Modelling Analogies and Analogical Reasoning: Connecting Cognitive Science Theory and NLP Research

Molly R. Petersen[1,2] and **Claire E. Stevenson**[3] and **Lonneke van der Plas**[1]
[1] IALS, IDSIA, Università della Svizzera italiana, Lugano, Switzerland
[2] NLP Lab, EPFL, Lausanne, Switzerland
[3] Psychological Methods, University of Amsterdam, The Netherlands
molly.petersen@epfl.ch, c.e.stevenson@uva.nl, lonneke.vanderplas@usi.ch

Abstract

Analogical reasoning is an essential aspect of human cognition. In this paper, we summarize key theories about the processes underlying analogical reasoning from the cognitive science literature and relate it to current research in natural language processing. While these processes can be easily linked to concepts in NLP, they are generally not viewed through a cognitive lens. Furthermore, we show how these notions are relevant for several major challenges in NLP research, not directly related to analogy solving. This may guide researchers to better optimize relational understanding in text, as opposed to relying heavily on entity-level similarity.

1 Introduction

Relational reasoning—and by extension analogy and analogical reasoning—has always held a prominent place in human psychology. Some have argued that analogy is central to the human cognitive experience (Hofstadter, 2001). The fact that analogies were used to demonstrate the emerging capabilities of the representations produced by Word2Vec (Mikolov et al., 2013a,b) when it was first introduced, is perhaps evidence that underscores the value we place on analogy and relational reasoning to understanding our world and the knowledge contained in it.

In cognitive science, perhaps the most well-known theory of analogical reasoning is Structural Mapping Theory (SMT) introduced by Gentner (1983). Here, Gentner differentiates types of similarity between two systems: a source concept and target concept (which can be thought of as a domain and codomain), for example, between the source concept *solar system* and target concept *atom*. Similarity can be measured along two axes: *attribute similarity* between entities (ex: comparing the size difference between a planet and an electron), and *relational similarity*, the extent to which relations between entities or predicates within each system are similar across systems (e.g., a planet orbits the sun in our solar system, and an electron orbits a nucleus in an atom). *Literal similarity* holds when both attribute and relational similarity between two systems is high, *mere appearance matches* when attribute similarity is high while relational similarity is low, and *analogy*, when attribute similarity is low but relational similarity is high. She also introduces the *systematicity principle*, where preference is given to preserving higher-order relations between two systems.

Because the extent of attribute and relational similarity between two systems is a continuum, the extent to which a comparison between any two systems fits within these categories is also a continuum, and depends on the context of the comparison taking place (Gentner, 1987). An electron and planet may both share the attribute of roundness, however they are dissimilar in many other attribute dimensions that neither mere appearance or literal similarity can accurately describe. Where they are similar is through their relations to other entities in their respective systems.

Regardless of whether one thinks that analogy is merely a common, useful tool for human cognition, or that it is at the core of it, the role it plays in cognition is profound and therefore the goal of making models smarter and more human-like should incorporate this type of reasoning.

In this paper, we cover key theories from the cognitive science literature regarding analogical reasoning, and review recent research regarding analogy solving in natural language processing (NLP). The goal of this paper is not to provide a detailed compilation of all analogical reasoning research in both the fields of cognitive science and computer science, but instead create an

Transactions of the Association for Computational Linguistics, vol. 14, pp. 711–732, 2026. https://doi.org/10.1162/TACL.a.632
Action Editor: Alessandro Lenci. Submission batch: 2/2025; Revision batch: 7/2025; Published 5/2026.

introduction for NLP researchers. We focus specifically on the processes of analogical reasoning, each of which have a large body of research, as well as their own unresolved questions. We chose to focus on this aspect of analogical reasoning since most of these processes are more or less already defined concepts in NLP. However, in practice they are not often viewed through a cognitive lens, but purely by their definitions in the computer science field. We also limit our discussion to computational methods of NLP using neural word embeddings, as reviews of other methods have already been published elsewhere (Holyoak and Thagard, 1989b; Gentner and Forbus, 2011; Mitchell, 2021). Since our focus is on NLP, we focus only on text-based analogies.

Lastly we argue that analogical reasoning is highly relevant to NLP at large, and can potentially be a way in which common limitations with current NLP models can be addressed. We do not provide specific answers to these problems, nor do we think analogical reasoning is the ultimate answer to addressing these problems. We instead wish to create a starting point for the NLP research community to engage in research that is theoretically motivated by the cognitive science literature.

Our paper is organized as follows. **Section 2:** we summarize theory from cognitive science regarding the analogical reasoning processes in humans, **Section 3:** we discuss different types of analogies as well as their specific considerations, **Section 4:** we summarize recent work in analogy solving and more broadly analogical reasoning in NLP, connecting them to the analogical reasoning processes discussed in Section 2, and finally, **Section 5:** we discuss domain-specific applications for analogies, as well as re-frame common NLP tasks and challenges through an analogical lens.

2 Analogical Reasoning in Cognitive Science

In the cognitive science literature, analogical reasoning is split into different processes. These are 1) RETRIEVAL, 2) MAPPING, 3) REREPRESENTATION, 4) ABSTRACTION, and 5) ENCODING. The literature will sometimes vary on which processes are included (Gentner, 1987; Holyoak and Thagard, 1989a; Gentner et al., 1993; Thagard et al., 1990; Gentner and Loewenstein, 2002; Kokinov and French, 2003; Gentner and Forbus, 2011; Gentner and Smith, 2012), however RETRIEVAL and MAPPING are always discussed. MAPPING specifically is considered the central process of analogy-making (Gentner and Forbus, 2011).

Additionally, there is often a distinction made between near and far analogies—near analogies being when both systems of an analogy come from a similar domain (and therefore the elements inside it are likely semantically similar), and far analogy being where they come from two disparate domains (Barnett and Ceci, 2002). It has been demonstrated in humans that solving near analogies is generally easier than far analogies (Bunge et al., 2005; Jones et al., 2022). However, far analogies have been demonstrated to better encourage novel problem solving and promote relational reasoning (Cagan et al., 2011; Vendetti et al., 2014; Chan and Schunn, 2015; Walker et al., 2018).

In this section, we will briefly discuss all five processes, and use the example of *planet:sun::electron:nucleus*, where the single *:* stands for *orbits*, to more concretely explain the processes. It's important to note that these processes are not necessarily independent of each other or performed step-by-step. They can be done simultaneously and are interdependent (Holyoak and Koh, 1987), therefore the order we present them in does not hold any special significance other than the ease of explanation. The processes and examples are summarized in Table 1.

2.1 Retrieval

In an ideal scenario, one may be given a relevant example and told to map it to a pre-specified target domain, or pre-defined list of options to select an analogy from, therefore eliminating the need for RETRIEVAL to begin with. More typically, when presented with a scenario, in order to form a useful analogy one must retrieve an instance from memory that is an appropriate analogue (Holyoak and Koh, 1987). In humans, this memory generally contains past experiences (episodic memory) and gained knowledge (semantic memory). For models, this would be the training data the model has seen, and occasionally also a knowledge base the model can query.

For an explicit example, when presented with the terms *planet:sun*, a cognitive system must explore their knowledge base to retrieve either *electron:nucleus* or another relevant pair. Notably,

Process	Example
RETRIEVAL	When given the pair *planet:sun*, collecting potential matches from memory that may contain the *orbits* relation, such as *electron:nucleus*, *satellite:Earth*, and *the World:you*
MAPPING	Identifying that *planet* plays the same role as *electron* in the solar system and an atom, respectively
REREPRESENTATION	Altering the representation of each system to improve the analogy. One does not need to represent each planet in the solar system or each electron in an atom individually, these entities can be grouped
ABSTRACTION	Representing the solar system as a *central force system*, where the particular entities in the general concept of a central force system are not specified
ENCODING	How an instance is represented in memory. Understanding the solar system as a *central force system* vs. a *gravitationally-bound system*

Table 1: Summarization of the processes of analogical reasoning using the example *planet:sun:: electron:nucleus* and the relation *orbits*.

there are other possible pairs that could be retrieved which will vary on their validity and utility for any given use case. In theory other potential candidates could be *satellite:Earth* or *the World:you*. Whether potential matches are appropriate for the specific use case is handled during evaluation.

RETRIEVAL poses a large challenge within the analogical reasoning process. Not only does the pool of candidates for selection include all knowledge or experience present in the given cognitive system, but there is also the issue of how to retrieve a relevant instance when potential relevant instances may have little or no semantic or surface-level similarity to the target domain. It has been shown in human studies that surface similarity affects people's abilities to retrieve relevant analogs despite people's preference for relational similarity in analogies (Gentner et al., 2003, 1993; Holyoak and Koh, 1987; Minervino and Trench, 2024). This has been attributed to how information is processed in both the ABSTRACTION and ENCODING steps, which we describe later (Gentner et al., 2003).

2.2 Mapping

MAPPING is considered the crux of analogical reasoning—without MAPPING between two systems there is no analogy. In our solar system example, this would be MAPPING the entity *planet* to *electron* and *sun* to *nucleus* by noticing that they play the same roles in relation to *orbits*. One can use this initial mapping to perform additional mappings between the two systems, such as connecting the role of *gravity* to that of *electromagnetism*. As stated previously, it is difficult to discuss analogy without mentioning Gentner's SMT (Gentner, 1983) outlined in the Introduction, however many people have expanded on or proposed alternatives to her original theory.

Holyoak and Thagard (1989a) introduced three MAPPING constraints gathered from common themes that appear in the analogy literature. The authors specify that these are not requirements, but instead act as ''pressures'' to be optimized during the analogical reasoning process. The first constraint is *isomorphism*, or the idea that mappings should be one-to-one and complete (bijective). This forms a structural constraint between two systems. Ideally, matches are one-to-one and complete, but this is only a constraint and not an absolute criterion. So, not all components must be matched, and one-to-many and many-to-one matches are also possible. This can partially be addressed in the REREPRESENTATION process covered in Section 2.3. The second constraint is that of *semantic similarity*, where the similarity in meaning between entities in the two systems should be considered to some extent.

Lastly, they argue that purpose should guide the MAPPING process with the constraint of *pragmatic centrality*. Analogical reasoning and

MAPPING always take place in the context of an overarching goal or purpose, and all steps of analogical reasoning, MAPPING included, will depend on the intended purpose and must take that into account. In theory, a large number of components in two systems can be mapped, or there could be different combinations of mappings between two systems, however, not all are universally relevant to all use cases.

2.3 Rerepresentation

REREPRESENTATION (not to be confused with *representation* as typically used in NLP, which is probably closest to the ENCODING process described in Section 2.4) is the process of altering a representation of either or both the source or target domain to improve a match between the two systems (Gentner and Forbus, 2011). Knowledge regarding a particular system can change over time due to experience or education, requiring updates to how it is represented in memory (Yan et al., 2003).

Yan et al. (2003) defined several methods to assist REREPRESENTATION: *truth-preserving transformation* to the structure (e.g., a planet being smaller than the sun can also be expressed as the sun being larger than a planet); *decomposition* of relations which still preserves the relevant features of the original relations (e.g., electrons do not have elliptical orbits like planets, but this feature may not be relevant to the analogy); *entity collecting*, grouping entities when multiple entities play equivalent roles (our solar system has eight planets, which can be grouped when comparing to electrons); and *entity splitting*, which helps one-to-one MAPPING when a particular entity plays multiple roles by splitting these objects into individual parts.

2.4 Abstraction and Encoding

In the original SMT paper, ABSTRACTION was defined as a type of similarity separate from analogy, however in more recent papers ABSTRACTION has been considered part of the process of analogical reasoning (Gentner, 1983; Gentner and Hoyos, 2017; Gentner and Forbus, 2011). Gentner and Hoyos (2017) define ABSTRACTION as ''decreasing the specificity (and thereby increasing the scope) of a concept'' (p. 673). In analogy, the focus is generally on abstractions of relations, creating rules or frameworks of a general concept from more specific examples. Given that surface similarity generally corresponds to specific details between two systems, ABSTRACTION can then help generalize and decrease the influence of irrelevant and distracting details (Gick and Holyoak, 1980).

ENCODING is how a particular structure is represented in memory. This can be a single specific example or the schema resulting from ABSTRACTION (Mandler and Orlich, 1993). How an abstraction or instance is encoded has been suggested as the key to effective RETRIEVAL from memory as well as analogical transfer (Gick and Holyoak, 1983; Loewenstein, 2010; Mandler and Orlich, 1993).

For example, both a solar system and an atom can be abstracted as a central force system (Gentner, 1983). In each case, the object of central force differs, as does the orbiting object and the force creating the orbit itself. The specifics of the objects that play each role has been abstracted away, they are slots to be filled with specific entities. How this is encoded in memory will have an effect on how it can be retrieved and compared later—if the solar system was encoded as a gravitationally–bound system in memory, it may be harder to retrieve an atom as an appropriate analogy (Gick and Holyoak, 1983). Additionally, it can be encoded in memory as both, however which encoding is associated and retrieved can vary given the specific context (Gentner et al., 2004), and subsequently effects analogical transfer.

One potential challenge with ENCODING and ABSTRACTION is word choice and how that affects a cognitive system's perception of the information that is contained in text, or more generally, how information is presented to optimize extraction of relevant details (Yan et al., 2003; Ramscar and Yarlett, 2003; Gentner et al., 2003; Gick and Holyoak, 1980). Additionally, there is the limitation of how much knowledge a cognitive system contains in order to formulate a representation that includes relevant details that can be mapped (Yan et al., 2003; Hummel and Holyoak, 1997). These challenges can be addressed in the REREPRESENTATION stage (Forbus et al., 1998; Gentner et al., 2003).

3 Problem Types

We will now describe our classification of analogy types and discuss their specific considerations. There has been other work defining a taxonomy of different analogy types (Wijesiriwardene et al.,

2023a,b; Nagarajah et al., 2022). For example, Wijesiriwardene et al. (2023a) present a taxonomy of analogies based on the depth of knowledge and information a system needs. For the purpose of this paper, we define analogies as belonging to one of three types: symbolic, entity-level, or contextualized. This categorization is not meant to compete with or replace other taxonomies, but instead to define and discuss three broad categories based on data format, which determine the approaches NLP researchers may choose to handle them.

3.1 Symbolic Analogies

First, we discuss analogies that, while made up of text, generally contain no semantic information. They are strings of letters and symbols that contain patterns which are solvable by humans (Hofstadter et al., 1995). For example, given *abc* $\rightarrow$ *abd*, finding the transformation for *lmn*. Additionally, this type of analogy could also include MAPPING words in natural language to non-semantic strings of characters (Musker et al., 2024) as well as digit matrices (for example, a 3$\times$3 matrix with a blank at the position at [3,3] which needs to be filled by the solver) (Webb et al., 2023).

One challenging aspect of this type of analogy is the potential to apply a literal, but valid, rule to solve the analogy. For example, for *abc* $\rightarrow$ *abd*, the conceptual rule is to change the last letter in the sequence to the next letter in the alphabet, resulting in *lmo*. However, a potential response could be *lmd*, where the assumed rule would be to replace the last letter with *d*, regardless of the original letter (Lewis and Mitchell, 2024).

Research on symbolic analogies in NLP is generally less prevalent than the other categories of analogy, perhaps because as mentioned before they often do not involve semantic information, and by extension the ability to solve these sorts of analogies are probably not indicative of any sort of natural language understanding. They do, however, function as a tool to measure reasoning and pattern recognition.

3.2 Entity-Level Analogies

Also called proportional analogies, these are analogies that are expressed in the format *a:b::c:d*, verbally expressed as *a* is to *b* as *c* is to *d*. These can be extended to analogies that contain more than 4 entities (e.g., *planet:electron::sun:nucleus::gravity:electromagnetism*). However, they do not contain additional language that contextualizes the entities in a larger picture, or extra textual information that models would need to either consider or discard to identify the analogy. Solving these analogies involves either selecting the best option from a predefined list of potential answers, or generating an answer from existing knowledge.

3.2.1 Morpho-syntactic Analogies

In NLP, the most well known English language benchmarks for analogy are probably the Google Analogy Test Set (Mikolov et al., 2013a) and the Bigger Analogy Test Set (BATS) (Gladkova et al., 2016). These datasets are largely composed of analogies where the ‘‘is to’’ relation is a morphological relation such as pluralizing a singular noun (e.g., *member:members::fact:facts*). Versions of these analogy datasets have been released for a variety of languages (Ulčar et al., 2020; Karpinska et al., 2018; Krishnan and Ragavan, 2021; Grave et al., 2018).

While morphological analogy benchmarks fit under the definition of analogy, they are limited by the fact they aren't very representative of human analogical reasoning (Ushio et al., 2021b; Petersen and van der Plas, 2023). They are trivially easy to solve;if given the pair *try:trying* and then asked to complete the analogy with the made up verb *‘‘zoop’’*, the fourth term is clearly *‘‘zooping’’*.

Here *‘‘zoop’’* has no definition, nor is knowing the meaning of *‘‘zoop’’* required to solve the analogy. That is not to say these analogies have no use, there are potential creative uses of being able to correctly conjugate novel words or morph already existing words into different parts of speech. But the relational reasoning that can be tested with these benchmarks is ultimately limited (Ushio et al., 2021b).

3.2.2 Semantic Analogies

Datasets in this category include the Scientific and Creative Analogy dataset (SCAN) (Czinczoll et al., 2022), analogies created to test human analogical reasoning (termed psychometric analogies by Ushio et al. [2021b]) such as the SAT dataset (Turney et al., 2003) and the Knowledge-intensive Analogical Reasoning benchmark (E-KAR) (Chen et al., 2022), as well as some relations present in the Google Analogy Testset and BATS (e.g., encyclopedic semantics). With these analogies, one

must understand all terms present in the analogy, as well as how they relate to each other.

3.2.3 Vocabulary Beyond the Layperson's Dictionary

While entity-level analogies may be thought of as containing words that could be found in a dictionary compiled for use by the everyday person, this concept can be extended and applied in more creative ways to include domain-specific jargon and concepts. For example, Yamagiwa et al. (2024) used this level of analogy for drug-gene relations where the entities consisted of drugs and genes (e.g., *bosutinib:ABL1*). Blair-Stanek and Van Durme (2021) used entities to represent U.S. tax codes and concepts, in which different tax codes form analogies by pertaining to similar rules, regulations, topics, etc.

3.2.4 Entity-Level Analogies: Considerations

As mentioned previously, solving these analogies involves either selection or generation, where generation includes the RETRIEVAL process while selection does not. If the goal is to generate an answer, arguably many words can be correct and vary in degree of correctness, creating a challenge for evaluation (Rogers et al., 2017).

The lack of context poses a challenge with these analogies. Words can be polysemous, and the precise definition employed by an entity may not be obvious given the other entities. If there is no other entity in the concept, as will often be the case in analogies of the format *a:b::c:?*, the meaning has to be inferred from the source concept. The effect of lack of context will depend on the part of speech, word frequency, and other qualities of a particular entity (Gentner and France, 1988; Asmuth and Gentner, 2017; Fenk-Oczlon et al., 2010). These issues could potentially be addressed with the *decomposition* method for REREPRESENTATION.

It's known that analogies are permutable (for example, permuting *a:b::c:d* to *a:c::b:d*) (Marquer et al., 2022; Antić, 2022), however, these permutations do not hold for every analogy. Given the analogy *Lima:Peru::Cuzco:Peru*, where the relation is *''city in''*, permuting the analogy to *Lima:Cuzco::Peru:Peru*, clearly creates an invalid mapping, as *Peru* = *Peru*, while *Lima* $\neq$ *Cuzco*. Many may be non-injective functions, such as the example with Peruvian cities, for which these permutations can potentially create invalid mappings.

3.3 Contextual Analogies

These are analogies that involve lengths of text ranging from phrases to passages. In line with the entity-level analogy between the solar system and an atom previously mentioned, comparing paragraphs describing these two systems, or even comparing whole Wikipedia pages between these two systems to find the analogical components, would fit into this category.

This category is obviously much more broad then entity-level analogies, and encompasses most use cases of analogies that you find ''in the wild'', including narratives (Gick and Holyoak, 1980; Sourati et al., 2024; Nagarajah et al., 2022), procedural texts (Sultan and Shahaf, 2022), and legal texts (T.y.s.s. et al., 2024).

3.3.1 Contextual Analogies: Considerations

While the lack of context may present challenges with entity-level analogies, the presence of context can also pose a challenge. The text may include information that is irrelevant to an analogy, or that distracts from the analogy with surface-level similarities or differences. It may also leave room to rely on spurious correlations, which may hinder ABSTRACTION and generalization (Gentner and Rattermann, 1991; Wang and Culotta, 2020).

In line with the theory that analogy should take into account *pragmatic centrality*, getting the model to consider the goal when making decisions could be challenging. When given text, it is not necessarily obvious what entities should be included in the MAPPING process, which may change based on the objective. This is opposed to entity-level analogies where maximizing the mappings between all provided entities will generally be the goal. Additionally, the objective will define what instance should be retrieved from memory.

4 Current Research on Analogies in NLP

Given the role of analogies as benchmarks in NLP as well as their use as a measure of reasoning abilities in humans, the question naturally arises whether any aspect of analogical reasoning or types of analogies are considered ''solved'' in NLP. To some extent this is a difficult question to answer, as high performance on currently available benchmarks raises the question of whether or

not these specific benchmarks have been seen in a model's pre-training data (Hodel and West, 2024; Deng et al., 2024).

Generally speaking, results from recent models suggest that analogies at any level are not solved, with the exception of morpho-syntactic analogies, where some models and experiments have demonstrated almost perfect accuracy (Chan et al., 2022; Yuan et al., 2024a).

For example, after the introduction if GPT-3 (Brown et al., 2020), Webb et al. (2023) evaluated the model on a variety of analogy tasks (including tasks from all categories of analogy mentioned in the previous section), and found that GPT-3 performed the same or better than humans on a variety of analogy tasks, making the claim that GPT-3 has the capacity to perform general reasoning tasks in the zero-shot setting. However, a response from Hodel and West (2024) criticized the ability of their experimental setup to evaluate general, zero-shot reasoning in GPT-3, mentioning the potential that the analogies used as tests where prevalent in the training data. The analogies they used have generally been around for quite a long time, such as Dunker's radiation problem (Duncker and Lees, 1945), which was originally published over half a century ago and has been cited over 6000 times, and Hofstadter's copycat problems (Hofstadter et al., 1995), which were published in the 1990s. Hodel and West (2024) found that models failed when the original tasks were modified, and that the model could accurately describe and provide an example to the copycat problem when prompted.

We will now cover specific approaches various researchers in NLP have taken with analogy tasks and language models (LMs). While computational modeling of analogies has been done for decades (Falkenhainer et al., 1989; Holyoak and Thagard, 1989a; Thagard et al., 1990; Hofstadter et al., 1995; Hummel and Holyoak, 1997; Doumas et al., 2008), we focus on more recent advances of modeling analogies using neural word embeddings. We attempt to connect specific experiments to the processes outlined in Section 2. In general, we find most research with entity-level analogies focuses on the MAPPING process. For contextual analogies, while motivating their work with theory from analogical MAPPING, notably, these works often do not explicitly perform any MAPPING between entities or relations, but instead perhaps more closely address the ABSTRACTION and ENCODING processes.

4.1 Symbolic Analogies

Lewis and Mitchell (2024) tested several GPT models on symbolic analogies with the English alphabet, as well as with novel alphabets (permuted English alphabets and a symbolic alphabet), and digit matrices. They found that while humans could generally cope with alternative alphabets, GPT models struggled with them more compared to the standard alphabet. This suggests an issue with RETRIEVAL, ABSTRACTION, and REREPRESENTATION, as they are not robust to reasoning over novel alphabets despite seeing alphabets and likely letter string analogies with the standard English alphabet in their training data. Additionally, on the digit matrices, human performance did not change much if the blank of the matrix was moved from the [3,3] position to an alternate position, however models struggled with this change.

Musker et al. (2024) designed a symbolic analogy task that incorporated semantic information by creating analogies that are completed by taking sematic knowledge from words provided as *a*, *c*, *e* terms, etc., and converting them to non-semantic strings, which they refer to as a ''semantic content'' task. For example, provided pairs may include a list of animals mapped to symbols, such as *horse*:*, *cat*:*, *spider*:! and so on where all mammals are mapped to an asterisk, *, and all non-mammals are mapped to an exclamation point, !. They found that smaller models performed much worse than humans, and therefore focused primarily on GPT-4 (OpenAI et al., 2024), Claude 3 (Anthropic, 2024), and Llama-405B (Grattafiori et al., 2024), as well as tested performance in humans. They found that Claude 3 achieved human performance on the semantic content task, while the other two models fell below human performance. However, given that this is just one experimental setup, more investigation should be done in order to evaluate whether models can solve these sorts of analogies over a robust set of problems.

4.2 Solving Entity-Level Analogies with Neural Networks

The Vector Offset Method. Analogies are a popular benchmark to demonstrate that neural word embeddings encode knowledge beyond word similarity alone. The original method for solving analogies with Word2Vec—referred to as

3CosAdd by Rogers et al. (2017)—involves finding the word, *v*, in a vocabulary, *V*, that is most similar to the vector resulting from the equation *a+c-b=d'*, where *a, b, c* and *d*, are the vector representations of the terms in the analogy *a:b::c:d* and *d'* is the estimated vector representation for *d* given the representations of *a, b* and *c*. This is to be compared to the actual, static vector representations $v \in V$ for *d* estimated through training (Mikolov et al., 2013a; Drozd et al., 2016; Mikolov et al., 2013c). Similarity in this case is defined as cosine similarity, *cos(d, d')*. While the list of possible *d* terms are finite in that they are limited to the model's predefined vocabulary size, this method still involves the RETRIEVAL process as all items (or most) in the vocabulary known to the model are potential candidates.

Researchers have identified several problems with this method of solving analogies. Rogers et al. (2017) and Linzen (2016) found that performance using 3CosAdd is heavily related to the proximity of *a, b*, and *c* terms in semantic space, and that the ability for 3CosAdd to correctly identify analogies decreases as similarity between the entities in the relation decreases. This problem was found to hold with other vector offset methods presented in Drozd et al. (2016) and Levy and Goldberg (2014). Linzen (2016) also found that if you do not exclude *a, b*, and *c* from *V* as an estimate of *d*, as was done in the Mikolov et al. (2013c) paper, that often the vector offset method will predict *b* or *c*.

Additionally, this method generally only allows options for *d'* terms to be selected from a model's predefined vocabulary that was determined before training. Chan et al. (2022) address this problem by proposing a model that generates the *d'* terms allowing for out-of-vocabulary generation, as opposed to retrieving them from a finite list. Their approach involved training a BiLSTM model to take as input each character of a word to encode a representation of the word. The 3CosAdd method is then applied to get a single representation of the analogy, and sent through a LSTM decoder to generate the *d'* terms. They achieved almost perfect performance on morphological analogies from eight languages.

Analogies for Model Probing. Several papers have used semantic entity-level analogies as a probing task for LMs (Ushio et al., 2021b; Czinczoll et al., 2022; Chen et al., 2022; Yuan et al., 2023), often while releasing an entity-level specific dataset. Ushio et al. (2021b) and Czinczoll et al. (2022) both found that models generally perform worse on datasets composed of only semantic-level analogies than on those that contain syntactic relations. When fine-tuning models with the BATS dataset both Czinczoll et al. (2022) and Yuan et al. (2024a) found that performance is reduced on semantic analogy datasets, suggesting that these sorts of analogies are inherently different, and that semantic analogy datasets require the understanding of more abstract relations to solve.

Algorithms for MAPPING Analogies. Analogical MAPPING can additionally be performed between entities using algorithms that utilize information from LMs, without specifically performing MAPPINGS between the neural representations of entities themselves. Jacob et al. (2023) developed an algorithm to map between entities that utilizes LMs in several ways. First, they utilize GPT-3 as a knowledge base to extract relations between entities. They then use SBERT (Reimers and Gurevych, 2019) to calculate the similarity of relations between two systems, and to cluster relations within a system that are similar (the *decomposition* method for REREPRESENTATION presented in Section 2.3). Since the algorithm relies on similarities of relations between entities as opposed to similarity between entities specifically, they found that their algorithm may be more resilient to focusing on surface similarity between entities than focusing on similarities between the words themselves.

Fine-tuning Models with Analogies. Yuan et al. (2024a), Chen et al. (2022), and Petersen and van der Plas (2023) tested whether analogy solving is something that can be learned with training on mapped analogies. Yuan et al. (2024a) released a dataset of analogies created from knowledge graphs (ConceptNet [Speer et al., 2017] and Wikidata [Vrandečić and Krötzsch, 2014]), and found that fine-tuning on their dataset improved analogy generation on out-of-domain semantic analogies, with T5-large (Raffel et al., 2020) gaining 44-63 percentage points on test-sets after fine-tuning as compared to the vanilla model, reaching up to 80% accuracy. On an analogy recognition task, they found that RoBERTa-large (Liu et al., 2019) and DeBERTa-v3 (He et al., 2021) trained on their data also often gained on performance.

Petersen and van der Plas (2023) attempted to address the geometry between word embeddings specifically by training models on analogies to maximize the cosine similarity between the differences of the entities, *cos(a-b, c-d)*. Notably, this is different from the vector offset method, which estimates a specific *d* term using the equation *cos(c-a+b, d')* and is generally a method for evaluation. They found that while training models to identify analogies when comparing differences between entities within the same domain (e.g., *cos(a-b,c-d)*) improved performance on this task, detecting similarities between the longer distance connections (e.g., *cos(a-c,b-d)*) did not improve with training.

Model Prompting and Few Shot Learning. Research in this area for entitly-level analogies can be split into two parts: templates for non-causal LM's such as BERT to predict entities in a cloze-style test, and engineering prompts to elicit text generation.

Ushio et al. (2021a) introduce RelBERT, a RoBERTa model fine-tuned on triples formed from the SemEval-2012 task-2 (Jurgens et al., 2012) to generate relation embeddings between two entities. They build different RelBERT models with several prompting methods, testing both manual and learned prompts (AutoPrompt [Shin et al., 2020] and P-tuning [Liu et al., 2023]). While the model was trained to generate relation embeddings apart from the analogy setting, the authors test their model on its abilities to solve analogy datasets that were not seen during training. All analogies were multiple choice, therefore did not involve any RETRIEVAL process, only MAPPING, and did not require explicit verbalization of the relation embedding that was compared. Despite solving analogies in a zero-shot setting, RelBERT with manual prompting outperformed other baselines including few-shot GPT-3. All RelBERT models regardless of prompting choice outperformed few shot GPT-3.

When probing models with their novel analogy benchmark, SCAR, Yuan et al. (2023) tested a variety of prompting methods to explore how background information and chain-of-thought (CoT) prompting (Wei et al., 2022) affect LM's reasoning abilities (in their case, large language models mostly with at least around 7B parameters such as InstructGPT). They found that including CoT prompting or providing background information improved robustness to prompt templates across 11 instruction designs. Additionally, they found that including background examples were particularly useful for the Chinese version of the SCAR across LMs, which they attribute to models' difficulties with Chinese domain-specific entities.

Model Scale. When introducing Brown et al. (2020) GPT-3, the authors tested GPT-3 models of varying size, ranging from 125M to 175B parameters on the SAT dataset. They found that model size was correlated with performance, with the biggest gains with model size demonstrated in the few-shot prompting context, where the 125M parameter model achieving around 30% accuracy and the 175B reaching 65% accuracy. In the case of fine-tuning, when the dataset is small—as is often the case with analogy datasets—increased model size may not lead to increased performance, as demonstrated in Petersen and van der Plas (2023), where BERT-base was able to improve accuracy on an analogy identification task using a relatively small training dataset, while BERT-large was unable to learn.

Comparison to Human Performance. Aside from the Google Analogy Testset and the BATS, there is no widely used benchmark to assess analogical reasoning for LMs. However, there is a decent-sized body of research on analogical reasoning performance in humans, often that provide the utilized datasets as well as estimates of human performance.

Yuan et al. (2023) found, unsurprisingly, that larger models such as GPT-4 were able to match or exceed human performance on certain semantic and morphological benchmarks. However, when using the E-KAR dataset to test models and humans performance with regards to accuracy and a relational structure identification test (which tests the ability to correctly identify the relation involved in an analogy), they found that the overlap between being able to correctly solve an analogy and correctly identify the analogous relation was lower in models than in humans. The authors additionally tested domain transfer between analogies with a dataset they released (SCAR) and found that analogies between relatively similar yet still disparate domains saw higher accuracy in cross domain transfer than between more disparate domains. Stevenson et al. (2023) tested models'

abilities to solve analogies and how they compared to the performance of children aged 7–12. They found that the LMs they tested outperformed 7-year-olds on an analogy task, and that several models such as RoBERTa and GPT-3 performed at the level of 11 year-olds. However after some investigation, they found that models may rely on associations between potential *d* terms with the given *c* to solve some analogies, as opposed to utilizing analogical reasoning by considering the *a* and *b* terms, suggesting similar issues as those presented with the vector offset method for Word2Vec despite these models being more sophisticated.

4.3 Identifying Analogies in Context

Mapping at the Entity Level within Context. Sultan and Shahaf (2022) took an algorithmic approach to MAPPING entities that are analogous between procedural texts, which they call Finding Mapping by Question. They use QA-SRL (FitzGerald et al., 2018) to generate questions and answers regarding the sentences present in the procedural texts. The questions function as a way to identify similar entities between texts, with the assumption that similar entities would have similar questions. The authors also used a clustering algorithm to address coreference issues in the texts, which, much like the Jacob et al. (2023) paper, could be considered *decomposition*. Beam search was used to finalize MAPPING between systems.

Higher Level Analogical Abstraction. Some research has attempted to identify analogies between two texts ''in general'', i.e., no specific entities or relations contained in the text are mapped, but instead multiple texts are deemed analogous at a higher level. For example, Ghosh and Srivastava (2022) released ePiC, a dataset of narratives associated with various proverbs, and tested whether models could predict the corresponding proverb, with the analogy being between the moral of the narrative and the moral of the proverb. Additionally, Nagarajah et al. (2022), Sourati et al. (2024), and Jiayang et al. (2023) try to identify the presence of analogies between entities or relations in texts, but do not specifically identify and map them. This is perhaps more relevant to the ABSTRACTION and ENCODING processes than performing MAPPING, since they identify whether the overall structure has similarities.

Contextual Analogy Generation. Another task that has been attempted is that of analogy generation—formulated as providing a source concept for which the generated text should be analogous (Bhavya et al., 2022; Ding et al., 2023; Sultan et al., 2024). Like entity-level generation, this task is related to the RETRIEVAL process, as any generated analogy would be based on knowledge the model contains.

Fine-Tuning. Unfortunately, given the relatively small size of contextual-analogy datasets (until recently), experiments exploring fine-tuning on these sorts of analogies has been minimal and not particularly informative. Jiayang et al. (2023) addressed the lack of data issue, and introduced the StoryAnalogy dataset, which includes 24K contextual analogy pairs. On an Semantic Textual Similarity (STS) style analogy identification task, they found that fine-tuning RoBERTa models improves classification ability on a style task, and that fine-tuned models performed better on their novel analogy evaluation metric than larger LLMs such as LLaMa-65B, but overall correlation with human scores had room for improvement. They also found that fine-tuning FlanT5 models (Chung et al., 2024) improved analogy generation and the novelty of the generations over the model in the few-shot setting, but that plausibility decreased with tuning.

Model Prompting. Sultan et al. (2024) introduced a data generation pipeline using GPT-3.5, ParallelParc, to generate contextual analogies. They found that prompting GPT-3.5 to generate analogies with no guidance resulted in analogies that were repetitive or between similar topics. After including the base system for which GPT-3.5 needed to create an analogical paragraph in a single prompt, the model tended to generate paragraphs for target systems that differed mostly through changing nouns of the base system. They ultimately found that a two-step prompt, one that first identified an appropriate target concept and the analogous relations, and a second that generated the text, achieved the desired results.

Among humans, it is common for teachers to use analogies to explain a newly introduced or more unfamiliar concept. Yuan et al. (2024b) tested this in the LM setting, where analogies, including long-form analogies, generated by LM's were used as prompts that included background

information on two scientific question answering datasets for student models. They found that including long-form analogies in the prompt outperformed zero-shot and CoT prompting.

Model Scale. On their StoryAnalogy dataset, Jiayang et al. (2023) found that larger models like GPT-3.5 and ChatGPT did not outperform smaller encoder models on a the STS style tasks. Combs et al. (2025) tested 13 large models, the smallest being StableLM (Bellagente et al., 2024) with 1.6B parameters and the largest with known number of parameters being GPT-4 with 1.8T- on datasets taken from Gentner et al. (1993) and Wharton et al. (1994). They found that while GPT-4, GPT-4o, and Claude were consistently among the top performers in the experiments, model performance was not completely correlated with size.

Comparisons to Human Performance. Sourati et al. (2024) tested the ability of six models, which were instructed to choose the narrative that was analogous to a given source narrative, and were presented with different combinations of near/far analogies and near/far disanalogies. They also tested SBERT, using cosine similarity between the two narratives to determine whether the pairs were analogous. This task was formulated as a binary analogy selection task, and as mentioned before, arguably addresses ABSTRACTION and ENCODING instead of MAPPING.

While none of the models were able to match human performance, GPT-4 was able to approach human performance when the true analogy option was a near analogy. On average, models performed better when the true analogy was near and the disanalogy was far, and the worst when the true analogy was far, and the distractor was near.

5 Applications of Analogies

Analogical reasoning is not just an exercise in relational thinking performed for fun or used as a broad test of intelligence, it is a tangible way to formulate many domain-specific tasks. A perhaps obvious one would be for creative processes, such as metaphor generation in creative writing (Gentner et al., 2001).

Additionally analogy is often used in education, from early education to medical school (Heywood, 2002; Guerra-Ramos, 2011; Pena and de Souza Andrade-Filho, 2010; Gray and Holyoak, 2021). It also has application in health and medicine (Alsaidi et al., 2022; Guallart, 2014), science communication (Schwarz-Plaschg, 2018; Corner and Pidgeon, 2015; Elliott, 2016), innovation and creative problem solving (Gick and Holyoak, 1980; Hope et al., 2017; Markman et al., 2009), and law (Lamond, 2016; Condello, 2016).

5.1 Broader Incorporation of Analogy and Analogical Thinking in NLP

We would like to discuss potential avenues for incorporating analogical reasoning in NLP at large. Analogical reasoning can potentially address certain known issues in NLP, such as spurious correlations, bias, out of domain generalization, and explainability.

When discussing different categories of explainability methods, Lyu et al. (2024) cover similarity-based methods, where similar previous examples are used to justify why a model made a decision given the current input. They describe these methods as being similar to how humans use analogy. Identifying relevant previous examples is related to the RETRIEVAL process of analogical reasoning, however, these methods can be subject to spurious correlation, perhaps much like the RETRIEVAL processes in humans can be effected by surface similarity. Additionally, much like the research with contextual analogies detailed in Section 4.3, these methods do not always map between specific spans of text. One way to improve the reliability and transparency of similarity-based methods would be to incorporate a MAPPING process as opposed to just a RETRIEVAL process, where specific entities, concepts, or themes in the retrieved text can be mapped to the target text. Ming et al. (2019) introduced an explainability method, ProSeNet, which learns prototypical examples that are then used to explain predictions on subsequent data, highlighting relevant text. However, the prototypical examples learned are not necessarily actual examples present in the dataset, and are not retrieved on a case by case basis. This might be relevant for some applications that require more specificity, such as identifying similar ruling for individual court rulings in the law domain.

Analogical reasoning could also be used to address out-of-domain generalization and transfer learning. Transfer of knowledge or expertise in

humans is an area of cognitive science that is heavily investigated, often in the context of analogical reasoning, and generally involves the ENCODING and RETRIEVAL processes (Kimball and Holyoak, 2000). In order to transfer knowledge, one must be able to retrieve the relevant instance in memory. The ability to both retrieve and then transport the solution to a new situation depends on how both the original and target instances are encoded (Gick and Holyoak, 1983; Loewenstein, 2010). Understanding text not merely as a sum of its parts, but as representing ideas and solutions that can be generalized to alternate domains or situations, and allowing these representations to be updated or adapted when presented with new information, is arguably vital to cross-domain generalization (Doumas et al., 2022). Furthermore, there may be domains that have a scarce availability of training data, where being able to identify generalizable, domain agnostic abstractions could be beneficial. This does not have to be limited to the training or evaluation stage, for example, Huang et al. (2024) presented Analogical Reasoning-Augmented Interactive Data Annotation (ARAIDA), to address the annotation stage.

A known problem in NLP is the tendency of models to rely on spurious correlations, where certain words or terms are heavily associated with a label despite being irrelevant to the task at hand (Wang and Culotta, 2020). This can be thought of as the model relying on surface similarity to perform tasks. In order to overcome this, models need to learn to identify and attend to more abstract and relational information and higher level perceptual information contained in text (Chalmers et al., 1992).

Somewhat related is the relevance of multimodel input, situational learning, and incorporating our various human senses and perceptual capabilities into processing and understanding language data, which are ways in which human learning and machine learning fundamentally differ (Frank, 2023; Beuls and Van Eecke, 2024). Arguably all the processes involved in analogical reasoning (and cognitive processes in general) in humans involve incorporating perceptual data (Mitchell, 2021). Kotovsky and Fallside (2013) and Zamani and Richard (2000) suggested that how a concept is encoded is important to a cognitive system's ability to recognize analogy with visual stimuli. For humans, context in language is not limited purely to distributional semantics. Language takes place in the context of all our senses and interactions. With that said, this can potentially be a double edged sword. While additional, non-textual information is influential in helping us understand language, there is also the issue mentioned in Section 2.4, where external stimuli can also be distracting for quality ABSTRACTION of ideas.

Lastly, current NLP tasks could be reframed in analogical way, which has already been done in the literature to some extent. For example, Wang and Lepage (2020) formulated sentence generation as a sentence analogy task, where the desired sentences to be generated *d* were edited versions of sentences *c*, where the necessary edits were the differences between two other given sentences *a* and *b*. Wijesiriwardene et al. (2024, 2023a) reformulated natural language inference, specifically the entailment and negation labels, as solving analogy problems. One could also argue that coreference resolution is related to the REREPRESENTATION process, and that the methods outlined in Yan et al. (2003) could be applied to approaches addressing coreference resolution to tackle other tasks such as natural language inference.

6 Conclusion and Limitations

In this paper, we review the processes of analogical reasoning in humans and connect them with current research and methods in NLP. We found that experiments with certain types of analogies typically focus on specific processes (e.g., entity level and mapping). Additionally, we also suggest that analogical reasoning could potentially be an approach to address current limitations of NLP models, and be a way for researchers to focus more on optimizing relational understanding and similarity rather than relying on entity similarity.

One major limitation of this work is that we just brushed the surface of research in cognitive science regarding analogical reasoning. Other areas of analogical reasoning that could also be of interest to the NLP community which were not touched upon would be the development of analogical and relational reasoning over the humans lifespan and how this could help improve analogical transfer.

Additionally, we did not address research with knowledge graphs, relation embeddings, and other similar areas of research which are often incorporated with NLP to create more knowledge-rich representation. However, given the limitations of

manually crafted knowledge bases, such as being ultimately limited in scope and time-intensive to build (Schwartz and Gomez, 2009; Yuan et al., 2024a), being able to understand and extract relations in text can ultimately not depend on them.

Acknowledgments

We are grateful to the Swiss National Science Foundation (grant 205121_207437: C-LING) and the Fondazione Aldo e Cele Daccò for funding this work. Part of this work was done at the Idiap Research Institute in Martigny. We thank members of the Idiap NLU-CCL group for helpful discussions, and the anonymous reviewers for their fruitful comments and suggestions.

References

Safa Alsaidi, Miguel Couceiro, Esteban Marquer, Sophie Quennelle, Anita Burgun, Nicolas Garcelon, and Adrien Coulet. 2022. An analogy based framework for patient-stay identification in healthcare. In *ATA@ ICCBR 2022-Workshop Analogies: from Theory to Applications*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Christian Antić. 2022. Analogical proportions. *Annals of Mathematics and Artificial Intelligence*, 90(6):595–644. https://doi.org/10.1007/s10472-022-09798-y

Jennifer Asmuth and Dedre Gentner. 2017. Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology*, 70(10):2007–2025. https://doi.org/10.1080/17470218.2016.1219752, PubMed: 27485316

Susan M. Barnett and Stephen J. Ceci. 2002. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637. https://doi.org/10.1037/0033-2909.128.4.612, PubMed: 12081085

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, Meng Lee, Emad Mostaque, Michael Pieler, Nikhil Pinnaparju, Paulo Rocha, Harry Saini, Hannah Teufel, Niccolo Zanichelli, and Carlos Riquelme. 2024. Stable LM 2 1.6b technical report.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, pages 1–34. https://doi.org/10.1162/coli_a_00534

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy generation by prompting large language models: A case study of InstructGPT. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 298–312, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.inlg-main.25

Andrew Blair-Stanek and Benjamin Van Durme. 2021. AI for tax analogies and code renumbering.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Silvia A. Bunge, Carter Wendelken, David Badre, and Anthony D. Wagner. 2005. Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, 15(3):239–249. https://doi.org/10.1093/cercor/bhh126, PubMed: 15238433

Jonathan Cagan, Joel Chan, Katherine Fu, Christian Schunn, Kristin Wood, and Kenneth Kotovsky. 2011. On the effective use of design-by-analogy: The influences of analogical distance and commonness of analogous designs on ideation performance. In *DS 68-7: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design,*

Vol. 7: Human Behaviour in Design, Lyngby/ Copenhagen, Denmark, 15.-19.08. 2011, pages 85–96.

David J. Chalmers, Robert M. French, and Douglas R. Hofstadter. 1992. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3):185–211. https://doi.org/10.1080/09528139208953747

Joel Chan and Christian Schunn. 2015. The impact of analogies on creative concept generation: Lessons from an in vivo study in engineerzing design. *Cognitive Science*, 39(1):126–155. https://doi.org/10.1111/cogs.12127, PubMed: 24835377

Kevin Chan, Shane Peter Kaszefski-Yaschuk, Camille Saran, Esteban Marquer, and Miguel Couceiro. 2022. Solving morphological analogies through generation. In *IJCAI-ECAI Workshop on the Interactions between Analogical Reasoning and Machine Learning (IARML@ IJCAI-ECAI 2022)*, volume 3174 of *Proceedings of the IJCAI-ECAI Workshop on the Interactions between Analogical Reasoning and Machine Learning (IARML@IJCAI-ECAI 2022)*, pages 29–39.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics. 10.18653/v1/2022.findings-acl.311

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Kara Combs, Trevor Bihl, Spencer Howlett, and Yuki Adams. 2025. Zero-shot comparison of large language models (LLMs) reasoning abilities on long-text analogies. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2025.194

Angela Condello. 2016. Metaphor as analogy: Reproduction and production of legal concepts. *Journal of Law and Society*, 43(1):8–26. https://doi.org/10.1111/j.1467-6478.2016.00738.x

Adam Corner and Nick Pidgeon. 2015. Like artificial trees? The effect of framing by natural analogy on public perceptions of geoengineering. *Climatic Change*, 130:425–438. https://doi.org/10.1007/s10584-014-1148-6

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.153

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.482

Zijian Ding, Arvind Srinivasan, Stephen Macneil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. *Proceedings of the 15th Conference on Creativity and Cognition*, pages 489–505. https://doi.org/10.1145/3591196.3593516

Leonidas A. A. Doumas, John E. Hummel, and Catherine M. Sandhofer. 2008. A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1):1. https://doi.org/10.1037/0033-295X.115.1.1, PubMed: 18211183

Leonidas A. A. Doumas, Guillermo Puebla, Andrea E. Martin, and John E. Hummel. 2022. A theory of relation learning and cross-domain generalization. *Psychological Review*, 129(5):999–1041. https://doi.org/10.1037/rev0000346, PubMed: 35113620

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king − man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Karl Duncker and Lynne S. Lees. 1945. On problem-solving. *Psychological Monographs*, 58(5):i–113. https://doi.org/10.1037/h0093599

Richard Elliott. 2016. Communicating Biological Sciences: Ethical and Metaphorical Dimensions. Routledge. https://doi.org/10.4324/9781315572888

Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63. https://doi.org/10.1016/0004-3702(89)90077-5

Gertraud Fenk-Oczlon, August Fenk, and Pamela Faber. 2010. Frequency effects on the emergence of polysemy and homophony. *International Journal of Information Technologies and Knowledge*, 4(2):103–109.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060. https://doi.org/10.18653/v1/P18-1191

Kenneth D. Forbus, Dedre Gentner, Arthur B. Markman, and Ronald W. Ferguson. 1998. Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(2):231–257. https://doi.org/10.1080/095281398146842

Michael C. Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992. https://doi.org/10.1016/j.tics.2023.08.007, PubMed: 37659919

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Dedre Gentner. 1987. Mechanisms of analogical learning. Technical Report UIUCDCS-R-87-1381 / UICU-ENG-87-1770, University of Illinois at Urbana-Champaign, Department of Psychology, Champaign, IL.

Dedre Gentner, Brian Bowdle, Phillip Wolff, and Consuelo Boronat. 2001. *The Analogical Mind: Perspectives from Cognitive Science*, chapter Metaphor is like Analogy. The MIT Press. https://doi.org/10.7551/mitpress/1251.003.0010

Dedre Gentner and Kenneth D. Forbus. 2011. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):266–276. https://doi.org/10.1002/wcs.105, PubMed: 26302075

Dedre Gentner and Ilene M. France. 1988. The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In *Lexical Ambiguity Resolution*. Elsevier, pages 343–382. https://doi.org/10.1016/B978-0-08-051013-2.50018-5

Dedre Gentner and Christian Hoyos. 2017. Analogy and abstraction. *Topics in Cognitive Science*, 9(3):672–693. https://doi.org/10.1111/tops.12278, PubMed: 28621480

Dedre Gentner and Jeffery Loewenstein. 2002. *Encyclopedia of Education, Second Edition*, chapter Learning: Analogical reasoning.

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2):393. https://doi.org/10.1037/0022-0663.95.2.393

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2004. Analogical encoding: Facilitating knowledge transfer and integration. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Dedre Gentner and Mary Jo Rattermann. 1991. Language and the career of similarity. Center for the Study of Reading Technical Report; no. 533. https://doi.org/10.1017/CBO9780511983689.008

Dedre Gentner, Mary Jo Rattermann, and Kenneth D. Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25(4):524–575. https://doi.org/10.1006/cogp.1993.1013, PubMed: 8243045

Dedre Gentner and Linsey Smith. 2012. Analogical Reasoning. Oxford, UK: Elsevier. pages 130–136. https://doi.org/10.1016/B978-0-12-375000-6.00022-7

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.276

Mary L. Gick and Keith J. Holyoak. 1980. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355. https://doi.org/10.1016/0010-0285(80)90013-4

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

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2002

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. 2024. The Llama 3 herd of models. https://arxiv.org/abs/2407.21783

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maureen E. Gray and Keith J. Holyoak. 2021. Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, 15(3):250–263. https://doi.org/10.1111/mbe.12288

Nino Guallart. 2014. Analogical reasoning in clinical practice. *Systematic Approaches to Argument by Analogy*, pages 257–273.

https://doi.org/10.1007/978-3-319-06334-8_15

Maria Teresa Guerra-Ramos. 2011. Analogies as tools for meaning making in elementary science education: How do they work in classroom settings? *Eurasia Journal of Mathematics, Science and Technology Education*, 7(1):29–39. https://doi.org/10.12973/ejmste/75175

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *The International Conference on Learning Representations (ICLR)*.

Dave Heywood. 2002. The place of analogies in science education. *Cambridge Journal of Education*, 32(2):233–247. https://doi.org/10.1080/03057640220147577

Damian Hodel and Jevin West. 2024. Response: Emergent analogical reasoning in large language models.

Douglas R. Hofstadter. 2001. *Epilogue: Analogy as the Core of Cognition*. The MIT Press. https://doi.org/10.7551/mitpress/1251.003.0020

Douglas R. Hofstadter, Melanie Mitchell. 1995. The copycat project: A model of mental fluidity and analogy-making. *Advances in Connectionist and Neural Computation Theory*, 2:205–267.

Keith J. Holyoak and Kyunghee Koh. 1987. Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4):332–340. https://doi.org/10.3758/BF03197035, PubMed: 3670053

Keith J. Holyoak and Paul Thagard. 1989a. Analogical mapping by constraint satisfaction. *Cognitive Science*, 13(3):295–355. https://doi.org/10.1207/s15516709cog1303_1

Keith J. Holyoak and Paul Thagard. 1989b. A computational model of analogical problem solving. *Similarity and Analogical Reasoning*, 242266. https://doi.org/10.1017/CBO9780511529863.012

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 235–243, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3097983.3098038

Chen Huang, Yiping Jin, Ilija Ilievski, Wenqiang Lei, and Jiancheng Lv. 2024. ARAIDA: Analogical reasoning-augmented interactive data annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10660–10675, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.574

John E. Hummel and Keith J. Holyoak. 1997. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427. https://doi.org/10.1037/0033-295X.104.3.427

Shahar Jacob, Chen Shani, and Dafna Shahaf. 2023. Fame: Flexible, scalable analogy mappings engine. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.1023

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.706

Lara L. Jones, Matthew J. Kmiecik, Jessica L. Irwin, and Robert G. Morrison. 2022. Differential effects of semantic distance, distractor salience, and relations in verbal analogy. *Psychonomic Bulletin & Review*, 29(4):1480–1491. https://doi.org/10.3758/s13423-022-02062-8, PubMed: 35132581

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity.

In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2905

Daniel R. Kimball and Keith J. Holyoak. 2000. Transfer and expertise. *The Oxford Handbook of Memory*, 109–122. https://doi.org/10.1093/oso/9780195122657.003.0007

Boicho Kokinov and Robert M. French. 2003. Computational models of analogy-making. *Encyclopedia of Cognitive Science*, 1:113–118.

Kenneth Kotovsky and David Fallside. 2013. Representation and transfer in problem solving. In *Complex Information Processing*, pages 89–128. Psychology Press.

Arjun Sai Krishnan and Seyoon Ragavan. 2021. Morphology-aware meta-embeddings for Tamil. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 94–111, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-srw.13

Grant Lamond. 2016. Precedent and Analogy in Legal Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2016 edition. Metaphysics Research Lab, Stanford University.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1618

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *preprint on arXiv 2402.08955*, abs/2402.08955.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2503

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *arXiv:2103.10385*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv: 1907.11692v1*.

Jeffrey Loewenstein. 2010. How one's hook is baited matters for catching an analogy, *Psychology of Learning and Motivation*, volume 53, pages 149–182. Elsevier. https://doi.org/10.1016/S0079-7421(10)53004-4

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. volume 50, pages 657–723, Cambridge, MA. MIT Press. https://doi.org/10.1162/coli_a_00511

Jean M. Mandler and Felice Orlich. 1993. Analogical transfer: The roles of schema abstraction and awareness. *Bulletin of the Psychonomic Society*, 31(5):485–487. https://doi.org/10.3758/BF03334970

Arthur B. Markman, Kristin L. Wood, Julie S. Linsey, Jeremy T. Murphy, and Jeffrey P. Laux. 2009. *Tools for Innovation: The Science Behind the Practical Methods that Drive New Ideas*, chapter Supporting innovation by promoting analogical reasoning. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195381634.003.0005

Esteban Marquer, Pierre-Alexandre Murena, and Miguel Couceiro. 2022. Transferring learned

models of morphological analogy. In *ATA@ ICCBR2022-Analogies: From Theory to Applications (ATA@ ICCBR2022)*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Iprocessing Systems*, 26.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Ricardo A. Minervino and Máximo Trench. 2024. Surface matches prevail over distant analogs during retrieval. *Memory & Cognition*, 53(3):775–791. https://doi.org/10.3758/s13421-024-01605-9, PubMed: 38992247

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 903–913. ACM. https://doi.org/10.1145/3292500.3330908

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101. https://doi.org/10.1111/nyas.14619, PubMed: 34173249

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. 2024. Semantic structure-mapping in LLM and human analogical reasoning. *arXiv 2406.13803*. https://doi.org/10.48550/arXiv.2406.13803

Thiloshon Nagarajah, Filip Ilievski, and Jay Pujara. 2022. Understanding narratives through dimensions of analogy. *arXiv 2206.07167*. https://doi.org/10.48550/arXiv.2206.07167

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, et al. 2024. GPT-4 technical report.

Gil Patrus Pena and José de Souza Andrade-Filho. 2010. Analogies in medicine: Valuable for learning, reasoning, remembering and naming. *Advances in Health Sciences Education*, 15:609–619. https://doi.org/10.1007/s10459-008-9126-2, PubMed: 18528776

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? Investigating training objectives and comparisons to human performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16414–16425, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1022

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Michael Ramscar and Daniel Yarlett. 2003. Semantic grounding in models of analogy: An environmental approach. *Cognitive Science*, 27(1):41–71. https://doi.org/10.1207/s15516709cog2701_2

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-1017

H. Andrew Schwartz and Fernando Gomez. 2009. Acquiring applicable common sense knowledge from the web. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pages 1–9. https://doi.org/10.3115/1641968.1641969

Claudia Schwarz-Plaschg. 2018. Nanotechnology is like. . . The rhetorical roles of analogies in public engagement. *Public Understanding of Science*, 27(2):153–167. https://doi.org/10.1177/0963662516655686, PubMed: 27412576

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.346

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. In *ARN: Analogical Reasoning on Narratives*. volume 12, pages 1063–1086, Cambridge, MA. MIT Press. https://doi.org/10.1162/tacl_a_00688

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press. https://doi.org/10.1609/aaai.v31i1.11164

Claire E. Stevenson, Mathilde ter Veen, Rochelle Choenni, Han L. J. van der Maas, and Ekaterina Shutova. 2023. Do large language models solve verbal analogies like children do? *arXiv 2310.20384*. https://doi.org/10.48550/arXiv.2310.20384

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. ParallelPARC: A scalable pipeline for generating natural-language analogies. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5900–5924, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.329

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.232

Paul Thagard, Keith J. Holyoak, Greg Nelson, and David Gochfeld. 1990. Analog retrieval by constraint satisfaction. *Artificial Intelligence*, 46(3):259–310. https://doi.org/10.1016/0004-3702(90)90018-U

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. In *RANLP*, pages 101–110. https://doi.org/10.1075/cilt.260.11tur

Santosh T.y.s.s., Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306, Torino, Italia. ELRA and ICCL. https://doi.org/10.63317/544mxp9oahts

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.712

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.280

Michael S. Vendetti, Aaron Wu, and Keith J. Holyoak. 2014. Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, 25(4):928–933. https://doi.org/10.1177/0956797613518079, PubMed: 24463552

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. https://doi.org/10.1145/2629489

Caren M. Walker, Samantha Q. Hubachek, and Michael S. Vendetti. 2018. Achieving abstraction: Generating far analogies promotes relational reasoning in children. *Developmental Psychology*, 54(10):1833. https://doi.org/10.1037/dev0000581, PubMed: 30234337

Liyan Wang and Yves Lepage. 2020. Vector-to-sequence models for sentence analogies. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 441–446. IEEE. https://doi.org/10.1109/ICACSIS51025.2020.9263191

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.308

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541. https://doi.org/10.1038/s41562-023-01659-w, PubMed: 37524930

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Charles M. Wharton, Keith J. Holyoak, Paul E. Downing, Trent E. Lange, Thomas D. Wickens, and Eric R. Melz. 1994. Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, 26(1):64–101. https://doi.org/10.1006/cogp.1994.1003

T. Wijesiriwardene, A. Sheth, V. L. Shalin, and A. Das. 2023a. Why do we need neurosymbolic AI to model pragmatic analogies? *IEEE Intelligent Systems*, 38(05):12–16. https://doi.org/10.1109/MIS.2023.3305862

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023b. Analogical–a novel benchmark for long text analogy evaluation in large language models. *Findings of the Association for Computational Linguistics: ACL 2023*. https://doi.org/10.18653/v1/2023.findings-acl.218

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2024. On the relationship between sentence analogy identification and sentence structure encoding in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 451–457, St. Julian's, Malta. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-eacl.31

Hiroaki Yamagiwa, Ryoma Hashimoto, Kiwamu Arakane, Ken Murakami, Shou Soeda, Momose Oyama, Mariko Okada, and Hidetoshi Shimodaira. 2024. Predicting drug-gene relations via analogy tasks with word embeddings. *Scientific Reports*, 15(1):17240. https://doi.org/10.1038/s41598-025-01418-z, PubMed: 40383732

Jin Yan, Kenneth D. Forbus, and Dedre Gentner. 2003. A theory of rerepresentation in analogical matching, *Proceedings of the 25th Annual Cognitive Science Society*, pages 1265–1270. Psychology Press. https://doi.org/10.21236/ADA466013

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2446–2460, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.160

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024a. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.68

Siyu Yuan, Cheng Jiayang, Lin Qiu, and Deqing Yang. 2024b. Boosting scientific concepts understanding: Can analogy from teacher models empower student models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6026–6036, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.346

Mojdeh Zamani and Jean-François Richard. 2000. Object encoding, goal similarity, and analogical transfer. *Memory & Cognition*, 28:873–886. https://doi.org/10.3758/BF03198422, PubMed: 10983461