the low-level details). We saw the nuanced effects of partial purpose alignment on the results' goodness as analogs for inspiration. Through qualitative observations, we described how certain attributes of analogical mapping were perceived as more salient by participants, and that mismatches on them can have either a positive (i.e. generative insights) or a negative impact (i.e. critical misalignment) on creative adaptation. In contrast, keyword-based search resulted in more *full* purpose matches and a higher level of direct application. The value of keyword-based search and analogy-based search thus may complement each other, while keyword-based search can help researchers find 'exactly that', analogy-based search can help researchers to switch from a preservative mode (i.e. aiming to find results with maximal similarity to the query) to a generative mode (i.e. aiming to find analogs that are relevant despite the surface dissimilarity) of searching, and ultimately lead them to recognize unusual relations and come up with ways to creatively adapt existing ideas for novel domains.

We also demonstrated how improving the sequence-to-sequence purpose and mechanism identification model can remove the human-in-the-loop but maintain a similar level of accuracy on purpose-match by human judges. This improvement enabled us to develop a fully automated analogical search system to use as a probe to study searchers' more natural interaction with analogical results. Through a series of evaluation we first show that our automated analogical search pipeline can emulate human judgment of purpose match and that it finds partial purpose matches in top ranked results with a similar rate compared to the human-in-the-loop system used in Study 1. Then through case studies we find generalizable challenges that future analogical search engines may face, such as early rejection of alternative mechanism ideas and the difficulty of abstracting and representing purposes at the right level. From our studies we synthesize design implications for

future analogical search engines, such as supporting purpose representations at different levels of abstraction, supporting the iterative process of steering away from critically misaligned analogs and towards a fertile land of generative misalignment, and providing explanations on why certain analogical search results may be relevant. We envision that future studies will shed light on deeper cognitive sources of the challenges identified here. A fruitful avenue of research may be studying how the dual processing theory [69, 113] underlies or interacts with analogical search interaction. Studying also how simplification heuristics [84] may negatively bias interaction with analogical results and how it may be reduced for expert user populations may be an interesting future direction [17, 77].

7.2 Limitations and future work

7.2.1 Experimental design and improving its validity. Our findings have several limitations. First the design of our studies may be improved to increase the experimental validity. We believe that our coders of the ideation outcomes had a reasonable understanding of participants' research context from descriptions of current and past research topics, think-alouds with 45 papers, and end-of-experiment discussions, and that the procedure of coding reduced potential biases (e.g., the coders were blind to experimental conditions, relied on participants' statements of novelty and distance). Despite this, it is possible that they judged ideas differently from domain experts, for example coding more or fewer ideas as creative adaptation, or pre-filtering useful ideas in the human-in-the-loop stage. In addition, other quality dimensions such as potential for impact or domain-expert-judged idea quality are largely inaccessible within the studies presented here. Future research may improve on these limitations by iterating on the experimental design, collecting data for triangulating the results and capturing other quality dimension of the generated ideas.

Additionally, future work may add ablation studies to quantify the effects of human filtering in Study 1 on the ideation outcome as well as sensitivity studies to relate how much the increased token-level classification performance of trained models may reduce the burden of human filtering. Furthermore, additional experiments with baselines other than keyword-based search using the whole abstract will help pinpoint the potential advantages of representing and matching papers using extracted purposes and mechanisms. For example, Chan et al. [21] found that embedding all words from an abstract (using GloVe embeddings) resulted in retrieval of fewer analogical items than when extracted purposes and mechanisms were used. Replicating this result with additional approaches such as contextualized word embeddings and pre-trained language models (e.g., ELMo [90], BERT [29], and SciBERT [7]) will be valuable.

7.2.2 Potential sampling bias. The sampling strategy in Study 1 was purposefully unbalanced, where analogical papers were sampled twice as much as keyword papers to ensure participants' exposure to sufficiently diverse results. This was crucial for uncovering potential benefits and challenges of our analogical search engine and investigating its viability. This ratio was chosen purposefully, to balance the statistical power for detecting potentially significant differences between the conditions, while also limiting the number of papers that each participant had to review. Given the cognitive burden of reviewing a paper while thinking aloud, we decided on 45 in total with the 2:1 ratio to fit the practical time limits of interviews. However, this may have led to unanticipated effects on ideation outcomes despite having accounted for the difference in sample sizes by measuring the outcomes in ratios. For example, when the results were combined into a single blinded list, the over-representation of analogical results over more purpose-aligned keyword results may have shifted the users' overall perceived value of the list to be more or less positive. Users' perception of diverse results may have been further affected by their relative over-representation. For example, increased cognitive load for processing analogical mapping [51, 52, 102] may suggest that results

that fully match on the purpose search query may have been perceived even more favorably than analogical results, due to a negative spill over effect from the rest of the papers in the list, which were less likely matched on the purpose. Investigating whether such factors led to compounding effects beyond our ratio-based measures of usefulness remains an open question for future work.

7.2.3 Controlling the diversity of search results. Our work is also limited by the lack of controllability in sampling the search results beyond purpose similarity. As described in §2.2.1, from pilot tests in our corpus we discovered that even close purpose matches of scientific papers already had high variance in terms of the mechanisms they proposed which allowed us to focus our approach to sampling based solely on purpose similarity. The simplicity of this approach also means fewer hyper parameters in the sampling mechanism compared to other approaches [61, 62]. However, all the approaches including this work thus far lacked a mechanism for explicitly controlling the diversity in retrieved search results which remains a fruitful avenue for future work. For example, prior research has uncovered the nuanced effects of distance (e.g., near vs. far sources of inspiration [24, 97]), suggesting the benefit of targeting analogs at different distance from the source problem for the right context. Future research may also uncover further complexities in the relationship between novelty and purpose-match. The result of our mediation analysis (Table 5) showed that the novelty of content among the search results in Study 1 was not a significant factor to the same extent that the three levels of purpose match was. However, the relationship between novelty and purpose match may be more complex than the levels of manipulation presented in this work. For example, [30] suggested a greater importance of novelty than usefulness for predicting creativity scores. Future work may design mechanisms to manipulate the variance in content novelty and alignment in the purpose-mechanism schema to uncover dynamics between the two that go beyond the results from mediation analyses presented here (§3.3.3). Furthermore, challenges with the abstraction of purposes remain open, for example how core versus peripheral attributes of research purposes may be identified, and how they may be selectively matched at a specific level of the conceptual hierarchy. Finally, not all query formulations are created equal in terms of their suitability for analogical search. We observed in the case studies that participants wanted to express different query intent via reformulation (§5.2.3). While participants could reformulate their search queries and examine the returned results from our analogical search engine in real-time, it was unclear whether and how specific query formulations may lead to more or less diverse results, and how subsequent queries may be updated after reviewing them. As such, systems that assist users in the potentially tedious process of query reformulation [114] (for example, by way of automatic query expansion [18]) in the context of analogical search will be important.

7.2.4 Studying the effect of larger context of scientific innovation on analogical innovation. Due to our focus on ideation outcomes, our results do not explain how these ideas may be integrated, developed, and shared across the research communities. Studying the lifetime of ideas that goes beyond their inception will deepen our understanding of the factors that currently make analogical innovation such a rare event in sciences (for example, Hofstra et al. suggested that more semantically distant conceptual combinations receive far less uptake [58]). Through interviewing our study participants and other colleagues in academia we found emerging structures related to this challenge. Our interviews informed us that in general the context in which a scientist exists – such as the scientist's role in a project, the maturity of a project, and the broader academic culture – can ultimately change how they interact with and seek analogical inspirations. For example a third-year PhD student studying chemical engineering commented "In the current stage of my project it's more about parameter-tuning – running multiple experiments at once and comparing which configuration works the best... If I were a first year PhD student maybe I would be in a broader

field and exploration." In contrast, a PhD in biology who recently defended noted that "analogical inspirations would perhaps be more useful if you're looking for a postdoc or a faculty position."

In addition, the underlying career incentive structures in academia may also affect researchers' perception of and openness to analogical inspirations. One of the study participants commented "since I'm already a third year PhD student and my project is further along and more firmed up, I'm not really looking for really far inspirations... first we push the specific way we have in mind with many iterations on the experiments until, say, publication." In addition to the career-wise incentives there are other types of competitive resourcefulness (e.g., social resources such as the advisors' and colleagues' expertise that participants can easily tap into; physical and other forms resources such as tangible artifacts like previously developed code packages or experimental processes and setups). These factors can influence scientists' perception of their advantage and lead them to interpret analogical inspirations as more or less useful, feasible, and directly applicable to their research. This observation is further suggested by survey results that asked our participants: "*Could this paper be useful to you?*," their ratings were significantly higher for keyword papers than analogy papers despite them having come up with creative adaptation ideas more often with analogy papers. Therefore, future work that studies incentive structures, the quality of ideation outcome, their feasibility, the differences in research context e.g., frames of research contribution such as discovery-oriented vs. novel system development-oriented, and taking a longitudinal observation of the variation in such factors will add a significant depth to our understanding.

8 CONCLUSION

In this paper we present our novel human-in-the-loop and fully automated analogical search engines for scientific articles. Through a series of evaluations we found that analogous papers that our systems retrieved were novel and useful for sparking creative adaptation ideas. However, significant work is needed to continue this trajectory, including additional understanding of the context and incentives of scientists as well as advances in the data pipeline and interaction methods beyond those described here for a system to maximize its real-world impact.

We imagine a future in which scholars and designers could find inspirations based on deep analogical similarity that can drive innovation across fields. We hope this work will encourage scientists, designers, and system builders to collaborate across disciplinary boundaries to accelerate the rate of scientific innovation.

ACKNOWLEDGMENTS

We thank our study participants for their valuable insights and feedback. This work was supported by Center for Knowledge Acceleration, National Science Foundation (FW-HTF-RL, grant no. 1928631; IIS, grant no. 1816242; SHF, grant no.1814826), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 852686, SIAM) and NSF-BSF grant no. 2017741. This work is also based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

REFERENCES

[1] [n.d.]. Random projection in Locality-sensitive hashing. https://en.wikipedia.org/wiki/Locality-sensitive_hashing#Random_projection [Online; accessed 23-Jan-2022].

[2] [n.d.]. Annoy: How it works. https://github.com/spotify/annoy#how-does-it-work [Online; accessed 23-Jan-2022].

[3] Kevin D Ashley. 1991. Reasoning with cases and hypotheticals in HYPO. *International journal of man-machine studies* 34, 6 (1991), 753–796.

[4] Yoram Bachrach, Yehuda Finkelstein, Ran Gilad-Bachrach, Liran Katzir, Noam Koenigstein, Nir Nice, and Ulrich Paquet. 2014. Speeding up the Xbox Recommender System Using a Euclidean Transformation for Inner-Product Spaces. In

Proceedings of the 8th ACM Conference on Recommender Systems (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 257–264. https://doi.org/10.1145/2645710.2645741

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL]

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research* 3 (2003), 1137–1155.

[9] Justin M. Berg. 2014. The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes* 125, 1 (2014), 1–17. https://doi.org/10.1016/j.obhdp.2014.06.001

[10] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Barcelona, Spain, 214–217. https://www.aclweb.org/anthology/P04-3031

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

[12] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 66, 11 (2015), 2215–2222.

[13] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

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

[15] Jaime G Carbonell. 1983. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*. Springer, 137–161.

[16] Jaime Guillermo Carbonell. 1985. *Derivational analogy: A theory of reconstructive problem solving and expertise acquisition.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[17] E Moulton Carol-anne, Glenn Regehr, Maria Mylopoulos, and Helen M MacRae. 2007. Slowing down when you should: a new model of expert judgment. *Academic Medicine* 82, 10 (2007), S109–S116.

[18] Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)* 44, 1 (2012), 1–50.

[19] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[20] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[21] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (Nov. 2018), 21 pages. https://doi.org/10.1145/3274300

[22] 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

[23] 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

[24] 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*. 93–105.

[25] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[26] Dimitrios Darzentas, Raphael Velt, Richard Wetzel, Peter J Craigon, Hanne G Wagner, Lachlan D Urquhart, and Steve Benford. 2019. Card mapper: Enabling data-driven reflections on ideation cards. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[27] Nicola Davis. 2017. *Nasa needs you: space agency to crowdsource origami designs for shield.* https://www.theguardian.com/science/2017/jul/20/nasa-needs-you-space-agency-to-crowdsource-origami-designs-for-shield

[28] Derek J. de Solla Price. 1965. Networks of Scientific Papers. *Science* 149, 3683 (1965), 510–515. https://doi.org/10.1126/science.149.3683.510 arXiv:https://science.sciencemag.org/content/149/3683/510.full.pdf

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[30] Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C Neubauer. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts* 9, 1 (2015), 35.

[31] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing* 4, 4 (2005), 18–26.

[32] 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.

[33] Chris Eliasmith and Paul Thagard. 2001. Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science* 25, 2 (2001), 245–286.

[34] Maryam Fazel-Zarandi and Eric Yu. 2008. Ontology-Based Expertise Finding. In *Practical Aspects of Knowledge Management*, Takahira Yamaguchi (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 232–243.

[35] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. 1993. A description of think aloud method and protocol analysis. *Qualitative health research* 3, 4 (1993), 430–441.

[36] Kenneth Forbus. 2001. *Exploring analogy in the large.* MIT Press.

[37] Kenneth D Forbus, Ronald W Ferguson, and Dedre Gentner. 1994. Incremental structure-mapping. In *Proceedings of the sixteenth annual conference of the Cognitive Science Society*. 313–318.

[38] Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. 2017. Extending SME to handle large-scale cognitive modeling. *Cognitive Science* 41, 5 (2017), 1152–1201.

[39] 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.

[40] Katherine Fu, Joel Chan, Christian Schunn, Jonathan Cagan, and Kenneth Kotovsky. 2013. Expert representation of design repository space: A comparison to and validation of algorithmic output. *Design Studies* 34, 6 (2013), 729 – 762. https://doi.org/10.1016/j.destud.2013.06.002

[41] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Melbourne, Australia, 1–6. https://doi.org/10.18653/v1/W18-2501

[42] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

[43] D. Gentner, S. Brem, R. W. Ferguson, P. Wolff, A. B. Markman, and K. D. Forbus. 1997. Analogy and Creativity in the Works of Johannes Kepler. In *Creative thought: An investigation of conceptual structures and processes*, T. B. Ward, J. Vaid, and S. M. Smith (Eds.). American Psychological Association, Washington D.C., 403–459.

[44] Dedre Gentner and Russell Landers. 1985. ANALOGICAL REMINDING: A GOOD MATCH IS HARD TO FIND.. In *Unknown Host Publication Title*. IEEE, 607–613.

[45] Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. *Encyclopedia of human behavior* 2 (2012), 130–136.

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

[47] Mary L. Gick and Keith 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

[48] 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

[49] Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. 2013. Inspiration peak: exploring the semantic distance between design problem and textual inspirational stimuli. *International Journal of Design Creativity and Innovation* 1, 4 (2013), 215–232.

[50] 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.

[51] Graeme S Halford. 1992. Analogical reasoning and conceptual complexity in cognitive development. *Human Development* 35, 4 (1992), 193–217.
[52] Graeme S Halford, William H Wilson, and Steven Phillips. 1998. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences* 21, 6 (1998), 803–831.
[53] Ning Hao, Yixuan Ku, Meigui Liu, Yi Hu, Mark Bodner, Roland H Grabner, and Andreas Fink. 2016. Reflection enhances creativity: Beneficial effects of idea evaluation on idea generation. *Brain and cognition* 103 (2016), 30–37.
[54] M. Hesse. 1966. *Models and analogies in science.* Notre Dame, IN.
[55] Mary B Hesse. 1966. Models and analogies in science. (1966).
[56] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[57] Douglas R Hofstadter, Melanie Mitchell, et al. 1995. The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory* 2 (1995), 205–267.
[58] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9284–9291.
[59] Keith J Holyoak and Paul Thagard. 1989. Analogical mapping by constraint satisfaction. *Cognitive science* 13, 3 (1989), 295–355.
[60] K. J. Holyoak and P. Thagard. 1996. The analogical scientist. In *Mental Leaps: Analogy in Creative Thought*, K. J. Holyoak and P. Thagard (Eds.). Cambridge, MA, 185–209.
[61] 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* (Halifax, NS, Canada) *(KDD '17)*. ACM, New York, NY, USA, 235–243. https://doi.org/10.1145/3097983.3098038
[62] Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling Creative Inspiration with Fine-Grained Functional Facets of Product Ideas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3517434
[63] Hen-Hsen Huang and Hsin-Hsi Chen. 2017. DISA: A Scientific Writing Advisor with Deep Information Structure Analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 5229–5231. https://doi.org/10.24963/ijcai.2017/773
[64] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).
[65] John E Hummel and Keith J Holyoak. 2003. A symbolic-connectionist theory of relational inference and generalization. *Psychological review* 110, 2 (2003), 220.
[66] David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies* 12, 1 (1991), 3–11.
[67] Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. 2020. Generalizing Natural Language Analysis through Span-relation Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2120–2133. https://doi.org/10.18653/v1/2020.acl-main.192
[68] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3 (2010), 258–263.
[69] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.
[70] Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S Weld, Doug Downey, and Jonathan Bragg. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, 23 pages. https://doi.org/10.1145/3491102.3517470
[71] Mark T Keane, Tim Ledgeway, and Stuart Duff. 1994. Constraints on analogical mapping: A comparison of three models. *Cognitive Science* 18, 3 (1994), 387–438.
[72] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, Vol. 37. ACM New York, NY, USA, 18–28.
[73] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]
[74] Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. 2019. Scaling up analogical innovation with crowds and AI. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1870–1877.
[75] 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 (2020), 535–557.
[76] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E MacKay. 2020. ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proceedings*

of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–27.
[77] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.
[78] Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104, 2 (1997), 211.
[79] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 188–197. https://doi.org/10.18653/v1/D17-1018
[80] Clayton Lewis. 1982. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY.
[81] Salvador E Luria and Max Delbrück. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28, 6 (1943), 491.
[82] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.
[83] David W. McDonald and Mark S. Ackerman. 2000. Expertise Recommender: A Flexible Recommendation System and Architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 231–240. https://doi.org/10.1145/358916.358994
[84] Henry Mintzberg, Duru Raisinghani, and Andre Theoret. 1976. The structure of" unstructured" decision processes. *Administrative science quarterly* (1976), 246–275.
[85] Richar Van Noorden. 2014. Global scientific output doubles every nine years. http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html
[86] R. Oppenheimer. 1956. Analogy in science. *American Psychologist* 11, 3 (1956), 127–135.
[87] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
[88] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP '14)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162
[89] Edwin A Peraza-Hernandez, Darren J Hartl, Richard J Malak Jr, and Dimitris C Lagoudas. 2014. Origami-inspired active structures: a synthesis and review. *Smart Materials and Structures* 23, 9 (2014), 094001.
[90] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202
[91] Kristopher J. Preacher and Andrew F. Hayes. 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers* 36, 4 (01 Nov 2004), 717–731. https://doi.org/10.3758/BF03206553
[92] Sekharipuram S Ravi, Daniel J Rosenkrantz, and Giri Kumar Tayi. 1994. Heuristic and special case algorithms for dispersion problems. *Operations Research* 42, 2 (1994), 299–310.
[93] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 269–276.
[94] Neil Savage. 2016. *Nanofins Make a Better Hologram*. https://spectrum.ieee.org/tech-talk/semiconductors/optoelectronics/nanofins-make-a-better-hologram
[95] Tobias Schnabel, Paul N Bennett, and Thorsten Joachims. 2019. Shaping feedback data in recommender systems with interventions based on information foraging theory. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 546–554.
[96] Tobias Schnabel, Gonzalo Ramos, and Saleema Amershi. 2020. "Who doesn't like dinosaurs?" Finding and Eliciting Richer Preferences for Recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 398–407.
[97] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 83–92.
[98] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[99] L. Streeter and K. Lochbaum. 1988. An expert/expert-locating system based on automatic representation of semantic structure. In *Proceedings. The Fourth Conference on Artificial Intelligence Applications*. IEEE Computer Society, Los

Alamitos, CA, USA, 345,346,347,348,349,350. https://doi.org/10.1109/CAIA.1988.196129
[100] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf
[101] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.
[102] John Sweller, Paul Chandler, Paul Tierney, and Martin Cooper. 1990. Cognitive load as a factor in the structuring of technical material. *Journal of experimental psychology: general* 119, 2 (1990), 176.
[103] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The Perfect Search Engine is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 415–422. https://doi.org/10.1145/985692.985745
[104] ThermoCool. 2021. *SKIVED FIN HEAT SINKS*. https://thermocoolcorp.com/project/skived-fins/
[105] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. Mediation: R package for causal mediation analysis. (2014).
[106] Peter D Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research* 33 (2008), 615–655.
[107] MW Van Someren, YF Barnard, and JAC Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* (1994).
[108] MW Van Someren, YF Barnard, and JAC Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* (1994).
[109] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
[110] 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*. Springer, 115–122.
[111] Manuela M Veloso and Jaime G Carbonell. 1993. Derivational analogy in PRODIGY: Automating case acquisition, storage, and utilization. In *Case-Based Learning*. Springer, 55–84.
[112] James A Waltz, Albert Lau, Sara K Grewal, and Keith J Holyoak. 2000. The role of working memory in analogical mapping. *Memory & Cognition* 28, 7 (2000), 1205–1212.
[113] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.
[114] Ryen W White, Paul N Bennett, and Susan T Dumais. 2010. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1009–1018.
[115] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
[116] Rand R Wilcox and HJ Keselman. 2003. Modern robust data analysis methods: measures of central tendency. *Psychological methods* 8, 3 (2003), 254.
[117] Michael Yovanovich, Richard Culham, and Peter Teertstra. 2004. *Calculating interface resistance*. http://www.thermalengineer.com/library/calculating_interface_resistance.htm
[118] Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. 2008. CiteSense: Supporting Sensemaking of Research Literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 677–680. https://doi.org/10.1145/1357054.1357161
[119] Shannon A Zirbel, Mary E Wilson, Spencer P Magleby, and Larry L Howell. 2013. An origami-inspired self-deployable array. In *ASME 2013 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*. American Society of Mechanical Engineers Digital Collection.

APPENDIX A. REPRODUCIBILITY

Training and validation datasets. The original annotation dataset from [21] also includes Background and Findings annotations which we exclude due to their relatively high confusion rates among the annotators with the Purpose and Mechanism classes and to balance the number of available training examples per annotation class.

Model parameter selection. We experimented with changing the model capacity relative to the signal present in the training dataset by tuning the number of hidden layers and the nodes used in each model architecture. For Model 1 we found a hidden layer of 100 nodes was sufficient. We optimized this model using the cross-entropy loss and the Adam optimizer [73] with a 0.0001 learning rate. For Model 2, we found three hidden layers with 256 nodes led to an improved accuracy on the validation set. We trained this model with an L2 regularizer ($\alpha = 0.01$), dropouts with the rate of 0.3, and the Adam optimizer with a 0.001 learning rate.

Span-based model architecture. We adapt SpanRel [67] as architecture for the span-based Model 2. SpanRel combines the boundary representation (BiLSTM) and the content representation with a self-attention mechanism for finding the core words. More specifically, given a sentence $\boldsymbol{x} = [e_1, e_2, \cdots, e_n]$, of n token embeddings, a span $s_i = [\omega_{s_i}, \omega_{s_i+1}, \cdots, \omega_{f_i}]$ is a concatenation of the *content representation* $\boldsymbol{z_i}^c$ (weighted average across all token embeddings in the span; SelfAttn) and the *boundary representation* $\boldsymbol{z_i}^b$ of the start (s_i) and end positions (f_i) of the span:

$$\begin{aligned} \boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_n} &= \text{BiLSTM}(\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_n}) \\ \boldsymbol{z_i^c} &= \text{SelfAttn}(\boldsymbol{e_{s_i}}, \boldsymbol{e_{s_i+1}}, \cdots, \boldsymbol{e_{f_i}}) \\ \boldsymbol{z_i^b} &= [\boldsymbol{u_{s_i}}; \boldsymbol{u_{f_i}}] \\ \boldsymbol{z_i} &= [\boldsymbol{z_i^c}; \boldsymbol{z_i^b}] \end{aligned}$$

We use the contextualized ELMo 5.5B embeddings[15] for token representation, following the near state-of-the-art performance reported on the named entity recognition task on the Wet Lab Protocol dataset in [67]. We refer to [67, 79] for further details.

Other parameters. We use GloVe vectors for input feature representation for Model 1 with 300 dimensions, consistent with the prior work [11, 78, 88]. For Model 2, we use the contextualized ELMo 5.5B embeddings as described above which have pre-determined 1024 dimensions. We use Universal Sentence Encoder (USE) [20] for encoding purposes. A USE embedding vector has pre-determined 512 dimensions.

[15] https://allennlp.org/elmo