

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

Shutong Wu* Tianyu Zhang* Hecong Wang Yang Lu Jiankun Yang Jiamu Tang
Zhen Bai

University of Rochester

Abstract

Analogy, a quintessential human cognitive capability, has long been studied for its role in transferring knowledge across domains, from generating novel analogies to evaluating their quality. The field of artificial intelligence (AI) has long sought to model the analogical reasoning process computationally, from using logical representations to adopting connectionist methods. However, the rapidly improving capabilities of large language models (LLMs) have led to the creation of new families of LLM-powered analogy generation systems, creating a need for a comprehensive review that situates these developments within the broader historical context. Following the PRISMA framework, we systematically reviewed computational analogy research across computer science, AI, and NLP, yielding a final corpus of 83 papers spanning four decades. Our synthesis surfaces three observations: automatic and human evaluation are split by granularity, with automatic metrics concentrated at the lexical level and human evaluation at the compositional level; LLM analogy generation in practice rarely operates without a human-in-the-loop curation step; and the question of whether LLMs perform genuine relational reasoning remains contested when assessed by accuracy alone, with recent literature beginning to provide a complementary perspective. We outline concrete standards for future work, including evaluation across multiple task variations and protocols that span both lexical and compositional levels.

1 Introduction

Analogy plays a fundamental role in human learning, enabling individuals to comprehend unfamiliar concepts by drawing parallels with familiar ones (Gentner and Smith, 2013; Bartha, 2024). Computational approaches to analogical reasoning, which

identify and map relational correspondences between a known *source* and a novel *target*, have attracted sustained interest across artificial intelligence (AI) and natural language processing (NLP).

Within this source–target mapping framing, the literature addresses three distinct tasks. *Analogy retrieval*: given a target concept, problem, or situation, finding an existing source analog from memory or a corpus, e.g., case retrieval or nearest-neighbor search over structured or embedding representations (Forbus et al., 1995). *Analogy generation*: producing a novel analogy—at the lexical level (e.g., $A:B::C:D$ four-term completions and word pairs) or the compositional level (i.e., extended structured comparisons between situations, processes, or narratives)—rather than selecting one that already exists. *Analogy evaluation*: judging the quality of a given analogy, including the validity of its relational mapping, aptness, novelty, and usefulness, whether by automatic metrics, learned models, or human judgment. All three differ from *analogical reasoning* which uses or completes a given analogy, as in four-term benchmark tasks, rather than producing a new one.

Earlier symbolic methods such as Winston’s frame-based system (Winston, 1980) and the Structure-Mapping Engine (Falkenhainer et al., 1989) represented analogies as logical forms and evaluated them by relational alignment over hand-crafted examples. Subsequent neural approaches enabled models to learn analogical relationships from data, evaluated on word- and sentence-analogy datasets (Turney, 2008; Mikolov et al., 2013). More recently, pretrained language models like GPT (Ouyang et al., 2022) extended analogy generation beyond the word level to longer compositional forms, supporting applications from analogy mining (Bhavya et al., 2023) to personalized educational use (Cao et al., 2024).

The evaluation of analogies remains a central challenge. Structure Mapping Theory (Gentner,

*Equal contribution.

1983) formalized analogy as relational alignment, and relational and word similarity have since been dominant automatic metrics (Turney et al., 2006), complemented by human evaluation via expert judgment or crowdsourcing. LLM-based hybrid evaluation strategies have also recently emerged (Bhavya et al., 2024a).

Prior surveys have reviewed symbolic and neural analogy methods (Mitchell, 2021; Gentner and Forbus, 2011; French, 2002), but none comprehensively cover the rise of LLM-based approaches. Recent work (Bhavya et al., 2022; Yuan et al., 2024a) remains fragmented across isolated tasks such as prompt engineering, dataset creation, or specific applications. Rather than an LLM-only review, this work is a four-decade synthesis: it situates LLM-era techniques within the longer arc of computational analogy research and surfaces foundational challenges that LLMs have reframed.

In this paper, we present a systematic literature review (SLR) of computational approaches to analogy generation and evaluation in the domain of computer science (CS), AI, and NLP, conducted in accordance with the PRISMA guidelines (Moher et al., 2010). The literature review will address the following research questions:

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

We systematically searched and screened 5,841 records across six databases, resulting in a final corpus of 83 papers spanning four decades of computational analogy research.

Our contributions are: (1) a two-round PRISMA-aligned review with 66 primary papers; (2) three findings cutting across the corpus, including a lexical-automatic versus compositional-human evaluation split, a corpus-wide reliance on human-in-the-loop curation in LLM-based generation, and an open question on whether benchmark performance reflects genuine relational reasoning; and (3) concrete standards for future work.

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 cast analogy as *completion*—filling the missing term of $A:B::C:?$ or aligning a given source–target pair—tasks with a gold answer against which outputs could be checked. Open-ended analogy *generation* lacks a meaningful gold answer, thereby creating the evaluation problem at the center of this review.

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). In Newell and Simon’s generate-and-test terms (Newell and Simon, 1972), this balance has inverted: classical systems paired weak generators with strong structural tests, whereas LLM-era pipelines couple a prolific but unreliable generator to a still-underdeveloped test.

2.2 SLRs in Computational Analogy Model

Prior surveys have framed computational analogy models in complementary ways. French (2002) provides a historical overview, classifying models into symbolic, connectionist, and hybrid paradigms and surveying open problems, whereas Gentner and Forbus (2011) decomposes analogy into the subprocesses of retrieval, mapping, abstraction, and re-representation, emphasizing that analogical mappings favor systematic, higher-order relational correspondences. Both reviews, along with foundational works (Gentner, 1983; Hofstadter, 1995), ground analogical inference in relational representations.

Mitchell (2021) brings a more recent perspective, noting that today’s systems “are almost entirely

lacking the ability” to form humanlike abstractions or analogies. These reviews (French, 2002; Gentner and Forbus, 2011; Mitchell, 2021), however, predate the generative turn: Bhavya et al. (2022) demonstrates that InstructGPT (Ouyang et al., 2022) can be prompted to produce near-human-quality conceptual analogies, yet judging the creativity and validity of such open-ended outputs—beyond lexical pattern matching—requires benchmarks and human judgments that classical frameworks do not provide.

In summary, these reviews elucidate foundational theories, categorize early computational models, and highlight key cognitive mechanisms, yet none systematically examines the role of large language models or generative approaches in analogy generation or evaluation. In generate-and-test terms, the evaluator sets the yield of the loop: the proportion of generated analogies that constitute useful inference is only as knowable as the test is reliable. This gap motivates a 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 the PRISMA guidelines (Moher et al., 2010), we first used abstract, title, and keyword (ATK) search among established NLP and CS databases including ACM Digital Library¹, IEEE Xplore², SpringerLink³, ScienceDirect⁴, Wiley Online Library⁵, and ACL Anthology⁶ (Referred to as ACL throughout this paper). These sources collectively ensure a comprehensive coverage of both foundational and emerging research relevant to computational analogy generation and its evaluation.

For keyword search, we included the keywords *analogy*, *analogous*, and *analogical*, as those are common instances of analogy and its synonyms. We also included the related keyword *metaphor* to maximize recall, as the two terms are closely entangled in the literature. However, because analogy and metaphor have evolved into distinct research domains, each with its theoretical founda-

tions and frameworks (Rai and Chakraverty, 2020), records whose primary subject was metaphor rather than analogy were excluded during screening under our scope criterion. We then conducted a primary search including the keywords *generation*, *retrieval*, and *evaluation*, which covers 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 (Kitchenham et al., 2009). **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 (Kitchenham et al., 2009). **EC6**: None of the paper’s claimed contributions concern analogy generation or evaluation process.

3.2.2 Process

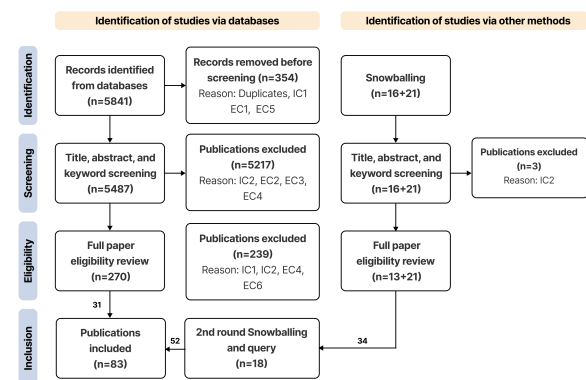


Figure 1: Pipeline of PRISMA framework used in this review process.

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/>

The keyword search was executed in April 2025 and identified 5,841 records across the six databases (IEEE Xplore 2,076; ScienceDirect 2,565; ACM Digital Library 734; ACL Anthology 223; SpringerLink 141; Wiley Online Library 102). Records were de-duplicated by registry-based DOI and title matching, leaving 5,680. We then removed 151 records published before 1980 (IC1), 23 grey-literature records (EC1), and 19 secondary studies (EC5), leaving 5,487 records for title, abstract, and keyword screening.

In the ATK screening stage, four authors randomly sampled 50 papers for a pilot screening round. Each author independently labeled the papers as *include*, *exclude*, or *uncertain*. The authors then met to resolve discrepancies and refine the inclusion and exclusion criteria. Each paper was reviewed by at least two authors to ensure consistency. This stage reduced the 5,487 records to 270, which advanced to full-text review.

Before proceeding to full screening, we assessed inter-annotator agreement on the pilot sample: four raters independently coded the 50 records against the inclusion and exclusion criteria, yielding a mean pairwise Cohen’s κ of 0.88, indicating strong agreement. All three disagreements concerned the boundary between analogy retrieval and analogy generation and were resolved by discussion. The adjudicated cases were codified as worked examples in our codebook.

Next, a full-text review was conducted to evaluate the selected papers. Another pilot round ($N = 20$) was conducted to further refine the inclusion and exclusion criteria. After this step, 31 of the 270 papers met the inclusion criteria and were included for analysis, with the remaining 239 excluded for documented reasons (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), adding 13 papers. In total, 44 papers (31 from the database search and 13 from snowballing) were included in the initial search and screening. Given the rapid pace of research in this area, we conducted a first update in April 2026 through forward and backward snowballing, applying the same inclusion and exclusion criteria; this added 21 papers, bringing the corpus to $N = 65$.

A second update, completed in August 2026, combined a fresh Anthology re-run, two rounds

of forward-backward snowballing, and a broadened title-level query (Appendix C.4), adding 18 papers for a final corpus of $N = 83$. The additions are concentrated in the automatic-accuracy evaluation cells and altered no conclusion of this review; per-observation effects are noted in §4. The full per-database query syntax and all record-level screening decisions are provided in Appendix C.

3.3 Data Extraction and Analysis

Data were extracted during the full-text eligibility review process described above. This included top-down coding for the following features: methodology (RQ1), evaluation metrics (RQ2), identified challenges (RQ3), and outcomes (RQ3) to answer each of the research questions. 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 Miro⁷, an online collaborative whiteboarding platform, as in prior reviews (Zhang et al., 2025). We held meetings to resolve disagreements and refine the groupings.

4 Results

Based on the 83 papers we examined during our SLR, we identified common themes and dimensions amongst the papers. Detailed results and paper selections are included in Appendix B.

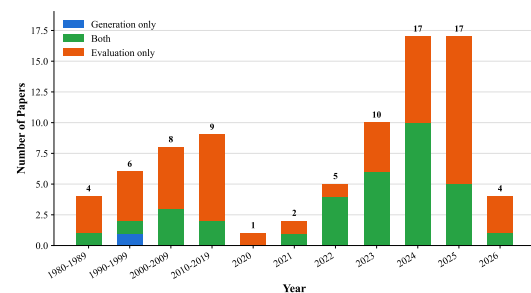


Figure 2: Publication years.

Publication Years. Fig.2 shows the distribution of publication years of all selected papers in the SLR. The trend indicates a gradual increase in publications from the 1980s through the 2010s. The slight decline during the early 2010s may be attributed to the rise of embedding-based models

⁷<https://miro.com/>

(Mikolov et al., 2013) that achieved strong performance on analogy tasks; during this period, much of the research shifted toward improving embedding techniques, which we excluded if their primary contribution was not directly related to analogy generation or evaluation. In recent years, however, there has been a noticeable resurgence of interest, driven by the emergence of LLMs (Ouyang et al., 2022).

4.1 Analogy Representation

We classify the computational models in our corpus into five types based on their underlying representation. This axis characterizes *architecture*—symbolic processing vs. embedding-based vs. LLM-internal representations—and is orthogonal to the task a paper addresses: the same generation or evaluation task can be attempted under any architecture. *Relational graph-based* models use explicit structured representations (knowledge graphs, ontologies) to align source and target domains, e.g., SME (Falkenhainer et al., 1989) performs graph-matching under one-to-one structural consistency constraints. *Distributional semantic* models compute relational similarity from corpus statistics, including word-embedding offsets that capture simple $A:B::C:D$ relations (Turney et al., 2006; Mikolov et al., 2013). *Cognitive/architectural* models, such as Copycat and LISA (Hofstadter and Mitchell, 1994; Hummel and Holyoak, 1996), simulate emergent binding or spreading activation, hybridizing symbolic and connectionist ideas. *Transformation-based* models interpret analogies as geometric relationships and use vector or character operations (Lepage and Ando, 1996; Plate, 2000). Finally, *learned relational representation* models learn dedicated relation embeddings or graph encoders for analogy, rather than relying on general distributional statistics or explicit symbol structures; DeepGAR (Ling et al., 2022) is the corpus exemplar, and the paradigm extends beyond our corpus to relation-embedding distillation and relation-embedding chains (Ushio et al., 2021a; Kumar and Schockaert, 2023).

We identified 20 papers in our corpus that fit into this taxonomy (see Tab. 3). Six papers use a relational graph-based method, nine take a distributional semantic approach, two use a cognitive/architectural model, two use a transformation-based model, and one learns dedicated relational representations. Our result shows that early

work relied on rule-based and graph-based methods, such as the relational graph-based model and the cognitive/architectural model; as machine learning advanced, learning-based distributional methods became more relevant and widely adopted. The remaining LLM-era papers in our corpus largely bypass all five explicit-representation paradigms; they instead rely on contextual embeddings or generative LLM-as-judge evaluation, raising new questions of measurement validity that we revisit in §4.5. Fig. 3 in Appendix B summarizes this taxonomy, situating the five explicit-representation paradigms alongside the corpus’s task-by-granularity structure and the LLM-prompting work that bypasses them.

4.2 Analogy Generation Framework

Generation by Granularity Based on our review, we identified two common targets of analogy generation, distinguished by the granularity and structure of the generated output: the first one is lexical-level analogy generation, which 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 autoregressive 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 story structures that preserve relational mappings (Mittal, 1992; Zhu and Ontanón, 2010).

Of the 35 works in our corpus concerning analogy generation, 19 are at the lexical level (14 of which use LLMs) and 17 at the compositional level, with one paper (Wang et al., 2024) addressing both (see Tab. 1). The compositional set spans story, scientific, historical, and educational analogies (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), analogy mining and mapping engines (Sultan and Shahaf, 2022; Jacob et al., 2023; Webb et al., 2023a; Ling et al., 2022), and a multimodal subset (Davies et al., 2008; Yaner and Goel, 2006; Sadeghi et al., 2015; Yiu et al., 2025; Jiang et al., 2024; Yilmaz et al., 2025; Lippolis et al., 2026), which we treat as compositional since it assesses relational structure across pair-wise inputs.

LLM-based Generation We identify 19 papers in our corpus using LLMs to generate analogies through prompting techniques (Refer to Tab. 2). In the identified collection, 12 papers use multi-step prompt pipelines to enhance generation quality, while seven earlier papers use single-shot prompting. This difference represents a shift from single-shot prompting to more structured prompt design processes. From the perspective of analogical reasoning, this shift aligns with the multi-stage cognitive process of analogical reasoning—retrieval, mapping, and transfer—suggesting that decomposing prompts into discrete steps may better guide LLMs and produce more coherent and relationally accurate analogies (Gentner and Forbus, 2011).

From the corpus, 11 of the 19 LLM-generation papers (58%) explicitly require a human-in-the-loop curation step rather than relying on the LLM’s autonomous output. The curation strategies vary: manual annotation and filtering (Sultan et al., 2024; Bernstein et al., 2024), user-information injection (Cao et al., 2024; Ju et al., 2025; Chen et al., 2024), and prompt tuning (Wang et al., 2024; Shao et al., 2025); one framework instead closes the loop without humans, iterating generate–evaluate–refine over automatic scoring functions (Bhavya et al., 2023). The common pattern is that human judgment is treated as a necessary stage of the generation pipeline, with multiple papers reporting quality issues in the initial generation (Bhavya et al., 2022; Ding et al., 2023; Cao et al., 2024; Shao et al., 2025). Among the 83 reviewed papers, autonomous LLM generation that reliably produces usable analogies has not yet been demonstrated; the strongest demonstrated paradigm is human curation of LLM candidates (11 of the 19 LLM-generation papers), and the limits of that paradigm remain unarticulated.

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 69 automatic evaluation approaches, 42 focus on the lexical level and 27 on the compositional level (See Tab. 4). On the human evaluation side, four papers use humans to evaluate lexical-level metrics and data, while 26 papers use experts or crowdsourcing to evaluate compositional-level metrics such as overall quality of the analogy gener-

ated (Sultan et al., 2024; Jiayang et al., 2023). 16 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 evaluation methodology of the reviewed literature: 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). We examine the consequences of this split in §4.5.

LLM-based Evaluation A recent slice of our corpus uses LLMs themselves as evaluators of analogies; Tab. 6 in Appendix B summarizes the four corpus members that do so. The approaches span direct judging of long-form analogies against ordinal criteria (Bhavya et al., 2024a), an LLM-in-the-metric design where GPT-4 decomposes events into dimensions before similarity scoring (Li et al., 2025), task-based evaluation that proxies analogy quality by a student model’s downstream gain (Yuan et al., 2024b), and a systematic study of whether LLMs can replace human analogy annotators (Zhang and Lyu, 2025). Reported agreement with humans ranges from modest (Krippendorff’s $\alpha \approx 0.35$) to strong (Spearman 0.73)—too unstable to certify LLM judges as substitutes for human evaluation at the compositional level. Adjacent to this line, AnaScore (Wang et al., 2025) offers a non-LLM automatic alternative that scores semantic parallelism with formalized, conceptual-knowledge-based explanations.

Quality Dimension Lastly, we categorized the reported evaluation metrics into: accuracy, validity, similarity, novelty, and human-preference/judgement metrics (See Tab. 5). Many automatic evaluation methods target one or more of these dimensions. Specifically, we identified 42 papers that use accuracy-based metrics, such as precision, recall, and F1-score; 22 papers that assess validity, including logical consistency checks and human validity judgments; and 23 papers that apply similarity-based metrics, such as BLEU, to evaluate generated analogies. Most recent LLM benchmarks in our corpus report 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’ judgement (Jiayang et al., 2023). Human preference/judgment is used in 20 papers within our corpus 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 analogy generated in 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. Only four of the 83 papers in our corpus 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 atom model (Gentner, 1983), were originally crafted by humans and then incorporated into computational models; in practice, both symbolic and neural systems frequently recycle 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 distribution.

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) report 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 design of 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 analogies generated are often complex to fit the 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 across the reviewed 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. This coverage limits the range of analogical phenomena that models can be evaluated on 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). A dedicated critique line within our corpus goes further, showing that the word-analogy test itself is an unreliable measure of relational knowledge: reported vector-offset successes depend on excluding the query terms, on hyperparameter choices, and on relation categories that reduce to word similarity (Schluter, 2018; Rogers et al., 2017; Fournier et al., 2020). 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 (42 of the 69 automatic evaluations) and human judgment dominates at the compositional level (26 of the 30 human evaluations), a method scoring well on lexical-automatic benchmarks (e.g., SAT analogy accuracy) carries 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). In short, no method in our corpus can yet automatically detect an analogy that is lexically strong but relationally shallow—the field still needs compositional-level automatic metrics or benchmarks that correlate meaningfully with human judgments.

Do LLMs Reason Analogically? A central question runs through the corpus’s recent 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 debate now sits inside our corpus. Webb et al. (2023b) report that GPT-3 matches or exceeds humans on letter-string, digit-matrix, four-term verbal, and story analogies; however, a non-archival reply (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. (2023b)’s four domains and documents brittleness to paraphrase, answer order, and counterfactual variants that controls do not eliminate.

Other corpus entries broaden the picture. In the human-comparison line, children generalize analogy solving across alphabets while frontier LLMs do not (Stevenson et al., 2026), and LLMs solve verbal analogies at roughly children’s level while leaning more on associative shortcuts (Johnson et al., 2025); LLMs match humans on many semantic structure-mapping conditions but diverge systematically in failure modes (Musker et al., 2025); the “analogical prompting” technique (Yasunaga et al., 2024) works about as well with random as with topical examples (Qin et al., 2025); and substantial cross-model variability complicates replication (de Varda et al., 2025).

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 in our corpus does not support a binary emergent-or-not verdict; it supports a partition in which relational tasks succeed and which fail. LLMs form invariant relational representations for some concept families, and succeed on the corresponding tasks, while falling back on surface similarity for others, with the partition tracking the prevalence of each relation in pretraining.

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., 2022) 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 11 of 19 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; Opielka 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.

In generate-and-test terms (Newell and Simon,

1972), we further recommend evaluating generators by their *yield*: the proportion of generated analogies that constitute useful, valid inference under a fixed evaluator. Our corpus approaches this quantity only obliquely—through novelty-as-distance scoring (Bhavya et al., 2023) and surface-versus-structural validity checks (Yuan et al., 2023)—but never operationalizes it as a property of the generator itself. Measuring yield presupposes a test that is reliable, scalable, and sensitive to relational structure, which the field currently lacks at the compositional level. The missing piece is thus an automated compositional-level evaluator: in the generate-and-test loop, evaluation, not generation, is the bottleneck.

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 and two subsequent updates, 83 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. 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.

Our corpus updates also have coverage limits. The broadened title-level query ran over the ACL Anthology only, so coverage of the five non-ACL databases after April 2025 relies on citation-based snowballing, and the precision-oriented conjunct query remains unmitigated there.

Acknowledgements

We thank the reviewers for feedback that substantially improved this review. We also thank Hangfeng He for his support. This work is supported by the National Science Foundation (NSF) RETTL program (Award No. 2238675).

References

- Stergos Afantenos, Miguel Couceiro, Emiliano Lorini, and Van-Duy Ngo. 2026. [Learning proportional analogies: Lightweight neural network vs large language models](#). In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 3592–3602. SCITEPRESS.
- 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*.
- Kara L. Combs, Isaiah Goble, Spencer V. Howlett, Yuki B. Adams, and Trevor J. Bihl. 2025. [Evaluating the tradeoff between analogical reasoning ability and](#)

- efficiency in large language models. *IEEE Transactions on Cognitive and Developmental Systems*, 17(6):1401–1410.
- 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.
- 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.
- Rikke Friis Dam and Teo Yu Siang. 2022. [Affinity diagrams: How to cluster your ideas and reveal insights](#). *Interaction Design Foundation—IxDF*.
- Mayukh Das and Wolf-Tilo Balke. 2025. [From prototypical to relational: How LLMs navigate complex analogies](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 465–485, Hanoi, Vietnam. Association for Computational Linguistics.
- Jim Davies, Ashok K Goel, and Patrick W Yaner. 2008. Proteus: Visuospatial analogy in problem-solving. *Knowledge-Based Systems*, 21(7):636–654.
- Andrea Gregor de Varda, Chiara Saponaro, and Marco Marelli. 2025. [High variability in LLMs’ analogical reasoning](#). *Nature Human Behaviour*, 9(7):1339–1341.
- 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.
- Kenneth D Forbus, Dedre Gentner, and Keith Law. 1995. Mac/fac: A model of similarity-based retrieval. *Cognitive science*, 19(2):141–205.
- Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. [Analogies minus analogy test: measuring regularities in word embeddings](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online. Association for Computational Linguistics.
- 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.
- 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.
- Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. [In-context analogical reasoning with pre-trained language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.
- William Hudson. 2013. [The encyclopedia of human-computer interaction, chapter card sorting](#). *The Interaction Design Foundation*, 2.

- John E Hummel and Keith J Holyoak. 1996. 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. Lawrence Erlbaum Associates.
- 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*, pages 16426–16442, Singapore. Association for Computational Linguistics.
- 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.
- Tamar Johnson, Mathilde ter Veen, Rochelle Choenni, Han van der Maas, Ekaterina Shutova, and Claire E Stevenson. 2025. [Do large language models solve verbal analogies like children do?](#) In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 627–639, Vienna, Austria. 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.
- Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. [Systematic literature reviews in software engineering – a systematic literature review](#). *Information and Software Technology*, 51(1):7–15.
- Nitesh Kumar and Steven Schockaert. 2023. [Solving hard analogy questions with relation embedding chains](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6224–6236, Singapore. Association for Computational Linguistics.
- 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.
- Chen Ling, Tanmoy Chowdhury, Junji Jiang, Junxiang Wang, Xuchao Zhang, Haifeng Chen, and Liang Zhao. 2022. [DeepGAR: Deep graph learning for analogical reasoning](#). In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1065–1070. IEEE.
- Anna Sofia Lippolis, Andrea Giovanni Nuzzolese, and Aldo Gangemi. 2026. [Enhancing multimodal analogical reasoning with logic augmented generation](#). *Information Fusion*, 133:104250.
- 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.
- Allen Newell and Herbert A Simon. 1972. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.
- 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](#). In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.
- 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.
- 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.
- 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.
- Oren Sultan and Dafna Shahaf. 2022. [Life is a circus and we are the clowns: Automatically finding analogies between situations and processes](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. 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.
- 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.
- 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.
- 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.
- Liyan Wang, Haotong Wang, and Yves LePage. 2025. [AnaScore: Understanding semantic parallelism in proportional analogies](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1175–1188, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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, Shuhao Fu, Trevor Bihl, Keith J. Holyoak, and Hongjing Lu. 2023a. [Zero-shot visual reasoning through probabilistic analogical mapping](#). *Nature Communications*, 14(1):5144.
- Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023b. [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. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.
- 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).

Xiaojing Zhang and Bochen Lyu. 2025. [Can language models serve as analogy annotators?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15853–15883, Vienna, Austria. Association for Computational Linguistics.

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

Category	Papers
Lexical-level	(Afantenos et al., 2026), (Bhavya et al., 2022), (Bhavya et al., 2023), (Boisson et al., 2025), (Bourrelly et al., 1983), (Chen et al., 2022), (Crouse et al., 2021), (Ding et al., 2023), (Jiayang et al., 2023), (Ju et al., 2025), (Salu, 1994), (Shao et al., 2025), (Shin et al., 2024), (Sultan et al., 2024), (Wang et al., 2024), (Wijesiriwardene et al., 2025), (Yasunaga et al., 2024), (Yuan et al., 2024a), (Yuan et al., 2023)
Compositional-level	(Bernstein et al., 2024), (Bhavya et al., 2024b), (Cao et al., 2024), (Chen et al., 2024), (Davies et al., 2008), (Jacob et al., 2023), (Li et al., 2005), (Li et al., 2025), (Ling et al., 2022), (Mittal, 1992), (Sadeghi et al., 2015), (Sultan and Shahaf, 2022), (Wang et al., 2024), (Webb et al., 2023a), (Yaner and Goel, 2006), (Yuan et al., 2024b), (Zhu and Ontanón, 2010)

Table 1: Analogy Generation by Granularity

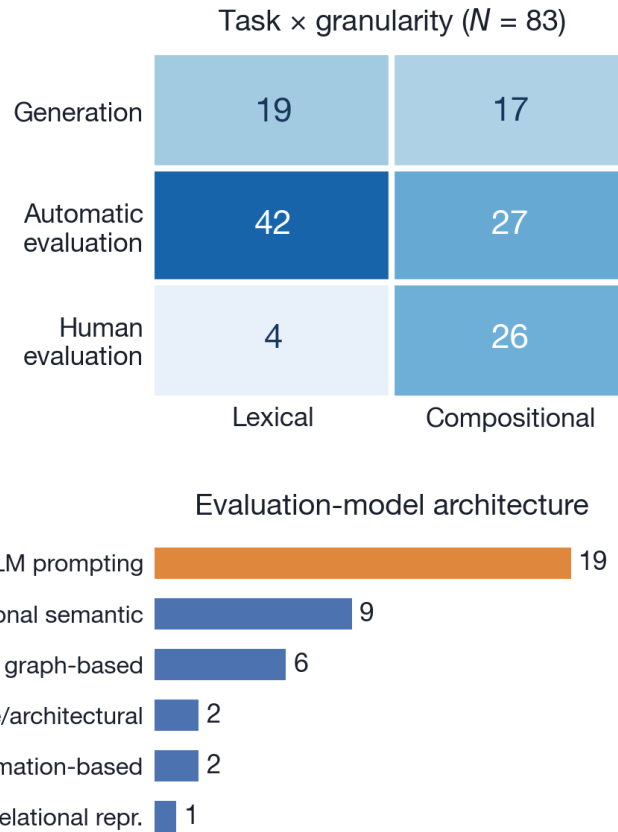


Figure 3: Taxonomy of the corpus ($N = 83$): tasks by granularity (top) and evaluation-model architectures (bottom). Task and architecture are independent dimensions; papers may contribute to multiple cells.

Type	Dimension	Papers
LLM	Human-in-the-loop	(Bernstein et al., 2024), (Boisson et al., 2025), (Cao et al., 2024), (Chen et al., 2024), (Jiayang et al., 2023), (Ju et al., 2025), (Shao et al., 2025), (Sultan et al., 2024), (Wang et al., 2024), (Yuan et al., 2024a), (Yuan et al., 2023)
	Multi-step Generation	(Bernstein et al., 2024), (Bhavya et al., 2023), (Boisson et al., 2025), (Cao et al., 2024), (Chen et al., 2024), (Jiayang et al., 2023), (Ju et al., 2025), (Li et al., 2025), (Shao et al., 2025), (Sultan et al., 2024), (Yuan et al., 2024a), (Yuan et al., 2024b)
	Single Prompt Generation	(Bhavya et al., 2022), (Ding et al., 2023), (Shin et al., 2024), (Wang et al., 2024), (Wijesiriwardene et al., 2025), (Yasunaga et al., 2024), (Yuan et al., 2023)
Non-LLM		(Afantenos et al., 2026), (Bhavya et al., 2024b), (Bourrelly et al., 1983), (Chen et al., 2022), (Crouse et al., 2021), (Davies et al., 2008), (Jacob et al., 2023), (Li et al., 2005), (Ling et al., 2022), (Mittal, 1992), (Sadeghi et al., 2015), (Salu, 1994), (Sultan and Shahaf, 2022), (Webb et al., 2023a), (Yaner and Goel, 2006), (Zhu and Ontanón, 2010)

Table 2: LLM vs. Non-LLM Generation

Category	Papers
Relational Graph-based	(Bourrelly et al., 1983), (Denaux and Gomez-Perez, 2019), (Falkenhainer et al., 1989), (Holyoak and Thagard, 1989), (Tsatsaronis et al., 2010), (Winston, 1980)
Distributional Semantic	(Bollegala, 2010), (Fournier et al., 2020), (Hope et al., 2017), (Li et al., 2023), (Rogers et al., 2017), (Schluter, 2018), (Turney et al., 2006), (Turney, 2008), (Wang et al., 2009)
Cognitive/architectural	(Hofstadter and Mitchell, 1994), (Hummel and Holyoak, 1996)
Transformation-based	(Lepage and Ando, 1996), (Plate, 2000)
Learned Relational Representations	(Ling et al., 2022)

Table 3: Evaluation Model Types

Mode	Granularity	Papers
Automatic	Lexical-level	(Afantenos et al., 2026), (Bhavya et al., 2022), (Bhavya et al., 2024b), (Bollegala et al., 2009), (Bollegala, 2010), (Chen et al., 2022), (Czinczoll et al., 2022), (Denaux and Gomez-Perez, 2019), (Fournier et al., 2020), (Hofstadter and Mitchell, 1994), (Hope et al., 2017), (Jiayang et al., 2023), (Johnson et al., 2025), (Lee et al., 2026), (Lepage and Ando, 1996), (Lewis and Mitchell, 2024), (Lewis and Mitchell, 2025), (Li et al., 2005), (Li et al., 2023), (Musker et al., 2025), (Plate, 2000), (Qin et al., 2025), (Rogers et al., 2017), (Sadeghi et al., 2015), (Schluter, 2018), (Shin et al., 2024), (Sultan et al., 2024), (Thagard et al., 1990), (Tsatsaronis et al., 2010), (Turney, 2008), (Ushio et al., 2021b), (Wang et al., 2024), (Wang et al., 2025), (Webb et al., 2025), (Wijesiriwardene et al., 2023), (Wijesiriwardene et al., 2025), (Winston, 1980), (Yang et al., 2025), (Yasunaga et al., 2024), (Yuan et al., 2024a), (Yuan et al., 2023), (Zhang and Lyu, 2025)
	Compositional-level	(Boisson et al., 2024), (Combs et al., 2025), (Crouse et al., 2021), (Das and Balke, 2025), (Falkenhainer et al., 1989), (Hu et al., 2023), (Hummel and Holyoak, 1996), (Jacob et al., 2023), (Jiang et al., 2024), (Ling et al., 2022), (Lippolis et al., 2026), (Plate, 2000), (Sourati et al., 2024), (Sultan and Shahaf, 2022), (Wang et al., 2009), (Webb et al., 2023b), (Webb et al., 2023a), (Wijesiriwardene et al., 2024), (Winston, 1980), (Yaner and Goel, 2006), (Ye et al., 2024), (Yilmaz et al., 2025), (Yiu et al., 2025), (Yuan et al., 2023), (Yuan et al., 2024b), (Zhang et al., 2017), (Zhu and Ontanón, 2010)
Human	Lexical-level	(Johnson et al., 2025), (Stevenson et al., 2026), (Turney, 2008), (Yuan et al., 2023)
	Compositional-level	(Bernstein et al., 2024), (Bhavya et al., 2022), (Bhavya et al., 2023), (Bhavya et al., 2024a), (Boisson et al., 2025), (Bourrelly et al., 1983), (Cao et al., 2024), (Chen et al., 2024), (Davies et al., 2008), (Ding et al., 2023), (Falkenhainer et al., 1989), (Hope et al., 2017), (Jiayang et al., 2023), (Ju et al., 2025), (Li et al., 2005), (Li et al., 2025), (Mittal, 1992), (Plate, 2000), (Shao et al., 2025), (Sourati et al., 2024), (Sultan and Shahaf, 2022), (Sultan et al., 2024), (Wang et al., 2024), (Webb et al., 2023b), (Yuan et al., 2024a), (Yuan et al., 2024b)

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

Category	Papers
Accuracy	(Afantenos et al., 2026), (Bhavya et al., 2024b), (Bhavya et al., 2024a), (Boisson et al., 2024), (Bollegala, 2010), (Chen et al., 2022), (Combs et al., 2025), (Crouse et al., 2021), (Czinczoll et al., 2022), (Das and Balke, 2025), (Hope et al., 2017), (Hu et al., 2023), (Jacob et al., 2023), (Jiang et al., 2024), (Johnson et al., 2025), (Lee et al., 2026), (Lewis and Mitchell, 2024), (Lewis and Mitchell, 2025), (Ling et al., 2022), (Lippolis et al., 2026), (Plate, 2000), (Qin et al., 2025), (Shin et al., 2024), (Sourati et al., 2024), (Stevenson et al., 2026), (Sultan and Shahaf, 2022), (Turney, 2008), (Ushio et al., 2021b), (Webb et al., 2023b), (Webb et al., 2023a), (Webb et al., 2025), (Wijesiriwardene et al., 2024), (Wijesiriwardene et al., 2025), (Yaner and Goel, 2006), (Yang et al., 2025), (Yasunaga et al., 2024), (Ye et al., 2024), (Yilmaz et al., 2025), (Yiu et al., 2025), (Yuan et al., 2024a), (Yuan et al., 2023), (Yuan et al., 2024b)
Similarity	(Bhavya et al., 2022), (Boisson et al., 2025), (Bollegala et al., 2009), (Denaux and Gomez-Perez, 2019), (Fournier et al., 2020), (Jiayang et al., 2023), (Lepage and Ando, 1996), (Li et al., 2023), (Plate, 2000), (Rogers et al., 2017), (Sadeghi et al., 2015), (Schluter, 2018), (Sultan et al., 2024), (Thagard et al., 1990), (Tsatsaronis et al., 2010), (Wang et al., 2009), (Wang et al., 2024), (Wang et al., 2025), (Wijesiriwardene et al., 2023), (Winston, 1980), (Ye et al., 2024), (Yuan et al., 2023), (Zhu and Ontanón, 2010)
Validity	(Bhavya et al., 2024a), (Bourrelly et al., 1983), (Crouse et al., 2021), (Hofstadter and Mitchell, 1994), (Holyoak and Thagard, 1989), (Hope et al., 2017), (Hummel and Holyoak, 1996), (Jiayang et al., 2023), (Lee et al., 2026), (Lewis and Mitchell, 2024), (Lewis and Mitchell, 2025), (Li et al., 2005), (Mittal, 1992), (Musker et al., 2025), (Qin et al., 2025), (Wang et al., 2025), (Webb et al., 2025), (Winston, 1980), (Yuan et al., 2024a), (Yuan et al., 2024b), (Zhang et al., 2017), (Zhang and Lyu, 2025)
Novelty	(Bhavya et al., 2022), (Bhavya et al., 2023), (Hope et al., 2017), (Jiayang et al., 2023)
Human Preference/Judgment	(Bernstein et al., 2024), (Bhavya et al., 2022), (Boisson et al., 2025), (Cao et al., 2024), (Chen et al., 2024), (Davies et al., 2008), (Ding et al., 2023), (Falkenhainer et al., 1989), (Ju et al., 2025), (Li et al., 2005), (Li et al., 2025), (Shao et al., 2025), (Sourati et al., 2024), (Stevenson et al., 2026), (Sultan and Shahaf, 2022), (Sultan et al., 2024), (Turney, 2008), (Wang et al., 2024), (Yuan et al., 2024a), (Yuan et al., 2024b)

Table 5: Analogy Evaluation Dimension

Paper	Judge model(s)	What is judged	Agreement w/ humans
Bhavya et al. (2024a)	GPT-4 (prompted baseline)	Long-form analogy quality on seven ordinal criteria	Krippendorff’s $\alpha \approx 0.35$ (pre-pilot)
Li et al. (2025)	GPT-4 (inside a multi-dimensional similarity metric)	Historical analogies along topic, background, process, and result	Pearson 0.72, Spearman 0.73 vs. human rankings
Yuan et al. (2024b)	Student LMs (GPT-3.5, Llama-3-8B, Mistral-7B, Vicuna-7B/13B)	Utility of teacher-LM analogies, proxied by scientific-QA gains	Not reported (task-based, indirect)
Zhang and Lyu (2025)	GPT-4o, Llama-3.1-8B/70B, Qwen-2.5-14B/32B/72B	Story-pair analogy annotation into six SMT-grounded categories	Scored against human gold labels (P/R/F1)

Table 6: LLM-based analogy evaluation approaches in our corpus.

C Search and Screening Protocol

The complete record-level registry behind this protocol, including the database export sheets, all screening decisions, the update-protocol hit lists, and the evidence table, is publicly available⁸.

C.1 Databases, Query, and Retrieval

All databases were searched in April 2025 with a single canonical Boolean query,

(analogy OR analogous OR analogical OR metaphor) AND (generation OR evaluation OR retrieval OR generate OR evaluate)

adapted to each portal’s field-search constraints. Tab. 7 lists, per database, the fields searched, the query form, the retrieval date, and the number of records exported into the screening registry. The *metaphor* stem was included for recall (§3.1). ScienceDirect returns relevance-ranked results: the April 2025 export

⁸<https://github.com/Scriptwonder/analogy-slr-registry>

captured the top 2,565 of 4,078 portal-reported hits; the update protocol (§C.4) partially mitigates this gap through citation snowballing.

Database	Field scope	Query	Retrieval date	Records exported
IEEE Xplore	title, abstract, author keywords	canonical	Apr. 2025	2,076
ScienceDirect	title, abstract, author keywords	canonical	Apr. 2025	2,565
ACM Digital Library	title, abstract, author keywords	canonical	Apr. 2025	734
ACL Anthology	title, abstract, keywords (regex over bulk export)	canonical (regex form)	Apr. 2025	223
SpringerLink (first export)	title, keywords	canonical	Apr. 2025	67
SpringerLink (second export)	title, keywords	canonical	Apr. 2025	74
Wiley Online Library	abstract	canonical	Apr. 2025	102
Total identified				5,841

Table 7: Search protocol per database. All rows execute the canonical Boolean query of §C.1, restricted to the fields each portal supports.

C.2 Deduplication and Screening Criteria

Registry-based deduplication (DOI matching, then normalized-title matching) removed 161 duplicate records: 5,841 $\rightarrow$ 5,680. All subsequent decisions were coded against the inclusion and exclusion criteria of §3.2, reproduced in Tab. 8 with the codes used in the screening registry.

Code	Criterion
IC1	Published in or after 1980, the year of a foundational computational-analogy work (Winston, 1980); the original search window closed in April 2025 and the update protocol (§C.4) extends coverage through 2026.
IC2	The research topic is primarily in NLP/AI/CS, and the contribution is relevant to computational analogy generation and its evaluation (novel systems or improvements upon previous work).
EC1	Grey literature: work-in-progress, workshop paper, poster, demo, extended abstract, or patent.
EC2	Not written in English.
EC3	Not archival (e.g., preprint-only).
EC4	The computational method lacks a concrete artifact (system, algorithm) or relies solely on human labor (e.g., crowdsourcing).
EC5	Solely a secondary study (survey, review, book).
EC6	None of the paper’s claimed contributions concern analogy generation or the evaluation process.

Table 8: Inclusion (IC) and exclusion (EC) criteria (§3.2) with registry codes.

C.3 Screening Stages

Pre-screening. From the 5,680 deduplicated records we removed 151 pre-1980 records (IC1), 23 non-full-paper records (EC1), and 19 secondary studies (EC5), leaving 5,487 records for title–abstract–keyword (ATK) screening.⁹

ATK screening. Four authors first piloted the criteria on a common sample of 50 records (inter-annotator agreement reported in §3.2; mean pairwise Cohen’s $\kappa = 0.88$). Every record was then screened by at least two authors; 270 records advanced to full-text review.

Full-text review. Each full text received a decision of *Accept*, *Reject*, or *Uncertain*. Of the 270 decisions, 46 (17%) were marked Uncertain and each was adjudicated by a second author (15 include / 31 exclude); borderline accepts were additionally spot-cross-checked. This stage yielded 31 papers. One round of backward snowballing added 13 papers, and a first update in April 2026, using forward and backward snowballing, added another 21, for 65 papers in total.

⁹The registry ledger recorded 5,681 and 5,488 for these stages; recounts with the ledger’s header row excluded give 5,680 and 5,487, the latter verified free of duplicate DOIs and titles.

C.4 Second Update Protocol (August 2026)

The second update extended the first (snowballing-only) update to a four-part protocol; all additions were screened under the criteria of Tab. 8.

1. **Anthology re-run (2025–26).** The original query re-executed over a fresh ACL Anthology snapshot for 2025–2026 yielded 47 new candidates, of which 3 met all criteria (Wang et al., 2025; Zhang and Lyu, 2025; Das and Balke, 2025). A targeted recall check over analogy-titled Anthology papers lacking a task-verb stem (2018–2024) added a further 7 (Ushio et al., 2021b; Sultan and Shahaf, 2022; Jacob et al., 2023; Czinczoll et al., 2022; Hu et al., 2023; Fournier et al., 2020; Johnson et al., 2025).
2. **Snowballing, round 1.** Forward and backward citation chasing (Semantic Scholar) from all 65 corpus papers produced 6,130 registry-deduplicated candidates; screening added 4 papers (Webb et al., 2023a; Ling et al., 2022; Afantenos et al., 2026; Combs et al., 2025); applying the same criteria to the archival studies of the capability debate added 2 (Webb et al., 2023b; Yang et al., 2025) (non-archival replies remain context citations under EC3).
3. **Snowballing, round 2.** Repeating the procedure from every newly added paper (16 seeds) produced 1,140 candidates, 813 of them new after deduplication; screening added one further paper (Lippolis et al., 2026), suggesting the search had reached saturation.
4. **Broadened title-level query.** To cover analogy-central papers phrased without a task-verb stem (*identifying, mining, mapping, solving, probing*), we ran *analog*-in-title* with no second conjunct over the full 1980–2026 Anthology snapshot: 193 hits (87 pre-2018 hits never previously screened; 106 hits from 2018–2026, all reconciled with earlier decisions). One paper was added (Rogers et al., 2017); title-query candidates were retained only when cited by at least one paper already in the corpus.

Together, the August 2026 update added 18 papers (3 from the re-run, 7 from the recall check, 4 and 1 from the two snowballing rounds, 2 capability-debate studies, and 1 from the title-level query), for a final corpus of $N = 83$.

D Codebook

This codebook consolidates the coding criteria behind Tab. 1–5. The coding unit is the paper; every paper was coded by at least two authors, with disagreements resolved by discussion. For each dimension we state whether its categories are *mutually exclusive* (one code per paper) or *inclusive* (a paper may carry several), define each category, and give an example from the corpus.

D1: Task (exclusive). Coded at the level of the paper’s *primary contribution*; internal components (e.g., a generator’s built-in evaluator) do not change the task code. **Generation:** the system produces a novel analogy artifact — a string, pair, or mapping not present in its input resources (e.g., prompting an LLM to author natural-language analogies (Bhavaya et al., 2022)). **Evaluation:** the contribution is a metric, protocol, benchmark, or critique for assessing analogy quality (e.g., the AnaScore metric (Wang et al., 2025)). **Retrieval/mining:** the system locates or ranks *existing* analogies or analogous pairs in corpora or knowledge bases (e.g., mining analogous process pairs from text collections (Sultan and Shahaf, 2022)). **Capability evaluation:** an empirical or mechanistic study of whether computational systems possess analogical ability (e.g., (Webb et al., 2023b)). *Boundary example:* Hu et al. (2023) is coded capability evaluation, *not* generation: it measures whether pre-trained language models solve textualized Raven’s Progressive Matrices in context and produces no analogy artifacts.

D2: Generation granularity (exclusive, one logged exception). Applies to papers whose systems output analogy artifacts (Tab. 1). **Lexical:** output at the word or short-phrase level, prototypically proportional analogies $a:b::c:d$ (e.g., (Afantenos et al., 2026)). **Compositional:** output expressing relational structure over systems of entities — sentences, narratives, or explicit mappings (e.g., the FAME mapping engine (Jacob et al., 2023)). *Exception:* Wang et al. (2024) generates at both levels and is the single paper coded in both categories.

D3: Generator type (exclusive), with sub-axes. **LLM vs. non-LLM** (Tab. 2); non-LLM covers symbolic, neural pre-LLM, and probabilistic systems (e.g., probabilistic analogical mapping (Webb et al., 2023a)). LLM generation carries a sub-axis: **multi-step** (a pipeline of prompts, retrieval calls, or refinement stages, e.g., (Ju et al., 2025)) vs. **single-prompt** (one prompt produces the artifact, e.g., (Ding et al., 2023)). The **human-in-the-loop (HITL)** flag marks pipelines in which a human curation or filtering step is *required* to produce the final outputs; offline human labor (annotating training data, calibrating scorers) does not qualify. *Contrast pair:* Cao et al. (2024) is HITL — its pipeline injects user information with a human curation step; Bhavya et al. (2023) is *not* HITL — candidate analogies are filtered by automatic scorers, and humans only train those scorers offline.

D4: Evaluation mode (inclusive). **Automatic:** evaluation computed without human judgment of outputs (e.g., (Ushio et al., 2021b)). **Human:** evaluation by human raters (e.g., (Davies et al., 2008)). A paper reporting both is coded in both (e.g., Johnson et al. (2025) combines benchmark accuracy with human comparison).

D5: Evaluation granularity (inclusive across a paper’s evaluations). **Lexical:** evaluation targets word/phrase-level correctness, e.g., answer selection on *a:b::c:d* items (Schluter, 2018). **Compositional:** evaluation targets relational structure over sentences, narratives, or scenes (e.g., (Ye et al., 2024)). A paper whose evaluations span both levels appears in both (e.g., (Yuan et al., 2023)). *Caption rule* (Tab. 4): multimodal benchmarks — KiVA (Yiu et al., 2025), MARVEL (Jiang et al., 2024), VOILA (Yilmaz et al., 2025) — are coded compositional, since they assess relational structure across image pairs.

D6: Evaluation model paradigm (exclusive). Applies to the subset of papers whose evaluated system embodies a representational paradigm (Tab. 3). **Relational graph-based:** explicit symbolic predicate/graph structures aligned by structure mapping (e.g., SME (Falkenhainer et al., 1989)). **Distributional semantic:** relational similarity computed in corpus-derived vector spaces (e.g., (Turney et al., 2006)). **Cognitive-architectural:** process models simulating human analogy-making within a cognitive architecture (e.g., Copycat (Hofstadter and Mitchell, 1994)). **Transformation-based:** analogy modeled as an explicit algebraic or vector transformation between representations (e.g., (Lepage and Ando, 1996)). **Learned relational representations:** models trained end-to-end to learn dedicated relation embeddings or relational graph encoders for analogical matching; exemplar: DeepGAR (Ling et al., 2022). Papers whose primary contribution is generic relation-representation learning without an analogy generation or evaluation contribution (the ReBERT line and successor relation-embedding work) fall under EC6 and are cited as context.

D7: Evaluation dimension (inclusive). **Accuracy:** agreement with gold answers on a benchmark (e.g., (Crouse et al., 2021)). **Similarity:** graded relational-similarity scores, often correlated with human ratings (e.g., (Bollegala et al., 2009)). **Validity:** whether an analogy is structurally and semantically sound — the mapping actually holds (e.g., (Wang et al., 2025)). **Novelty:** distance of an output from known or corpus-attested analogies (e.g., (Bhavya et al., 2023)). **Human preference/judgement:** direct human ratings, rankings, or acceptance of outputs (e.g., (Sultan et al., 2024)).

Boundary rule: retrieval vs. generation. All three disagreements in the 50-record pilot (§3.2) sat on this boundary, and adjudication produced the rule applied throughout: retrieving or ranking an *existing* analog is retrieval; producing a *novel* analogy string or mapping is generation — the novel-artifact criterion governs. *Example:* a case-based system that retrieves candidate source cases and ranks them for analogical fit is coded retrieval, because the delivered artifact already existed; a system that retrieves candidate source domains and then *verbalizes a new source–target comparison* is coded generation (with a retrieval component), because the delivered comparison is novel.

Inter-annotator agreement for these criteria is reported in §3.2 ($\kappa = 0.88$).

E Characteristics of Included Papers

Table 9: Characteristics of the included papers ($N = 83$), sorted by year.

Paper	Task	Granularity	Architecture	Domain	Evaluation	Human study
Winston 1980	Mining	Lexical	Symbolic	General	similarity, validity	—
Bourrelly et al. 1983	Generation	Lexical	Symbolic	General	validity	Expert
Falkenhainer et al. 1989	Mining	Compositional	Symbolic	Scientific	validity, preference	Expert
Holyoak and Thagard 1989	Mining	Lexical	Symbolic	General	validity	—
Thagard et al. 1990	Mining	Lexical	Symbolic	General	similarity	—
Mittal 1992	Generation	Compositional	Symbolic	General	validity, preference	Expert
Hofstadter and Mitchell 1994	Mining	Lexical	Cognitive	General	validity	—
Salu 1994	Generation	—	Cognitive	General	—	—
Hummel and Holyoak 1996	Mining	Compositional	Cognitive	General	validity	—
Lepage and Ando 1996	Mining	Lexical	Transformation	General	similarity	—
Plate 2000	Mining	Lexical	Transformation	General	accuracy, similarity	Expert
Li et al. 2005	Generation	Compositional	Symbolic	General	validity, preference	Expert
Turney et al. 2006	Mining	Lexical	Distributional	General	accuracy, similarity	Comparison
Yaner and Goel 2006	Generation	Compositional	Symbolic	Multimodal	accuracy	—
Davies et al. 2008	Generation	Compositional	Symbolic	Multimodal	preference	Expert
Turney 2008	Mining	Lexical	Distributional	General	accuracy, similarity, preference	Comparison
Bollegala et al. 2009	Mining	Lexical	Distributional	General	similarity	—
Wang et al. 2009	Mining	Compositional	Distributional	General	similarity	—
Bollegala 2010	Mining	Lexical	Distributional	General	accuracy, similarity	—
Tsatsaronis et al. 2010	Mining	Lexical	Symbolic	General	similarity	—
Zhu and Ontanón 2010	Generation	Compositional	Symbolic	Story	similarity	—
Sadeghi et al. 2015	Generation	Compositional	Learned relational	Multimodal	accuracy, similarity	—
Hope et al. 2017	Mining	Compositional	Distributional	Scientific	accuracy, validity, novelty, preference	Crowd
Rogers et al. 2017	Evaluation	Lexical	Distributional	General	similarity	—
Zhang et al. 2017	Mining	Compositional	Symbolic	Historical	validity	—
Schluter 2018	Mining	Lexical	Distributional	General	similarity	—
Denaux and Gomez-Perez 2019	Mining	Lexical	Symbolic	General	similarity	—
Fournier et al. 2020	Evaluation	Lexical	Distributional	General	similarity	—
Crouse et al. 2021	Generation	Lexical	Learned relational	General	accuracy, validity	—
Ushio et al. 2021b	Capability evaluation	Lexical	LLM prompt-ing	General	accuracy	—
Bhavya et al. 2022	Generation	Lexical	LLM prompt-ing	Educational	accuracy, similarity, novelty, preference	Crowd
Chen et al. 2022	Generation	Lexical	Symbolic	General	accuracy	—
Czinczoll et al. 2022	Capability evaluation	Lexical	LLM prompt-ing	Scientific	accuracy	—
Ling et al. 2022	Generation	Compositional	Learned relational	General	accuracy	—
Sultan and Shahaf 2022	Generation	Compositional	Symbolic	Story	accuracy, preference	Crowd
Bhavya et al. 2023	Generation	Lexical	LLM hybrid	Educational	novelty	Crowd
Ding et al. 2023	Generation	Compositional	LLM prompt-ing	Educational	preference	Expert
Hu et al. 2023	Capability evaluation	Compositional	LLM prompt-ing	General	accuracy	—
Jacob et al. 2023	Generation	Compositional	Symbolic	General	accuracy	—
Jiayang et al. 2023	Generation	Lexical	LLM hybrid	Story	similarity, validity, novelty, preference	Crowd
Li et al. 2023	Mining	Lexical	Distributional	General	accuracy, similarity	—
Webb et al. 2023b	Capability evaluation	Compositional	LLM prompt-ing	General	accuracy	Comparison
Webb et al. 2023a	Generation	Compositional	Symbolic	Multimodal	accuracy	—
Wijesiriwardene et al. 2023	Mining	Lexical	LLM prompt-ing	General	similarity	—
Yuan et al. 2023	Generation	Lexical	LLM prompt-ing	Scientific	accuracy, similarity, preference	Crowd
Bernstein et al. 2024	Generation	Compositional	LLM hybrid	Educational	preference	Crowd

continued on next page

(continued)

Paper	Task	Granularity	Architecture	Domain	Evaluation	Human study
Bhavya et al. 2024b	Generation	Compositional	Symbolic	General	accuracy	Crowd
Bhavya et al. 2024a	Evaluation	Compositional	LLM prompt- ing	Educational	accuracy, validity	Crowd
Boisson et al. 2024	Evaluation	Compositional	LLM prompt- ing	General	accuracy	—
Cao et al. 2024	Generation	Compositional	LLM prompt- ing	Educational	preference	Crowd
Chen et al. 2024	Generation	Compositional	LLM hybrid	Educational	preference	Crowd
Jiang et al. 2024	Capability evaluation	Compositional	LLM prompt- ing	Multimodal	accuracy	—
Lewis and Mitchell 2024	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy, validity	—
Shin et al. 2024	Generation	Lexical	LLM prompt- ing	General	accuracy	—
Sourati et al. 2024	Evaluation	Compositional	LLM prompt- ing	Story	accuracy, preference	Comparison
Sultan et al. 2024	Generation	Lexical	LLM hybrid	General	similarity, preference	Crowd
Wang et al. 2024	Generation	Both	LLM hybrid	General	accuracy, similarity, preference	Expert
Wijesiriwardene et al. 2024	Evaluation	Compositional	LLM prompt- ing	General	accuracy	—
Yasunaga et al. 2024	Generation	Lexical	LLM prompt- ing	General	accuracy	—
Ye et al. 2024	Evaluation	Compositional	LLM prompt- ing	General	accuracy, similarity	—
Yuan et al. 2024a	Generation	Lexical	LLM hybrid	General	accuracy	Crowd
Yuan et al. 2024b	Generation	Compositional	LLM hybrid	Scientific	accuracy, validity, preference	Crowd
Boisson et al. 2025	Generation	Lexical	LLM hybrid	Story	accuracy, similarity, preference	Crowd
Combs et al. 2025	Capability evaluation	Compositional	LLM prompt- ing	General	accuracy	—
Das and Balke 2025	Capability evaluation	Compositional	LLM prompt- ing	General	accuracy	—
Johnson et al. 2025	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy	Comparison
Ju et al. 2025	Generation	Lexical	LLM hybrid	Story	preference	Crowd
Lewis and Mitchell 2025	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy, validity	—
Li et al. 2025	Generation	Compositional	LLM hybrid	Historical	preference	Crowd
Musker et al. 2025	Capability evaluation	Lexical	LLM prompt- ing	General	validity	Comparison
Qin et al. 2025	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy, validity	—
Shao et al. 2025	Generation	Lexical	LLM hybrid	Scientific	accuracy, preference	Crowd
Wang et al. 2025	Evaluation	Lexical	LLM prompt- ing	General	similarity, validity	—
Webb et al. 2025	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy, validity	—
Wijesiriwardene et al. 2025	Generation	Lexical	LLM prompt- ing	General	accuracy	—
Yang et al. 2025	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy	—
Yilmaz et al. 2025	Capability evaluation	Compositional	LLM prompt- ing	Multimodal	accuracy	—
Yiu et al. 2025	Capability evaluation	Compositional	LLM prompt- ing	Multimodal	accuracy	—
Zhang and Lyu 2025	Evaluation	Lexical	LLM prompt- ing	General	validity	—
Afantenos et al. 2026	Generation	Lexical	Symbolic	General	accuracy	—
Lee et al. 2026	Capability evaluation	Lexical	LLM prompt- ing	General	accuracy, validity	—
Lippolis et al. 2026	Capability evaluation	Compositional	LLM hybrid	Multimodal	accuracy	—

continued on next page

(continued)

Paper	Task	Granularity	Architecture	Domain	Evaluation	Human study
Stevenson et al. 2026	Capability evaluation	Lexical	LLM prompt-ing	General	accuracy	Comparison