A Systematic Review of Analogy Generation and Evaluation: Methods, Metrics, and Challenges

Anonymous ACL submission

Abstract

Analogy is a quintessential human cognitive capability that supports knowledge transfer across domains. While researchers have long studied analogy generation, evaluation, and reasoning through symbolic and connectionist approaches, the rise of large language models (LLMs) has introduced new generation systems and evaluation methods. However, a systematic understanding of this emerging landscape and its connection to prior work remains underexplored. To address this gap, we conduct a PRISMA-based systematic literature review spanning four decades across six databases. We identify three key observations: first, generation and evaluation methods are strongly shaped by analogy granularity, such as lexical versus compositional analogies; second, human-in-the-loop design plays an important role in LLM-based analogy generation systems; and third, the field remains divided over whether LLMs perform genuine analogical reasoning. Based on our review, we propose practical guidelines for future work on analogy generation, including potential evaluation dimensions for lexical and compositional tasks.

1 Introduction

Analogy plays a fundamental role in human learning by helping individuals understand unfamiliar concepts through familiar ones (Gentner and Smith, 2013; Bartha, 2024). Computational approaches to analogy generation, evaluation, and reasoning have long attracted interest in artificial intelligence (AI) and natural language processing (NLP).

Earlier symbolic approaches, such as Winston's frame-based system (Winston, 1980) and the Structure-Mapping Engine (Falkenhainer et al., 1989), represented analogies in logical form and evaluated them through relational alignment on hand-crafted examples. Later connectionist approaches learned analogical relationships from data and were evaluated on word- and sentence-level analogy datasets (Turney, 2008; Mikolov et al., 2013). More recently, pretrained large language models (LLMs), such as GPT (Ouyang et al., 2022), have extended analogy generation to longer compositional forms, supporting applications such as analogy mining (Bhavya et al., 2023) and personalized learning (Cao et al., 2024).

Evaluating analogies remains a central challenge. Structure Mapping Theory (Gentner, 1983), which characterizes analogy as the mapping of relational structure between a source and target domain, provides a foundation for evaluation through relational alignment. Automatic methods have largely relied on relational and word similarity metrics (Turney et al., 2006), often complemented by human evaluation through expert judgment or crowdsourcing (Sultan et al., 2024). More recently, LLM-based hybrid evaluation strategies have also emerged (Bhavya et al., 2024a).

Prior surveys have examined symbolic and neural approaches to analogy (Mitchell, 2021; Gentner and Forbus, 2011; French, 2002); however, the rapid emergence of LLM-based methods has since reshaped the field. Recent LLM-era studies (Bhavya et al., 2022; Yuan et al., 2024a) have made significant advances in areas such as prompt engineering, dataset construction, and downstream applications. Yet as these efforts grow across tasks and applications, the field lacks a synthesis of how LLM-based approaches relate to one another and to earlier symbolic and neural systems. This review seeks to provide such a synthesis, situating LLM-based analogy research within the field's broader historical trajectory and identifying trends and challenges of analogy generation and evaluation.

In this work, we present a systematic literature review (SLR) of computational approaches to analogy generation and evaluation in AI and NLP, following the PRISMA guidelines (Moher et al., 2010). Our review addresses three research questions:

- **RQ1:** What computational methods have been developed for analogy generation?
- **RQ2:** What existing methods are used to evaluate the analogy's quality?
- **RQ3:** What are the key challenges, limitations, and future directions for analogy generation and evaluation?

We systematically searched and screened 4,641 papers across six databases in two rounds, resulting in a final corpus of 65 papers spanning four decades of computational analogy research.

Our review makes three main contributions: (1) we conduct a PRISMA-based SLR on analogy generation and evaluation techniques; (2) we synthesize three observations across the corpus: analogy granularity shapes how generation and evaluation methods are designed, human-in-the-loop paradigm plays an important role in LLM-based analogy generation systems, and the field remains divided over whether LLMs perform genuine analogical reasoning; and (3) we propose practical guidelines for future work, including evaluation dimensions for both lexical and compositional analogy generation tasks.

2 Related Work

2.1 Analogy Generation and Evaluation

Computational models of analogy date back to the 1980s, beginning with Winston's model (Winston, 1980) and followed by the influential SMT (Gentner, 1983). These early works focused on modeling human analogical reasoning and investigating how computational systems could replicate this process to retrieve and evaluate analogies.

Early generation methods relied on handcrafted rules (Falkenhainer et al., 1989), evolving through symbolic structure-matching (Turney, 2008) and statistical embeddings (Mikolov et al., 2013) to prompt-based LLM pipelines (Ding et al., 2023; Yuan et al., 2024a). Each paradigm reduces reliance on handcrafted representations while introducing new challenges in coherence, relational fidelity, and bias (Wijesiriwardene et al., 2023; Yuan et al., 2023; Bhavya et al., 2024a). Evaluation has followed a parallel trajectory: from rule-based soundness checks (Holyoak and Thagard, 1989; Falkenhainer et al., 1989) through automatic metrics (precision, recall, BLEU, ROUGE) to hybrid protocols that pair lexical-level automatic similarity with compositional-level human judgment (Yuan et al., 2023; Jiayang et al., 2023).

2.2 SLRs in Computational Analogy Model

Prior surveys have framed computational analogy models in complementary ways. French (2002) provides a historical overview, classifying computational analogy models into symbolic, connectionist, and hybrid paradigms. By contrast, Gentner and Forbus (2011) analyzed analogies through the lens of computational models. The work decomposes the analogy into subprocesses: retrieval, mapping, abstraction, and re-representation. Gentner and Forbus (2011) emphasizes that analogical mappings favor systematic and higher-order relational correspondences. Both reviews (French, 2002; Gentner and Forbus, 2011), along with several other works (Gentner, 1983; Hofstadter, 1995), underscore that analogical inference relies on relational representations but differ in focus: French (2002) surveys broad model families and open problems, whereas Gentner and Forbus (2011) dives into the computational models of mapping.

Mitchell (2021) brings a recent AI perspective, noting that today's systems "are almost entirely lacking the ability" to form humanlike abstractions or analogies. However, these reviews (French, 2002; Gentner and Forbus, 2011; Mitchell, 2021) predate the recent explosion of neural generative models for language and reasoning. In the LLM era, pretrained transformers can generate seemingly analogies. For example, Bhavya et al. (2022) demonstrates that InstructGPT (Ouyang et al., 2022) can be prompted to produce meaningful conceptual analogies and explanations: with careful prompts, LLMs can achieve near-human quality on analogy-generation tasks. At the same time, such a method exposes new challenges: for instance, evaluating the creativity and validity of LLM-generated analogies (beyond lexical pattern matching) requires new benchmarks and human judgments, not addressed in classical frameworks.

In summary, these reviews make valuable contributions by elucidating foundational theories, categorizing early computational models, and highlighting key cognitive mechanisms involved in analogical reasoning. However, there is a lack of systematic research on the role of large language models or generative approaches in analogy generation or evaluation. This gap motivates the need for a new, systematic review that bridges classic symbolic and

connectionist theories with recent LLM-based and deep generative methods for analogy generation and evaluation.

3 Methodology

3.1 Identification

Following PRISMA (Moher et al., 2010), we first used abstract, title, and keyword (ATK) search among the online NLP and other CS databases including ACM Digital Library[1], IEEE Xplore[2], SpringerLink[3], ScienceDirect[4], Wiley Online Library[5], and ACL Anthology[6] (Referred to as ACL throughout this paper). These sources provide comprehensive coverage of both foundational and emerging research on computational analogy generation and evaluation.

For the keyword search, we used *analogy*, *analogous*, and *analogical* to capture common variants of the term. We excluded the related term *metaphor* because our preliminary review suggested that analogy and metaphor have developed into distinct research areas, each with its own theoretical foundations and frameworks (Rai and Chakraverty, 2020). We then combined these terms with *generation*, *retrieval*, and *evaluation* to cover the main computational subprocesses.

3.2 Screening, Eligibility, and Inclusion

3.2.1 Inclusion and Exclusion Criteria

- Include: **IC1**: Published between 1980 and 2025 to ensure we cover established computational models(Winston, 1980; Falkenhainer et al., 1989); **IC2**: The research topic is primarily in NLP/AI/CS, and the contribution is relevant to computational analogy generation and its evaluation; this could be proposing novel systems or improvements upon previous works.

- Exclude: **EC1**: The paper is grey literature, such as a work-in-progress, workshop, poster, demo, an extended abstract, or a patent (Handoyo and Sensuse, 2017). **EC2**: The paper is not written in English; **EC3**: The paper is not archival; **EC4**: The computational method lacks a concrete artifact (e.g., system, algorithm) or relies solely on human labor (e.g., crowdsourcing), since we focus on computational methods and systems. **EC5**: The paper constitutes solely of secondary studies, as our focus is on the existing methods in primary research (Handoyo and Sensuse, 2017). **EC6**: None of the paper's claimed contributions concern analogy generation or the evaluation process.

3.2.2 Process

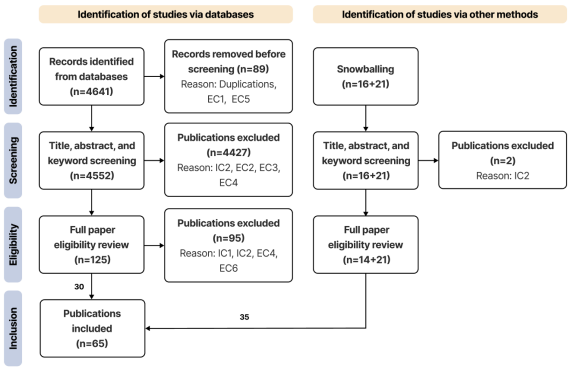

Figure 1: Pipeline of PRISMA framework used in this review process.

The initial keyword search across six databases yielded 4,641 papers. After removing duplicates, papers published before 1980, non-full papers, and secondary studies, 4,552 papers remained and were subsequently included in the screening process. In the ATK screening stage, four authors randomly sampled 50 papers for a pilot round. Each author independently labeled each paper as *include*, *exclude*, or *uncertain*. The authors then resolved discrepancies and refined the inclusion and exclusion criteria. To ensure consistency, each paper was reviewed by at least two authors. This stage resulted in 125 papers. Next, a full-text screening was conducted to evaluate the remaining papers. Another pilot round (N=20) was conducted to further refine the inclusion and exclusion criteria. After this step, 31 papers met the inclusion criteria and were kept for analysis (See Fig. 1 for the entire pipeline). One round of backward snowballing was conducted to identify relevant studies that may have been missed during the initial search (Jalali and Wohlin, 2012), resulting in 13 additional papers identified.

We conducted the initial PRISMA search and screening in April of 2025. Given the rapid pace of research in this area, we conducted a second

[1]https://dl.acm.org/
[2]https://ieeexplore.ieee.org/
[3]https://link.springer.com/
[4]https://www.sciencedirect.com/
[5]https://onlinelibrary.wiley.com/
[6]https://aclanthology.org/

round of forward and backward snowballing in April of 2026 to identify relevant recent studies. We applied the same inclusion and exclusion criteria and identified 21 additional papers, bringing the final corpus to 65 papers.

3.3 Data Extraction and Analysis

Data were extracted during the full-text eligibility review process. This included top-down coding for the following dimensions: methodology (RQ1), evaluation metrics (RQ2), identified challenges (RQ3), and outcomes (RQ3). To comprehensively analyze current approaches, we employed open coding and affinity diagramming techniques (Dam and Siang, 2022; Hudson, 2013; Spencer, 2009) to categorize the identified generation and evaluation methods. The extracted data were grouped and cross-validated by three authors using Miro[7], an online collaborative whiteboarding platform (Zhang et al., 2025). We held meetings to resolve disagreements and refine the groupings.

4 Results

Based on the 65 papers we examined in our SLR, we identified common themes and dimensions among them. We present our paper selections and summary statistics in Appendix B.

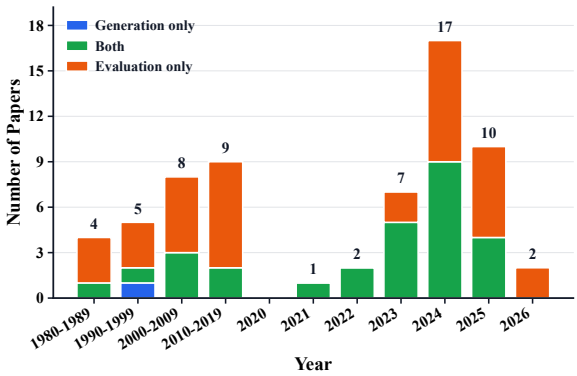

Figure 2: Publication years.

Fig. 2 shows the publication-year distribution of the papers included in our SLR. The field saw a gradual increase in publications from the 1980s through the 2010s, followed by a decline in the early 2010s. This decline may partly be an artifact of our inclusion and exclusion criteria: although embedding-based models (Mikolov et al., 2013) achieved strong performance on analogy tasks during this period, we excluded many such papers when their primary contribution centered on embedding methods rather than analogy generation or evaluation. More recently, the emergence of LLMs (Ouyang et al., 2022) has coincided with renewed interest in computational analogy research.

4.1 Analogy Representation

Before examining analogy generation and evaluation, we first characterize the representational assumptions underlying some systems in our corpus. Although representation is not posed as a standalone research question, it provides the basis for interpreting both RQ1 and RQ2: generation methods differ in what form of analogical structure they operate over, and evaluation methods differ in what kinds of relational correspondence they can measure. Across the reviewed papers, we identify four main paradigms from systems that explicitly store or process analogies.

Relational graph-based models use graph-based structured representations (e.g., knowledge graphs, ontologies) to align source and target domains. A canonical example is the Structure-Mapping Engine (Falkenhainer et al., 1989), which evaluates analogies by graph matching under one-to-one structural consistency constraints. *Distributional semantic* models estimate relational similarity from corpus statistics, such as word-embedding offsets that capture simple `A:B::C:D` relations (Turney et al., 2006; Mikolov et al., 2013). *Cognitive architecture* models, such as Copycat and LISA (Hofstadter and Mitchell, 1994; Hummel and Holyoak, 2019), model analogy through symbolic processing. Finally, *transformation-based* models treat analogies as abstract transformations, using vector-based operations to map between domains (Lepage and Ando, 1996; Plate, 2000).

We identified 17 papers that fit this taxonomy (see Tab. 3). Six use relational graph-based methods, seven use distributional semantic methods, two use cognitive architecture models, and two use transformation-based models. This distribution reflects a broader historical shift in computational analogy research. Earlier work tends to rely on explicit, rule-based representations, including relational graph-based and cognitive architecture models, where analogical structure is directly encoded and manipulated. As machine learning techniques developed, distributional semantic methods became more prominent, replacing hand-specified relational structures with corpus-derived represen-

[7] https://miro.com/

tations of similarity. Many LLM-based papers in our corpus, however, do not fit cleanly into these four representation types because they rely on the model's internal contextual representations and generative behavior rather than an explicitly specified analogy representation.

4.2 Analogy Generation Framework

Generation by Granularity Based on our review, we identified two common targets of analogy generation, distinguished by their granularity and the structure of the generated output. *Lexical-level* analogy generation focuses on producing analogies involving individual words or short phrases (e.g., `king:man::queen:woman`) (Bourrelly et al., 1983). We also include LLM-based work in this category for the word-by-word prediction nature of the method (Bhavya et al., 2024b). The second category is *compositional-level* analogy generation, which involves generating analogies at a higher level of abstraction, such as detailed analogies of scientific concepts, or coherent story structures that preserve relational mappings across larger contexts (Mittal, 1992; Zhu and Ontanón, 2010).

Of the 29 works concerning analogy generation, 18 are at the lexical level (14 of which use LLMs) and 12 at the compositional level, with one paper (Wang et al., 2024) addressing both (see Tab. 1). The compositional set spans story generation, scientific-concept analogies, historical analogies, educational explanations (Bhavya et al., 2024b; Mittal, 1992; Li et al., 2005, 2025; Chen et al., 2024; Bernstein et al., 2024; Yuan et al., 2024b; Zhu and Ontanón, 2010), and a small multimodal subset that operates on image pairs (Davies et al., 2008; Yaner and Goel, 2006; Sadeghi et al., 2015), alongside evaluation benchmarks (Yiu et al., 2025; Jiang et al., 2024; Yilmaz et al., 2025); we treat the multimodal works as compositional since they assess relational structure across pair-wise inputs.

LLM-based Generation We identify 18 papers that use LLMs to generate analogies through prompting techniques (see Tab. 2). Among these, 11 use multi-step prompting pipelines to improve generation quality, while seven earlier papers rely on single-step prompting. This pattern suggests a shift from simple prompting toward more elaborate prompt designs, enabled by increasingly capable LLMs and the growing availability of LLM-focused tooling. From the perspective of analogical reasoning, this shift also aligns with the multi-stage cognitive account of analogy—retrieval, mapping, and transfer—suggesting that decomposing prompts into discrete steps may better guide LLMs through these stages and produce more coherent, relationally accurate analogies (Gentner and Forbus, 2011).

From the corpus, 10 of the 18 LLM-generation papers (56%) explicitly incorporate a human-in-the-loop curation step rather than relying solely on autonomous LLM output. These interventions take several forms, including manual annotation and filtering (Sultan et al., 2024; Bhavya et al., 2023; Bernstein et al., 2024), injection of user-related information (Cao et al., 2024; Ju et al., 2025; Chen et al., 2024), and prompt tuning (Wang et al., 2024; Shao et al., 2025). Across these strategies, human judgment functions not as a post-hoc addition but as an integral stage of the generation pipeline. Several papers report quality issues in initial LLM outputs without such intervention (Bhavya et al., 2022; Ding et al., 2023; Cao et al., 2024; Shao et al., 2025), suggesting that fully autonomous LLM analogy generation has not yet been shown to produce consistently usable outputs. In practice, the strongest demonstrated paradigm is human curation of LLM-generated candidates, though its limits and trade-offs remain underexplored.

4.3 Analogy Evaluation Methods

Evaluation by Granularity We identify the evaluation methods using the same granularity (lexical-level and compositional-level) and differentiate the human evaluation approach from the automatic one. Across 51 automatic evaluation approaches, 33 focus on the lexical level and 18 on the compositional level (See Tab. 4). On the human evaluation side, three papers use human evaluation of lexical-level metrics and data, while 24 papers use experts or crowdsourced participants to evaluate compositional-level metrics, such as the overall quality of the generated analogy (Sultan et al., 2024; Jiayang et al., 2023). 13 papers include both automatic and human evaluation approaches. One paper does not include a formal evaluation using metrics as it proposes a theoretical model (Salu, 1994). This result reveals a structural split in the field's evaluation methodology: automatic metrics are concentrated where analogies are formed simplest (lexical-level), while human evaluation is concentrated where analogies carry the most relational

meaning (compositional-level). With only 13 of 65 papers measuring at both levels, the field lacks a universal way to detect when a method scoring well on lexical-automatic benchmarks fails to produce relationally deep compositional analogies.

Quality Dimension We also categorized the reported evaluation metrics into five dimensions: accuracy, validity, similarity, novelty, and human preference/judgment (See Tab. 5). Many automatic evaluation methods target one or more of these dimensions. Specifically, we identified 28 papers that use accuracy-based metrics, such as precision, recall, and F1-score; 20 papers that assess validity, including logical consistency checks and human validity judgments; and 20 papers that apply similarity-based metrics, such as BLEU, to evaluate generated analogies. Most recent LLM benchmark reports accuracy as the primary metric, and we will continue the discussion of whether this metric is well-suited to compositional, novel, or relationally deep analogies in §4.5.

Novelty is an often mentioned metric in the work we identified (Bhavya et al., 2023; Jiayang et al., 2023), and while some work evaluates novelty through measuring word distance from existing analogy (Bhavya et al., 2023), some directly use crowdsourcers' judgment (Jiayang et al., 2023). Human preference/judgment is used in 19 papers as a primary evaluation method. These works typically employ crowdsourcers (Jiayang et al., 2023; Sultan et al., 2024) or experts (Cao et al., 2024; Shao et al., 2025) to evaluate or validate the generated analogy in the context of their work.

4.4 Analogy Generation Challenge

Novel Analogy Generation Generating novel analogies remains one of the field's most under-evaluated and unsolved problems. Our corpus contains only four papers (out of 65) that include a dedicated novelty metric (Tab. 5), making novelty a niche evaluation dimension. The under-evaluation reflects an underlying difficulty rather than disinterest. Many canonical analogies, such as the comparison between the solar system and Rutherford's atomic model (Gentner, 1983), were originally crafted by humans and later incorporated into computational models; in practice, both symbolic and neural systems frequently reuse well-established conceptual mappings, resulting in limited novelty and diversity (Bhavya et al., 2022). One study notes that LLMs tend to produce "mostly known analogies that are explicitly mentioned on the Web" (Bhavya et al., 2023), and it remains unclear how to elicit truly creative new analogies from them. Furthermore, novel analogies often require structural or underlying similarity with little to no literal similarity, which makes them hard to generate and capture using corpus-based or embedding-only models (Yuan et al., 2023)—a tension that is unresolved for LLMs, whose generative capacity is itself bounded by their training corpus.

Generation with LLM A key challenge in LLM-based analogy generation is the model's limited ability to capture deep relational similarity consistently. Multiple reports (Jiayang et al., 2023; Yuan et al., 2023, 2024a; Chen et al., 2022) find LLMs often conflate literal similarity with actual analogical structure, frequently generating analogies that are either repetitive or shallow. This misalignment undermines the goal of analogy generation, which centers on abstract, structural mapping.

Moreover, generation quality is susceptible to the choice of LLM and the prompt. Even carefully crafted prompts can yield outputs that are misleading, incorrect, or overly simplistic (Bhavya et al., 2023). Even with detail-designed prompts, LLM can still generate analogies perceived as "oversimplified and lacking depth" (Cao et al., 2024). This issue extends to analogy applications, where the generated analogies often fail to match users' prior knowledge and are sometimes considered superficial or incomplete (Cao et al., 2024; Shao et al., 2025).

4.5 Analogy Evaluation Challenge

Metrics and Dataset Limitation Limitations in evaluation metrics and available datasets have been widely documented in the literature (Chen et al., 2022; Li et al., 2023; Turney et al., 2006). These issues affect not only traditional systems (e.g., symbolic, logic-based, and retrieval methods) but also recent LLM-driven approaches.

A central concern is the narrow scope of existing benchmarks. Commonly used datasets, such as the SAT word analogy dataset (Turney, 2008) and the Google analogy corpus (Mikolov et al., 2013), are relatively small and restricted to lexical-level analogies. This narrow coverage limits the range of analogical phenomena on which models can be evaluated and hinders cross-domain, multilingual, or multimodal marking.

In evaluation methodology, many studies rely on

binary classification accuracy or multiple-choice formats to assess model performance, especially in SAT-style tasks. While straightforward, these metrics fail to capture graded similarity, analogical strength, and explanatory coherence (Bollegala et al., 2009). Studies also use standard NLP metrics, such as BLEU, ROUGE, and BERTScore, to evaluate analogical quality. However, these metrics primarily assess surface-level textual overlap or vector-based semantic similarity, rather than relational alignment or structural correctness (Chen et al., 2022). As a result, there is a risk of overemphasizing surface similarity, encouraging models to generate trivial or formulaic analogies at the expense of deeper, more creative mappings (Bhavya et al., 2024a).

Evaluation with LLM Because automatic metrics dominate at the lexical level and human metrics dominate at the compositional level, a method that scores well on lexical-automatic benchmarks (e.g., SAT analogy accuracy) provides little information about whether it generates relationally deep compositional analogies. The decoupling poses two specific challenges.

First, LLMs sometimes generate outputs that are literally similar but relationally shallow, which can mislead both human and automatic evaluations (Yuan et al., 2023; Jiayang et al., 2023). Without a metric that detects divergence between surface and structural similarity, the field cannot reliably tell when a model is fitting the eval. Second, existing automatic metrics, especially those designed for word analogy tasks, struggle to evaluate analogies beyond the lexical or syntactic level (Yuan et al., 2023). Few automatic metrics account for creativity, novelty, or contextual coherence, all of which are central to human analogical reasoning (Hofstadter and Sander, 2013). This indicates the need to develop compositional-level automatic metrics or new benchmarks that correlate meaningfully with human judgments.

Do LLMs Reason Analogically? A central question runs through the identified literature: do LLMs perform genuine relational reasoning, or do they exploit training-distribution cues that surface as analogy-like outputs? The question is hard to settle because LLM evaluation is itself susceptible to confounds that mimic capability differences: LLMs succeed disproportionately on probable inputs and outputs at nominally identical tasks (McCoy et al., 2024), and emergent capabilities depend substantially on metric choice (Schaeffer et al., 2023; Hu and Frank, 2024). Accuracy alone cannot easily distinguish reasoning from confounded performance.

The corpus's debate illustrates this directly. Webb et al. (2023) report that GPT-3 matches or exceeds humans on letter-string, digit-matrix, four-term verbal, and story analogies; however, Hodel and West (2024) shows fragility on simple variants where humans remain consistent, and Lewis and Mitchell (2024) formalizes the critique into a counterfactual-tasks paradigm: GPT performance collapses on permuted- and symbol-alphabet variants even when comprehension checks pass. Webb et al. (2025) reply that GPT-4 closes most of this gap when given code execution for alphabet indexing, arguing the critique conflates analogy with auxiliary task demands. Lewis and Mitchell (2025) extends the paradigm to three of Webb et al. (2023)'s four domains and documents brittleness to paraphrase, answer order, and counterfactual variants that controls do not eliminate.

In response, recent work has begun to localize the question inside the model. Yang et al. (2025) identify in Llama-3-70B (replicated in Gemma-2 and Qwen-2.5) a three-stage emergent circuit—symbol-abstraction, symbolic-induction, and retrieval heads—that implements abstract-rule reasoning over letter-string and verbal-analogy tasks. Hofmann et al. (2025) shows that an exemplar-based cognitive model captures GPT-J's English derivational morphology better than a rule-based alternative, indicating analogical generalization at the mechanistic level. Opiełka et al. (2025) localize invariant linear concept vectors for verbal antonyms and categories in Llama-3.1-70B but find no stable representation for ordinal "previous"/"next"—which are also the relations on which counterfactual perturbations elicit the largest accuracy declines (Lewis and Mitchell, 2024, 2025).

The combined record does not converge on a single answer, but it does sharpen the question. Accuracy patterns admit competing causal accounts—analogy-specific brittleness or an auxiliary-task-demand bottleneck—that current protocols do not separate. The mechanistic results, when obtained, indicate a partition rather than a binary: LLMs form invariant relational representations for some concept families and rely on surface similarity for others, with the partition reflecting the prevalence of each relation in pretraining. Claims about LLMs'

capability for analogy cannot be supported solely on accuracy; we examine the consequences for evaluation practice in §5.

5 Discussions

Challenges and Opportunities in Analogy Generation A promising direction is the development of analogy-specific prompting strategies that mirror the cognitive stages of analogical reasoning—retrieval, mapping, and transfer (Gentner and Forbus, 2011). Multi-step prompting (Wei et al., 2023) has shown early success, but lacks a framework tailored to analogy tasks. Developing such a framework could also reduce reliance on human-in-the-loop curation, which 10 of 18 LLM-generation papers in our corpus explicitly require, narrowing the gap between iterative generation with curation and autonomous analogy generation.

Generating novel analogies remains the field's most under-evaluated and unsolved problem (Bhavya et al., 2022, 2024a). The deeper challenge lies in aligning computational definitions of novelty, typically dissimilarity to known examples or training corpora, with human judgments grounded in prior knowledge, domain experience, and perceived insight. Bridging this gap will likely require approaches that combine retrieval-augmented generation, knowledge-aware prompting, and structure-constrained decoding to guide models beyond surface-level similarity.

Future Directions in Analogy Evaluation A notable shift in the recent literature is that capability questions about analogy in LLMs are increasingly seen as not fully answerable at the level of accuracy alone (§4.5). Task-performance numbers have not resolved whether LLMs perform genuine relational reasoning. We accordingly recommend that capability claims be reported as performance distributions across paraphrase, alphabet, content, metric, and answer-order variants (Yang et al., 2025; Opiełka et al., 2025; Lee et al., 2026).

The split between lexical-automatic and compositional-human evaluation (§4.3) makes it difficult to detect when methods scoring well at the lexical level fail to produce relationally-deep compositional analogies. Closing this gap requires domain-diverse datasets that pair granularities with evaluation protocols that combine automatic metrics with human assessments of analogical relevance, novelty, and coherence (Bhavya et al., 2024a). Analogy evaluation should move beyond lexical correctness to capture the relational structure that defines analogical reasoning.

6 Conclusion

In this paper, we present a systematic literature review on analogy generation and evaluation, following the PRISMA framework. After two rounds of search and screening, 65 primary research papers spanning four decades of work were included in the final analysis. Our synthesis surfaces a structural split between lexical-automatic and compositional-human evaluation, a corpus-wide reliance on human-in-the-loop curation in LLM-based generation, and an open question on whether benchmark performance reflects genuine analogical reasoning. These findings motivate concrete standards for future work: capability claims should be supported by results across multiple variations of each task rather than a single benchmark score and, where possible, accompanied by analyses of how the model reaches its answer; evaluation protocols should span both the lexical and compositional levels, so that strong performance at one does not mask weakness at the other.

Limitations

As a review process that involves subjective judgments by individual authors, this study may be subject to bias. To mitigate this, we conducted pilot checks at each stage to iteratively refine the selection criteria. Additionally, all results were cross-checked by multiple authors, and meetings were held to resolve conflicts and ensure consistency. Despite careful formulation of the search strings, some relevant papers may have been missed. In particular, while this review focused primarily on analogy, there is conceptual overlap with metaphor, and relevant discussions of analogy may appear in metaphor-focused papers. Future work could undertake a more exhaustive investigation into the distinctions and interrelations between analogy and metaphor. Our subsequent update of the corpus relied on forward and backward snowballing rather than re-running the full keyword search, so new works that are not reachable by citation from the original corpus may be under-represented.

References

Paul Bartha. 2024. Analogy and Analogical Reasoning. In Edward N. Zalta and Uri Nodelman, editors, *The*

Stanford Encyclopedia of Philosophy, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Seth Bernstein, Paul Denny, Juho Leinonen, Lawrence Kan, Arto Hellas, Matt Littlefield, Sami Sarsa, and Stephen MacNeil. 2024. "like a nesting doll": Analyzing recursion analogies generated by CS students using large language models. In *Proceedings of the 2024 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE)*, pages 122–128.

Bhavya Bhavya, Chris Palaguachi, Yang Zhou, Suma Bhat, and ChengXiang Zhai. 2024a. Long-form analogy evaluation challenge. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 1–16.

Bhavya Bhavya, Shradha Sehgal, Jinjun Xiong, and ChengXiang Zhai. 2024b. Anade1. 0: A novel data set for benchmarking analogy detection and extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1723–1737.

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 (INLG)*, pages 298–312. Association for Computational Linguistics.

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2023. CAM: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, pages 3903–3914. Association for Computing Machinery.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2025. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 6692–6704. Association for Computational Linguistics.

Joanne Boisson, Asahi Ushio, Hsuvas Borkakoty, Kiamehr Rezaee, Dimosthenis Antypas, Zara Siddique, Nina White, and Jose Camacho-Collados. 2024. How are metaphors processed by language models? the case of analogies. In *Proceedings of the 28th Conference on Computational Natural Language Learning (CoNLL)*, pages 365–387.

Danushka Bollegala. 2010. A supervised ranking approach for detecting relationally similar word pairs. In *2010 Fifth International Conference on Information and Automation for Sustainability*, pages 323–328. IEEE.

Danushka T Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2009. Measuring the similarity between implicit semantic relations from the web. In *Proceedings of the 18th international conference on World wide web*, pages 651–660.

L Bourrelly, E Chouraqui, and M Ricard. 1983. Formalisation of an approximate reasoning: The analogical reasoning. *IFAC Proceedings Volumes*, 16(13):135–141.

Cassie Chen Cao, FANG Zoe, Y CAO Lydia, LIN Jionghao, and LI Ruizhe. 2024. Llm-generated personalized analogies to foster ai literacy in adult novices. In *International Conference on Computers in Education*.

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*, pages 3941–3955. Association for Computational Linguistics.

Qing Chen, Wei Shuai, Jiyao Zhang, Zhida Sun, and Nan Cao. 2024. Beyond numbers: Creating analogies to enhance data comprehension and communication with generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Maxwell Crouse, Constantine Nakos, Ibrahim Abdelaziz, and Ken Forbus. 2021. Neural analogical matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 809–817.

Rikke Friis Dam and Teo Yu Siang. 2022. Affinity diagrams: How to cluster your ideas and reveal insights. *Interaction Design Foundation—IxDF*.

Jim Davies, Ashok K Goel, and Patrick W Yaner. 2008. Proteus: Visuospatial analogy in problem-solving. *Knowledge-Based Systems*, 21(7):636–654.

Ronald Denaux and Jose Manuel Gomez-Perez. 2019. Assessing the lexico-semantic relational knowledge captured by word and concept embeddings. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 29–36.

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

Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63.

Robert M French. 2002. The computational modeling of analogy-making. *Trends in cognitive Sciences*, 6(5):200–205.

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

Dedre Gentner and Kenneth D Forbus. 2011. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3):266–276.

Dedre Gentner and Linsey A Smith. 2013. Analogical learning and reasoning.

Ikut Tri Handoyo and Dana Indra Sensuse. 2017. Knowledge-based systems in decision support context: A literature review. In *2017 4th International Conference on New Media Studies (CONMEDIA)*, pages 81–86.

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

Valentin Hofmann, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

Douglas R Hofstadter. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

Douglas R Hofstadter and Melanie Mitchell. 1994. The copycat project: A model of mental fluidity and analogy-making.

Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books.

Keith J Holyoak and Paul Thagard. 1989. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355.

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*, pages 235–243.

Jennifer Hu and Michael C. Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *Conference on Language Modeling (COLM)*. Outstanding Paper.

William Hudson. 2013. The encyclopedia of human-computer interaction, chapter card sorting. *The Interaction Design Foundation*, 2.

John E Hummel and Keith J Holyoak. 2019. Lisa: A computational model of analogical inference and schema induction. In *Proceedings of the eighteenth annual conference of the cognitive science society*, pages 352–357. Routledge.

Samireh Jalali and Claes Wohlin. 2012. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 29–38.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024. MARVEL: Multidimensional abstraction and reasoning through visual evaluation and learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS) Datasets and Benchmarks Track*.

Cheng Jiayang, Lin Qiu, Tsz Ho 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 (EMNLP)*, pages 11518–11537. Association for Computational Linguistics.

Hyojin Ju, Jungeun Lee, Seungwon Yang, Jungseul Ok, and Inseok Hwang. 2025. Toward affective empathy via personalized analogy generation: A case study on microaggression. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–31.

Taeyoon Lee, Minju Song, Chaewon Yoon, Jaewoo Park, and Jaewoo Kang. 2026. The curious case of analogies: Investigating analogical reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Forthcoming.

Yves Lepage and Shinichi Ando. 1996. Saussurian analogy: a theoretical account and its application. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society (CogSci)*.

Martha Lewis and Melanie Mitchell. 2025. Evaluating the robustness of analogical reasoning in large language models. *Transactions on Machine Learning Research*.

John Li, Deborah Nichols, and Allan Terry. 2005. Analogy, deduction and learning. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 294a–294a. IEEE.

Nianqi Li, Siyu Yuan, Jiangjie Chen, Jiaqing Liang, Feng Wei, Zujie Liang, Deqing Yang, and Yanghua Xiao. 2025. Past meets present: Creating historical analogy with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 3942–3957.

Shuyi Li, Shaojuan Wu, Xiaowang Zhang, and Zhiyong Feng. 2023. An analogical reasoning method based on multi-task learning with relational clustering. In *Companion Proceedings of the ACM Web Conference 2023*, pages 144–147.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Vibhu O Mittal. 1992. Generating analogical natural language object descriptions. In *Proceedings of the 30th annual ACM Southeast Regional Conference*, pages 239–246.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. 2010. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, 8(5):336–341.

Sam Musker, Aleksandra Duchnowski, Raphaël Millière, and Ellie Pavlick. 2025. LLMs as models for analogical reasoning. *Journal of Memory and Language*, 145:104676.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. 2025. Analogical reasoning inside large language models: Concept vectors and the limits of abstraction. ArXiv preprint.

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

Tony A Plate. 2000. Analogy retrieval and processing with distributed vector representations. *Expert systems*, 17(1):29–40.

Chengwei Qin, Wenhan Xia, Tan Wang, Fangkai Jiao, Yuchen Hu, Bosheng Ding, Ruirui Chen, and Shafiq Joty. 2025. Relevant or random: Can LLMs truly perform analogical reasoning? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23993–24010.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Fereshteh Sadeghi, C Lawrence Zitnick, and Ali Farhadi. 2015. Visalogy: Answering visual analogy questions. *Advances in Neural Information Processing Systems*, 28.

Yehuda Salu. 1994. A neural network for analogical reasoning. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 7, pages 4772–4777. IEEE.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36 (NeurIPS)*. Outstanding Paper.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 2 (Short Papers)*, pages 242–246. Association for Computational Linguistics.

Zekai Shao, Siyu Yuan, Lin Gao, Yixuan He, Deqing Yang, and Siming Chen. 2025. Unlocking scientific concepts: How effective are LLM-generated analogies for student understanding and classroom practice? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery.

Donghyeon Shin, Seungpil Lee, K. L. Kovacec, and Sundong Kim. 2024. From generation to selection: Findings of converting analogical problem-solving into multiple-choice questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6696–6708.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Donna Spencer. 2009. *Card sorting: Designing usable categories*. Rosenfeld Media.

Claire E Stevenson, Alexandra Pafford, Han LJ van der Maas, and Melanie Mitchell. 2026. Can large language models generalize analogy solving like children can? *Transactions of the Association for Computational Linguistics*, 14:612–626.

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.

Paul Thagard, Keith J Holyoak, Greg Nelson, and David Gochfeld. 1990. Analog retrieval by constraint satisfaction. *Artificial intelligence*, 46(3):259–310.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2010. Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research*, 37:1–39.

Peter D Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Peter D Turney et al. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Junjie Wang, Dan Yang, Binbin Hu, Yue Shen, Wen Zhang, and Jinjie Gu. 2024. Know your needs better: Towards structured understanding of marketer demands with analogical reasoning augmented llms. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5860–5871.

Wei Wang, Qinghua Zheng, and Yingying Chen. 2009. Knowledge element analogy relation recognition using text and graph structure. In *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–8. IEEE.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2025. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS Nexus*, 4(5):pgaf135.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL – a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549. Association for Computational Linguistics.

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.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Sreeram Vennam, Vinija Jain, Aman Chadha, Amitava Das, Ponnurangam Kumaraguru, and Amit Sheth. 2025. KnowledgePrompts: Exploring the abilities of large language models to solve proportional analogies via knowledge-enhanced prompting. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 3979–3996.

Patrick H Winston. 1980. Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703.

Patrick W Yaner and Ashok K Goel. 2006. Visual analogy: Viewing analogical retrieval and mapping as constraint satisfaction problems. *Applied Intelligence*, 25:91–105.

Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. 2025. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 70515–70549.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13060–13082.

Nilay Yilmaz, Maitreya Patel, Yiran Lawrence Luo, Tejas Gokhale, Chitta Baral, Suren Jayasuriya, and Yezhou Yang. 2025. VOILA: Evaluation of MLLMs for perceptual understanding and analogical reasoning. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. 2025. KiVA: Kid-inspired visual analogies for testing large multimodal models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

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. Association for Computational Linguistics.

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.

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 (EMNLP)*, pages 6026–6036.

Shutong Zhang, Tianyu Zhang, Jinghui Cheng, and Shurui Zhou. 2025. Who is to blame: A comprehensive review of challenges and opportunities in designer-developer collaboration. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Yating Zhang, Adam Jatowt, and Katsumi Tanaka. 2017. Temporal analog retrieval using transformation over dual hierarchical structures. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 717–726.

Jichen Zhu and Santiago Ontanón. 2010. Story representation in analogy-based story generation in riu. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, pages 435–442. IEEE.

A Disclosure of AI Assistant Use

We used AI-based writing assistants (e.g., Grammarly and ChatGPT) solely for grammar checking, phrasing refinement, and language polishing. All conceptual contributions, analyses, and writing content were original and authored by the researchers.

B Complete Result

Generation Granularity	Papers
Lexical-level	(Sultan et al., 2024), (Yuan et al., 2024a), (Bourrelly et al., 1983), (Ding et al., 2023), (Bhavya et al., 2023), (Wang et al., 2024), (Chen et al., 2022), (Jiayang et al., 2023), (Bhavya et al., 2022), (Salu, 1994), (Yuan et al., 2023), (Crouse et al., 2021), (Shao et al., 2025), (Ju et al., 2025), (Boisson et al., 2025), (Wijesiriwardene et al., 2025), (Shin et al., 2024), (Yasunaga et al., 2024)
Compositional-level	(Bhavya et al., 2024b), (Mittal, 1992), (Li et al., 2005), (Wang et al., 2024), (Li et al., 2025), (Chen et al., 2024), (Bernstein et al., 2024), (Yuan et al., 2024b), (Zhu and Ontanón, 2010), (Davies et al., 2008), (Yaner and Goel, 2006), (Sadeghi et al., 2015)

Table 1: Analogy Generation by Granularity

Type	Dimension	Papers
LLM	**Human-in-the-loop**	(Shao et al., 2025), (Ju et al., 2025), (Boisson et al., 2025), (Sultan et al., 2024), (Wang et al., 2024), (Yuan et al., 2024a), (Jiayang et al., 2023), (Yuan et al., 2023), (Chen et al., 2024), (Bernstein et al., 2024)
	Multi-step Generation	(Shao et al., 2025), (Ju et al., 2025), (Boisson et al., 2025), (Sultan et al., 2024), (Yuan et al., 2024a), (Jiayang et al., 2023), (Bhavya et al., 2023), (Li et al., 2025), (Chen et al., 2024), (Bernstein et al., 2024), (Yuan et al., 2024b)
	Single Prompt Generation	(Bhavya et al., 2022), (Ding et al., 2023), (Wang et al., 2024), (Yuan et al., 2023), (Wijesiriwardene et al., 2025), (Shin et al., 2024), (Yasunaga et al., 2024)
Non-LLM		(Salu, 1994), (Crouse et al., 2021), (Chen et al., 2022), (Bhavya et al., 2024b), (Bourrelly et al., 1983), (Mittal, 1992), (Davies et al., 2008), (Yaner and Goel, 2006), (Li et al., 2005), (Zhu and Ontanón, 2010), (Sadeghi et al., 2015)

Table 2: LLM vs. Non-LLM Generation

Model Type	Papers
Relational Graph-based	(Denaux and Gomez-Perez, 2019), (Falkenhainer et al., 1989), (Winston, 1980), (Tsatsaronis et al., 2010), (Bourrelly et al., 1983), (Holyoak and Thagard, 1989)
Distributional Semantic	(Turney et al., 2006), (Turney, 2008), (Li et al., 2023), (Wang et al., 2009), (Hope et al., 2017), (Bollegala, 2010), (Schluter, 2018)
Cognitive/architectural	(Hofstadter and Mitchell, 1994), (Hummel and Holyoak, 2019)
Transformation-based	(Plate, 2000), (Lepage and Ando, 1996)

Table 3: Evaluation Model Types

Type	Level	Papers
Automatic	**Lexical-level**	(Schluter, 2018), (Tsatsaronis et al., 2010), (Sadeghi et al., 2015), (Denaux and Gomez-Perez, 2019), (Wang et al., 2024), (Chen et al., 2022), (Plate, 2000), (Li et al., 2005), (Jiayang et al., 2023), (Turney, 2008), (Winston, 1980), (Yuan et al., 2023), (Bhavya et al., 2022), (Li et al., 2023), (Yuan et al., 2024a), (Wijesiriwardene et al., 2023), (Bollegala, 2010), (Hope et al., 2017), (Sultan et al., 2024), (Bhavya et al., 2024b), (Thagard et al., 1990), (Bollegala et al., 2009), (Lepage and Ando, 1996), (Hofstadter and Mitchell, 1994), (Wijesiriwardene et al., 2025), (Shin et al., 2024), (Yasunaga et al., 2024), (Lewis and Mitchell, 2024), (Webb et al., 2025), (Lewis and Mitchell, 2025), (Musker et al., 2025), (Qin et al., 2025), (Lee et al., 2026)
	Compositional-level	(Wang et al., 2009), (Zhang et al., 2017), (Falkenhainer et al., 1989), (Crouse et al., 2021), (Yuan et al., 2023), (Zhu and Ontanón, 2010), (Yaner and Goel, 2006), (Winston, 1980), (Plate, 2000), (Hummel and Holyoak, 2019), (Wijesiriwardene et al., 2024), (Boisson et al., 2024), (Sourati et al., 2024), (Ye et al., 2024), (Yuan et al., 2024b), (Yiu et al., 2025), (Jiang et al., 2024), (Yilmaz et al., 2025)
Human	**Lexical-level**	(Yuan et al., 2023), (Turney, 2008), (Stevenson et al., 2026)
	Compositional-level	(Falkenhainer et al., 1989), (Li et al., 2005), (Wang et al., 2024), (Plate, 2000), (Bourrelly et al., 1983), (Boisson et al., 2025), (Cao et al., 2024), (Mittal, 1992), (Shao et al., 2025), (Ju et al., 2025), (Sultan et al., 2024), (Jiayang et al., 2023), (Yuan et al., 2024a), (Davies et al., 2008), (Ding et al., 2023), (Bhavya et al., 2024a), (Bhavya et al., 2023), (Hope et al., 2017), (Bhavya et al., 2022), (Li et al., 2025), (Chen et al., 2024), (Bernstein et al., 2024), (Yuan et al., 2024b), (Sourati et al., 2024)

Table 4: Analogy Evaluation by Granularity. Multimodal benchmarks (KiVA, MARVEL, VOILA) are categorized as compositional-level since they assess relational structure across image pairs.

Evaluation Granularity	Papers
Accuracy	(Crouse et al., 2021), (Turney, 2008), (Chen et al., 2022), (Yuan et al., 2024a), (Bhavya et al., 2024b), (Hope et al., 2017), (Yuan et al., 2023), (Yaner and Goel, 2006), (Bollegala, 2010), (Bhavya et al., 2024a), (Plate, 2000), (Wijesiriwardene et al., 2024), (Wijesiriwardene et al., 2025), (Shin et al., 2024), (Yasunaga et al., 2024), (Boisson et al., 2024), (Sourati et al., 2024), (Ye et al., 2024), (Lewis and Mitchell, 2024), (Webb et al., 2025), (Lewis and Mitchell, 2025), (Stevenson et al., 2026), (Qin et al., 2025), (Lee et al., 2026), (Yiu et al., 2025), (Jiang et al., 2024), (Yilmaz et al., 2025), (Yuan et al., 2024b),
Similarity	(Bollegala et al., 2009), (Sultan et al., 2024), (Thagard et al., 1990), (Li et al., 2023), (Wang et al., 2024), (Lepage and Ando, 1996), (Plate, 2000), (Wang et al., 2009), (Schluter, 2018), (Jiayang et al., 2023), (Yuan et al., 2023), (Sadeghi et al., 2015), (Tsatsaronis et al., 2010), (Denaux and Gomez-Perez, 2019), (Winston, 1980), (Wijesiriwardene et al., 2023), (Zhu and Ontanón, 2010), (Bhavya et al., 2022), (Boisson et al., 2025), (Ye et al., 2024),
Validity	(Crouse et al., 2021), (Jiayang et al., 2023), (Hope et al., 2017), (Li et al., 2005), (Zhang et al., 2017), (Yuan et al., 2024a), (Winston, 1980), (Bhavya et al., 2024a), (Hofstadter and Mitchell, 1994), (Holyoak and Thagard, 1989), (Hummel and Holyoak, 2019), (Bourrelly et al., 1983), (Mittal, 1992), (Lewis and Mitchell, 2024), (Webb et al., 2025), (Lewis and Mitchell, 2025), (Musker et al., 2025), (Qin et al., 2025), (Lee et al., 2026), (Yuan et al., 2024b),
Novelty	(Bhavya et al., 2022), (Jiayang et al., 2023), (Hope et al., 2017), (Bhavya et al., 2023),
Human Preference /Judgement	(Sultan et al., 2024), (Yuan et al., 2024a), (Davies et al., 2008), (Ding et al., 2023), (Bhavya et al., 2022), (Turney, 2008), (Falkenhainer et al., 1989), (Li et al., 2005), (Wang et al., 2024), (Shao et al., 2025), (Cao et al., 2024), (Ju et al., 2025), (Boisson et al., 2025), (Li et al., 2025), (Chen et al., 2024), (Bernstein et al., 2024), (Yuan et al., 2024b), (Sourati et al., 2024), (Stevenson et al., 2026),

Table 5: Analogy Evaluation Dimension