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

Anonymous ACL submission

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 (CS), AI, and natural language processing (NLP), focusing on methods for analogy generation and evaluation. We categorized existing approaches across various dimensions, from symbolic, embedding-based, to LLM-driven methods, and identified core challenges, including difficulties in generating novel analogies, conflating relational and literal similarity, and limitations in current evaluation metrics and datasets. Based on this analysis, we propose future directions aimed at enhancing both the generation process and the quality of outputs in analogy generation and evaluation systems.

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). As a core cognitive function, analogical reasoning has been the subject of extensive research over several decades, focusing on how analogies are formed and how their quality can be assessed (Hofstadter, 1995).

Computational approaches to analogical reasoning, which involve identifying and mapping relational correspondences between a known *source* and a novel *target* to transfer insights, have attracted sustained interest in artificial intelligence (AI) and natural language processing (NLP). Earlier methods, such as Winston's frame-based system (Winston, 1980a) and Structure-Mapping Engine (SME) (Falkenhainer et al., 1989), used symbolic approaches by representing input analogies as structured sets of logical statements. These approaches primarily relied on hand-crafted, human-annotated analogies, which were evaluated based on the alignment of relational structures.

The advances in machine learning (ML) have enabled neural models to learn and predict analogical relationships, particularly in the form of word and proportional analogies (e.g., `France is to Paris as Italy is to Rome`) (Mikolov et al., 2013). Neural architectures such as embedding models are trained and evaluated through word-analogy and sentence-analogy datasets (Turney, 2008; Mikolov et al., 2013) filled with such questions to show their ability to perform analogical reasoning in a confined selection (Ushio et al., 2021).

Pretrained language models, such as GPT (Ouyang et al., 2022), have introduced generative capabilities, extending analogy generation beyond the word-level to more complex forms. These advancements support a wide range of applications, including automatic analogy mining applied to information retrieval (Bhavya et al., 2023) and personalized analogy generation tailored to individual users (CAO et al., 2024).

The evaluation of analogies remains a central and ongoing challenge in the field. Structure Mapping Theory (SMT) (Gentner, 1983) has been a major theoretical framework for analogy evaluation. It emphasizes mappings between the relations of two entities. Since then, relational similarity and word similarity have been a primary automatic evaluation metric for analogy (Turney et al., 2006). Complementary to these approaches, human evaluation,

conducted via expert judgment or crowdsourcing, continues to play a significant role. More recently, large language models (LLMs) have enabled hybrid evaluation strategies, in which analogies are rated using model-generated assessments across multiple criteria (Bhavya et al., 2024a).

Although prior surveys have reviewed analogical reasoning methods, including symbolic and neural approaches (Mitchell, 2021; Gentner and Forbus, 2011; French, 2002), there is a lack of comprehensive reviews that capture advanced approaches since the rise of LLMs. A growing body of recent research explores analogy generation and evaluation with LLMs (Bhavya et al., 2022; Yuan et al., 2023b; CAO et al., 2024); however, these studies remain fragmented—focused on isolated tasks like prompt engineering, dataset creation, or specific application—without a cohesive narrative linking early symbolic and distributional approaches to modern LLM-based methods. This review will situate new advanced techniques within the field's historical arc, reveal how foundational challenges have been reframed, and guide researchers toward unified best practices.

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 4,641 papers among six databases, resulting in reviewing 45 papers to discern the directions and methods in the domain of computational analogy generation and evaluation. We categorized existing approaches across multiple dimensions and identified key challenges such as the difficulty of generating novel analogies, the conflation of relational and literal similarity, and the limitations of current evaluation metrics and datasets.

In summary, our paper makes the following contributions: (1) a systematic literature review on the existing research related to analogy generation and evaluation; (2) a summary of four main categories of computational analogy generation methods and their corresponding evaluation metrics; (3) a highlight on challenges faced by generation and evaluation, and multiple future research directions.

2 Related Work

2.1 Analogy Generation and Evaluation

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

Early analogy generation methods relied on handcrafted rules to detect and evaluate analogies (Falkenhainer et al., 1989). Over time, these approaches evolved to symbolic structure, such as LRME (Turney, 2008), uses explicit graph-matching to align relational schemas between a source and target domain, to statistical embedding approaches (Mikolov et al., 2013). Most recently, prompt-based LLM pipelines leverage large pre-trained models to generate rich, context-sensitive analogies with minimal human effort (Ding et al., 2023; Yuan et al., 2023b). Each generation paradigm progressively reduces reliance on handcrafted representations while increasing flexibility and domain coverage, yet also introduces new challenges in controlling output coherence, ensuring relational fidelity, and mitigating model biases (Wijesiriwardene et al., 2023a; Yuan et al., 2023a; Bhavya et al., 2024a).

Early systems primarily focused on evaluating given analogies through rules and restraints (Holyoak and Thagard, 1989; Falkenhainer et al., 1989), and they relied heavily on human validation and relational similarity checks to ensure soundness. While human evaluation remains an essential component, the growth of machine learning has led to widespread use of automatic metrics, such as precision, recall, and F1-score, as well as similarity-based metrics like BLEU and ROUGE. Recent evaluations often use a combination of automatic and human evaluation, where automatic metrics are used to test the relational similarity of analogies on a lexical level, and human judges are used to evaluate the total soundness of the analogies (Yuan et al., 2023a; Jiayang et al., 2023).

2.2 SLRs in Computational Analogy Model

Prior surveys have framed computational analogy models in complementary ways. French (2002) provides a historical overview, classifying computational analogy models into symbolic, connectionist, and hybrid paradigms. By contrast, Gentner and Forbus (2011) analyzed analogies through the lens of computational models. The work decomposes the analogy into subprocesses: retrieval, mapping, abstraction, and re-representation. Gentner and Forbus (2011) emphasizes that analogical mappings favor systematic and higher-order relational correspondences. Both reviews (French, 2002; Gentner and Forbus, 2011), along with several other work (Gentner, 1983; Hofstadter, 1995) underscore that analogical inference relies on structured, relational representations and selective correspondence, but differ in focus: French (2002) surveys broad model families and open problems, whereas Gentner and Forbus (2011) drill into the computational models of mapping (for example, comparing symbolic systems like MAC/FAC (Forbus et al., 1995) and SME versus cognitive-inspired models like LISA (Hummel and Holyoak, 2019a) and DORA (Doumas et al., 2008)).

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

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

3 Methodology

3.1 Identfictaion

Following the PRISMA guidelines (Moher et al., 2010), we first used abstract, title, and keyword (ATK) search among the online NLP and other CS databases including ACM Digital Library [1], IEEE Xplore [2], SpringerLink [3], ScienceDirect [4], Wiley Online Library [5], and ACL Anthology (Referred to as ACL throughout this paper) [6]. ACL is recognized as a primary repository for NLP research. IEEE represents a leading community that contains the pioneering research in Engineering and Technology. ACM represents the comprehensive work in Human Computer Interaction (HCI) and other CS related fields. ScienceDirect contains interdisciplinary work across CS and cognitive science domains. SpringerLink and Wiley offer access to both theoretical and applied research across artificial intelligence, computational linguistics, and psychology, which are essential for understanding analogical reasoning from both computational and cognitive perspectives. 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 did not include the related keyword *metaphor* because our preliminary investigation revealed that analogy and metaphor have evolved into distinct research domains, each with its theoretical foundations and frameworks (Rai and Chakraverty, 2020). We then conducted a primary search including the keywords *generation*, *retrieval*, and *evaluation*. The term retrieval was selected because it captures both cognitive and computational processes fundamental to analogical reasoning, particularly in models that simulate memory or information access (Kolodner, 2014; Falkenhainer et al., 1989).

[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/

Generation has gained prominence in the era of LLMs, where producing analogies is often framed as a generative task (Bhavya et al., 2022; Sultan et al., 2024). Similarly, evaluation is essential for assessing the quality and effectiveness of generated or retrieved analogies, especially in empirical or automated settings.

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, 1980b; Falkenhainer et al., 1989); **IC2**: The research topic is primarily in NLP/AI/CS, and the contribution is relevant to computational analogy generation and its evaluation; this could be proposing novel systems or improvements upon previous works.

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

3.2.2 Process

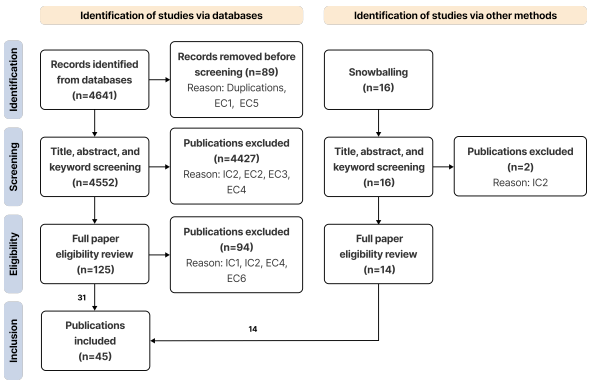

Figure 1: Pipeline of PRISMA framework used in this review process.

The initial keyword search across six databases yielded 4641 papers. After removing duplicates, papers published before 1980, non-full papers, and secondary studies, 4552 papers remained and were subsequently included in the screening process.

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. During the main screening process, each paper was reviewed by at least two authors to ensure consistency. This stage resulted in 125 papers.

Next, a full-text review was conducted to assess the eligibility of the selected papers. Another pilot round (N=20) was used to refine the inclusion and exclusion criteria further. After this step, 31 papers met the inclusion criteria and were included for analysis. Refer to Fig. 1 for the entire filtering pipeline.

Additionally, one round of backward snowballing was conducted to identify relevant studies that may have been missed during the initial search (Jalali and Wohlin, 2012). During the full paper review, we examined the related works of the included papers to identify relevant papers that align with our research questions. This process yielded 16 additional papers, 14 of which were eligible for full-text review. All 14 were included in the final corpus. Eventually, a total of 45 papers were included in the final analysis.

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 including 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 Miro[7], an online collaborative whiteboarding platform (Zhang et al., 2025). We held meetings to resolve disagreements and refine the groupings.

[7] https://miro.com/

4 Results

Based on the 45 papers we examined during our SLR, we identified common themes and dimensions amongst the papers. Due to limited space, we mainly provide a high-level summary of our findings. Detailed results and paper selections are included in Appendix B.

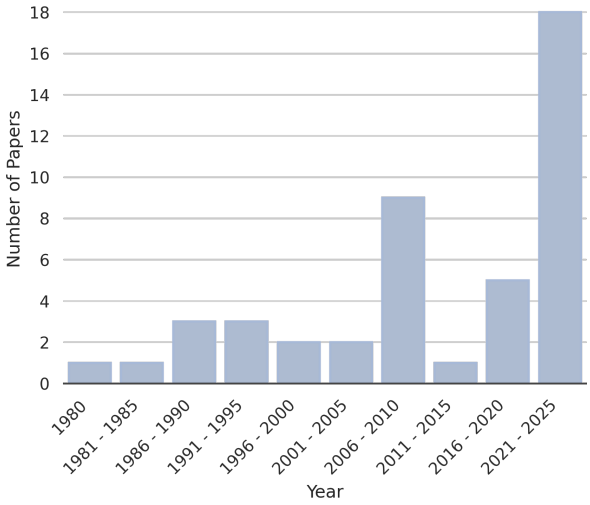

Figure 2: Publication years.

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 (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 in interest, driven by the emergence of LLMs (Ouyang et al., 2022), highlighting renewed attention and growing popularity in computational analogy research.

4.1 Analogy Generation Framework

We first examined papers from the 45 selected corpus, concerning analogy generation characterized by producing or generating one or multiple analogy pairs based on a target concept or domain (RQ1). During the bottom-up analysis process, we categorize the 22 related papers into two dimensions: the granularity (i.e., lexical-level and compositional-level) and the generation method (i.e., LLM-based).

4.1.1 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 word-by-word prediction nature of the method (Bhavya et al., 2024b). The second category is compositional-level analogy generation, which involves generating analogies at a higher level of abstraction, such as detailed analogies of scientific concepts, or coherent story structures that preserve relational mappings across larger contexts (Mittal, 1992; Zhu and Ontanón, 2010).

Of the 22 works concerning analogy generation, 15 proposed generation methods reside at the lexical level. This includes 11 papers that utilized LLMs for generation. Besides these works, we found four papers working on compositional-level analogy generation (Zhu and Ontanón, 2010; Li et al., 2005; Bhavya et al., 2024b; Mittal, 1992), including story generation and explanation generation (See Table.1).

We also noticed three visual analogy works apart from textual analogy generation (Davies et al., 2008; Yaner and Goel, 2006; Sadeghi et al., 2015). Visual problem solving is an essential aspect of analogical reasoning (Lovett and Forbus, 2017), and has been viewed as a way to measure Artificial General Intelligence [8]. These works demonstrate generation approaches beyond textual modality and highlight a promising direction towards a more intelligent multimodal analogy generation.

4.1.2 LLM-based Generation

We identify 11 papers using LLMs to generate analogies through prompting techniques (Refer to Table.2). In the identified collection, seven papers use multi-step prompt pipelines to enhance the generation quality, while four earlier papers use single-shot prompting to generate. This difference represents a shift from single-shot prompting to more structured prompt design processes as LLMs become more powerful and researchers become more familiar with AI-assistance tools. From the perspective of analogical reasoning, this shift aligns with the multi-stage cognitive process of analogi-

[8] https://arcprize.org/arc-agi

cal reasoning—retrieval, mapping, and transfer—suggesting that decomposing prompts into discrete steps may better guide LLMs through these stages and produce more coherent and relationally accurate analogies (Gentner and Forbus, 2011).

Throughout these LLM papers, we found eight papers that use human-in-the-loop approaches to enhance the generation quality through strategies such as manual annotation and filtering (Sultan et al., 2024; Bhavya et al., 2023), user-related information injecting (CAO et al., 2024; Ju et al., 2025), and prompt tuning (Wang et al., 2024; Shao et al., 2025). The lack of quality in the initial generation is mentioned multiple times (Bhavya et al., 2022; Ding et al., 2023; CAO et al., 2024; Shao et al., 2025), specifying a need for human-centered iterations.

4.2 Analogy Evaluation Methods

In our analogy evaluation report, we first identify papers relevant to RQ2. Evaluating analogy quality is a highly researched question, and we present our review result through the three lenses below.

4.2.1 Evaluation by Granularity

We identify the evaluation methods using the same granularity(lexical-level and compositional-level) used in generation analysis and differentiate the human evaluation approach from the automatic one. Through 34 automatic evaluation approaches, 24 are focused on the lexical level, and 10 are concentrated on the compositional level evaluation (See Table.4). On the human evaluation side, two papers use humans to evaluate lexical-level metrics and data, while 19 papers use experts or crowdsourcing to evaluate compositional-level metrics such as overall quality of the analogy generated (Sultan et al., 2024; Jiayang et al., 2023). 11 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 shows that the automatic approach is commonly used at assessing analogies on the lexical-level, while human evaluation is conducted to test analogies on the compositional-level.

4.2.2 Model Type

We identify four different models to evaluate analogy. First, relational graph-based models use explicit structured representations (graphs such as knowledge graph and entity-relation representation of texts or trees such as ontology and lexicon) to align a familiar source domain with a novel target domain. A primary example is SME (Falkenhainer et al., 1989), which performs graph-matching under a one-to-one structural consistency constraint.

Second, distributional semantics models find analogies via statistical representations, such as word embeddings, capturing co-occurrence information or heuristic distance metrics (e.g., distance in an ontology). Notably, word-embedding models (Mikolov et al., 2013) demonstrated that vector offsets can capture simple `A:B::C:D` relations. These methods (Turney et al., 2006; Mikolov et al., 2013) compute relational similarity using corpus statistics.

Third, cognitive/architectural models are inspired by human cognition. This includes Copycat-style architecture and LISA which simulate emergent binding or spreading activation to generate analogies from sub-symbolic processes (Hofstadter and Mitchell, 1994; Hummel and Holyoak, 2019b). These systems often hybridize symbolic and connectionist ideas and emphasize emergent, context-sensitive mapping.

Fourth, transformation-based models build analogies by vector-based or character-based operations. In this framework, analogies are interpreted as geometric relationships, and they use various distance metrics, normally in high-dimensional space, and vector operations to evaluate the analogies (Lepage and Ando, 1996; Plate, 2000).

We identified 17 papers that fit into such a taxonomy (See Table.3). We report six papers that use the relational graph-based method, seven papers that take the distributional semantic approach; we also found two papers that use the cognitive/architectural model and two papers that use the transformation-based model. Our result shows that early work relies on rule-based and graph-based methods, such as the relational graph-based model and cognitive/architectural model; as ML advances, a learning-based method, which requires a corpus and data to train, becomes relevant and adapted, such as the distributional semantic model.

4.2.3 Quality Dimension

Lastly, we categorized the reported evaluation metrics into: accuracy, validity, similarity, novelty, and human-preference/judgement metrics (See Table.5). Many automatic evaluation methods target one or more of these dimensions. Specifically, we identified 11 papers that use accuracy-based metrics, such as precision, recall, and F1-score; 13

papers that assess validity, including logical consistency checks and human validity judgments; and 19 papers that apply similarity-based metrics, such as BLEU, to evaluate generated analogies.

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/judgement is used in 13 papers as a primary evaluation method. These works typically employ crowdsourcers (Jiayang et al., 2023; Sultan et al., 2024) or experts (CAO et al., 2024; Shao et al., 2025) to evaluate or validate the analogy generated in their work.

4.3 Analogy Generation Challenge

To address RQ3, we identify relevant papers that discuss challenges and limitations encountered during analogy generation or evaluation. Our findings are summarized below.

4.3.1 Novel Analogy Generation

Across the literature, there is broad consensus that generating novel analogies remains a significant challenge. Many canonical analogies, such as the comparison between the solar system and Rutherford's atom model (Gentner, 1983), have historically been crafted by humans, 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). For example, 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., 2023a).

4.3.2 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., 2023a,b; 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).

4.4 Analogy Evaluation Challenge

4.4.1 Metrics and Dataset Limitation

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

A central concern is the narrow scope of existing benchmarks. Commonly used datasets, such as the SAT word analogy dataset (Turney, 2008) and the Google analogy corpus (Mikolov et al., 2013), are relatively small and restricted to lexical-level analogies. This narrow coverage limits the range of analogical phenomena 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). 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).

4.4.2 Evaluation with LLM

Throughout our report, we found that human evaluation is mainly conducted on the compositional-level, while automatic evaluation is primarily performed on the lexical-level.

This mismatch poses two challenges. First, LLMs sometimes generate outputs that are literally similar but relationally shallow, which can mislead both human and automatic evaluations (Jiayang et al., 2023). Second, existing automatic metrics, especially those designed for word analogy tasks, struggle to evaluate analogies beyond the lexical or syntactic level (Yuan et al., 2023a). Moreover, few existing automatic metrics account for creativity, novelty, or contextual coherence, all of which are central to human analogical reasoning (Hofstadter and Sander, 2013).

5 Discussions

5.1 Challenges and Opportunities in Analogy Generation

As computational analogy generation finds broader applications across domains, 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). While techniques such as multi-step prompting and chain-of-thought reasoning (Wei et al., 2023) have shown early success, they currently lack a standardized framework tailored to analogy tasks. Advancing in this direction could enable AI systems to perform more human-like, structured analogical reasoning.

Another major challenge is the generation of novel analogies (Bhavya et al., 2022, 2024a). One open challenge lies in aligning computational definitions of novelty (e.g., dissimilarity to known examples or training corpora) with human judgments, which rely on prior knowledge, domain experience, and perceived insight. Bridging this gap may require hybrid approaches that combine retrieval-augmented generation (RAG), knowledge-aware prompting, and structure-constrained decoding to guide the model beyond surface-level similarity.

In addition, future work should explore cross-modal analogy generation, expanding current methods beyond text to include visual, auditory, or symbolic modalities. This multimodal approach could unlock new forms of analogical reasoning, particularly in domains such as visual analogy tasks, educational simulations, and conceptual design.

5.2 Future Directions in Analogy Evaluation

To achieve a more robust and standardized evaluation, future work should also focus on developing large-scale, domain-diverse analogy datasets that include metaphorical, visual, and scientific reasoning examples. Evaluation protocols should incorporate automatic metrics and human-centered assessments (e.g., analogical relevance, novelty, and coherence), as suggested in (Bhavya et al., 2024a). Ultimately, analogy evaluation needs to move beyond lexical correctness to capture the full depth and function of analogical reasoning.

6 Conclusion

In this paper, we present a systematic literature review on analogy generation and evaluation, following the PRISMA framework. We began with a keyword search across six academic databases, yielding 4,641 papers. After applying the ATK screening process and conducting a full-text eligibility review, a total of 45 papers were included in the final analysis. We conducted a detailed analysis of these works, categorizing existing computational methods for analogy generation and evaluation. Additionally, we highlighted key challenges in the field and outlined future research directions to advance analogy generation and evaluation systems.

Limitations

As a review process that involves subjective judgments by individual authors, this study may be subject to bias. To mitigate this, we conducted pilot checks at each stage to iteratively refine the selection criteria. Additionally, all results were cross-checked by multiple authors, and meetings were held to resolve conflicts and ensure consistency.

Despite careful formulation of the search strings, it is possible that some relevant papers were 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. To address this, we conducted one round of backward snowballing to identify and include potentially overlooked studies. Future work could undertake a more exhaustive investigation into the distinctions and interrelations between analogy and metaphor.

References

Paul Bartha. 2024. Analogy and Analogical Reasoning. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

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. *arXiv preprint arXiv:2210.04186*.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. Cam: A large language model-based creative analogy mining framework (www'23). association for computing machinery, new york, ny, usa, 12 pages.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2024. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. *arXiv preprint arXiv:2412.15375*.

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. *arXiv preprint arXiv:2203.08480*.

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

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

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

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

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

Leonidas AA Doumas, John E Hummel, and Catherine M Sandhofer. 2008. A theory of the discovery and predication of relational concepts. *Psychological review*, 115(1):1.

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.

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

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

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

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

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

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.

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

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

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

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

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*.

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.

Janet Kolodner. 2014. *Case-based reasoning*. Morgan Kaufmann.

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*.

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.

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.

Andrew Lovett and Kenneth Forbus. 2017. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60.

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.

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

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

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

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

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

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? *arXiv preprint arXiv:2502.16895*.

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

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. Parallelparc: A scalable pipeline for generating natural-language analogies. *arXiv preprint arXiv:2403.01139*.

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, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.

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

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

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

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023a. Analogical–a novel benchmark for long text analogy evaluation in large language models. *arXiv preprint arXiv:2305.05050*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023b. Analogical–a novel benchmark for long text analogy evaluation in large language models. *arXiv preprint arXiv:2305.05050*.

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

Patrick H. Winston. 1980b. Learning and reasoning by analogy. *Commun. 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.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023a. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. *arXiv preprint arXiv:2305.12660*.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023b. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*.

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

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

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

A Disclosure of AI Assitant Use

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

B Complete Result

Generation Granularity	Papers
Lexical-level	(Sultan et al., 2024), (Yuan et al., 2023b), (Bourrelly et al., 1983), (Ding et al., 2023), (Bhavya et al., 2023), (Wang et al., 2024), (Chen et al., 2022), (Jiayang et al., 2023), (Bhavya et al., 2022), (Salu, 1994), (Yuan et al., 2023a), (Crouse et al., 2021), (Shao et al., 2025), (Ju et al., 2025), (Boisson et al., 2024)
Compositional-level	(Bhavya et al., 2024b), (Mittal, 1992), (Li et al., 2005), (Wang et al., 2024)

Table 1: Analogy Generation by Granularity

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

Table 2: LLM vs. Non-LLM Generation

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

Table 3: Evaluation Model Types

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

Table 4: Analogy Evaluation by Granularity

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

Table 5: Analogy Evaluation Dimension