The Structure of Analogy Matters: Progressive Analogy Generation for Better LLM Reasoning

Anonymous ACL submission

Abstract

Analogical reasoning enables knowledge transfer by mapping relational structures between familiar and novel domains. This work investigates whether large language models (LLMs) can generate and benefit from structured analogies that improve scientific reasoning. We propose Progressive Analogy Generation (PAG), a cognitively inspired approach that constructs analogies as stepwise, coherent narratives, and pair it with an automated LLM-as-Judge framework to assess analogy quality and revealness. Experiments on the GPQA dataset show that PAG substantially improves reasoning performance over single-concept analogies and non-analogical hints, especially in smaller models, while producing clearer and more pedagogically effective explanations without increasing answer leakage. These results demonstrate that the structure of an analogy is crucial for enabling LLMs to reason more effectively. We will provide our code and prompts upon publication.

1 Introduction

Analogy is a core mechanism of human cognition: by projecting relational structure from a source domain (familiar situation) onto a target domain (novel problem), people transfer knowledge, form explanations, and learn abstract concepts (Gentner, 1983; Holyoak and Thagard, 1995; Gentner and Markman, 1997). Such relational mapping, rather than surface similarity, underpins effective explanation in STEM education and scientific reasoning.

Large language models (LLMs) can both generate analogies and benefit from them as scaffolds for problem-solving. Recent work, such as SCUA (Yuan et al., 2024), shows that analogical explanations boost performance by having a "teacher" model generate an analogy for a "student" model. However, SCUA relies on single-concept analogies, whereas many scientific problems involve interdependent concepts. Simply extending analogies into longer text blocks risks overloading models or diluting key relations. Empirically, reasoning quality declines with longer, less structured inputs (Ye et al., 2024; Liu et al., 2024), suggesting that how an analogy is structured matters as much as whether one is provided. Model size may further moderate these effects: larger models possess stronger abstraction and integration abilities, while smaller ones may depend more on external scaffolds.

We further argue that progress in analogy-assisted reasoning requires measuring analogy quality, not just downstream accuracy. Without explicit criteria, improvements may stem from incidental cues or subtle answer leakage. Drawing on cognitive theory (Gentner, 1983; Holyoak and Thagard, 1995; Gentner and Markman, 1997), a high-quality analogy should preserve deep relational structure, exhibit systematicity and coherence, and offer clear, plausible, and explanatory support. However, prior LLM studies typically report performance gains without operationalizing these dimensions or auditing revealness (Li et al., 2024; Ye et al., 2024), underscoring the need for a principled, theory-grounded evaluation of analogy quality.

To address these gaps, we first introduce **Progressive Analogy Generation (PAG)**, which structures analogies as stepwise expansions of a coherent source-domain narrative. Rather than isolating a single concept, PAG incrementally builds relational connections as new ideas are introduced, mirroring the layered nature of scientific reasoning while maintaining clarity and coherence. We compare PAG against single-concept analogies and non-analogical hints by evaluating the performance of several student models (i.e., from the Llama family) when analogies are given by a teacher model (i.e., from the Gemini and Mixtral families).

To interpret the results more rigorously, we complement effectiveness testing with an automated LLM-as-Judge evaluation that quantifies two control factors: (i) revealness, to ensure analogies do

not disclose the correct answer, and (ii) quality, capturing relational mapping, coherence, and explanatory strength. This integrated framework links analogy generation, evaluation, and reasoning performance, enabling controlled analysis of how analogy structure shapes LLM behavior.

Our main contributions are: (i) We introduce PAG, a cognitively inspired method that constructs analogies as stepwise, coherent narratives. (ii) We propose a principled LLM-as-Judge evaluation that measures analogy quality and revealness, grounding model assessment in cognitive science theory. (iii) We provide comprehensive empirical evidence showing that PAG substantially improves scientific reasoning accuracy in GPQA benchmark (Rein et al., 2023) compared to related baselines—especially in smaller models.

2 Related Work

Analogical reasoning is a central mechanism in human cognition, enabling abstraction, learning, and creativity (Gentner, 1983; Holyoak and Thagard, 1995). The Structure-Mapping Theory (Gentner, 1983; Gentner and Markman, 1997) emphasizes that effective analogies rely on deep relational correspondences rather than surface similarity. These cognitive principles—relational mapping, systematicity, and explanatory coherence—provide the foundation for evaluating analogical reasoning in both humans and LLMs, and directly inform our framework for analogy assessment.

Recent studies have examined whether LLMs exhibit similar analogical abilities. Benchmarks such as AnaloBench (Ye et al., 2024) and AnalogyKB (Dong et al., 2023) reveal partial competence but also fragility when reasoning over extended or abstract analogies (Chen et al., 2023). Broader evidence further shows that reasoning quality depends on context organization and input structure, with phenomena such as the "Lost in the Middle" effect (Liu et al., 2024) highlighting the limits of unstructured prompts. Building on insights from cognitive scaffolding (Gentner and Smith, 2012), our Progressive Analogy Generation (PAG) framework structures analogies into coherent steps, and our automated LLM-as-Judge evaluation operationalizes cognitive dimensions of analogy quality—relational mapping, clarity, and explanatory power—while controlling for answer leakage.

3 Methodology

Our framework (Figure 1) investigates how LLMs generate, assess, and learn from analogies in scientific and reasoning tasks. It is composed of three main stages: 1) Progressive analogy generation (PAG), 2) LLM-as-Judge quality evaluation and 3) Effectiveness evaluation.

3.1 Progressive Analogy Generation (PAG)

PAG enables knowledge transfer by structuring analogies as evolving sequences of conceptual mappings. Given a reasoning problem, a teacher model identifies core concepts, orders them from basic to complex, and generates a multi-step analogy that incrementally extends a coherent source-domain narrative without revealing task-specific terms or answers. This progressive structure helps models build understanding through guided conceptual expansion rather than a single abstract leap (see Figure 2). We compare PAG with two baselines: (1) Single Analogy Generation (SAG) (Yuan et al., 2024), which produces one isolated analogy, and (2) Hint, a short non-analogical cue. Not all generated analogies are pedagogically sound—some may leak answers or lack coherence. To evaluate them systematically, we employ an LLM-as-Judge framework with two evaluators: a revealness judge and a quality judge. Detailed generation and evaluation prompts are provided in Appendix A.

Revealness Judge . This evaluator scores how much an analogy discloses the final answer on a 0–4 scale: 0—pure metaphor; 1—mixed source/-target language; 2—explanatory but non-revealing; 3—implicitly reveals answer; 4—explicitly states it. Analogies rated 3 or 4 are automatically regenerated, enforcing a non-revealing threshold (≤ 2). This filtering ensures analogies guide reasoning without leaking solutions.The details of the regeneration procedure used for revealness evaluation are provided in the Appendix D.

Quality Judge Analogy quality is assessed along six cognitively grounded dimensions (Gentner, 1983; Holyoak and Thagard, 1995; Gentner and Markman, 1997): (1) relational mapping, (2) systematicity, (3) clarity, (4) creativity, (5) plausibility, and (6) explanatory power. An LLM-as-Judge prompt scores each dimension from 0–2 (poor, moderate, strong), yielding a total score (0–12) and a normalized quality score in [0, 1]. The judge also provides brief justifications for each rating, support-

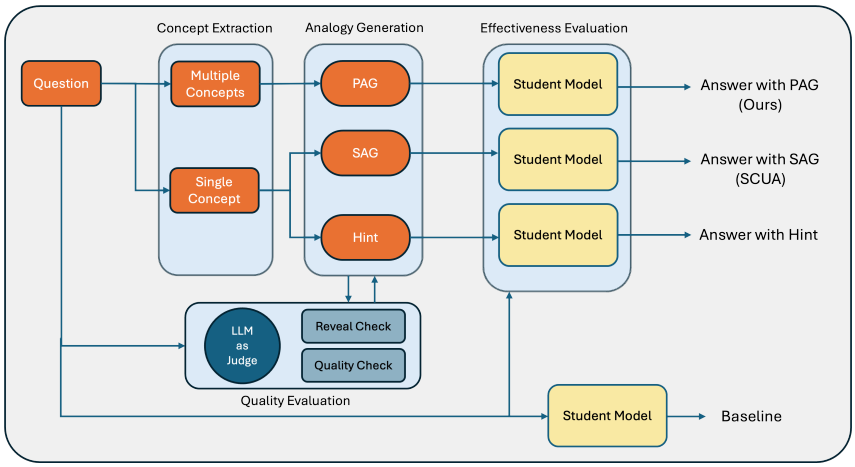


Figure 1: Teacher models generate analogies or hints for scientific reasoning questions, which are automatically evaluated by LLM-as-Judge modules for both revealness and quality. Analogies that expose the correct answer are regenerated to enforce non-revealing behavior, while all analogies receive quality scores for later analysis. These outputs are then given to student models to solve the same questions, enabling a unified evaluation of analogy quality, fairness, and pedagogical effectiveness across model families and scales.

Student Model	Baseline	HINT	SAG	PAG
Teacher: Gemini 2.5 Flash				
LLaMA-3-8B	42.63	50.00	58.04	**71.21**
LLaMA-3.1-8B	50.67	51.34	54.46	**68.08**
LLaMA-3.1-70B	39.51	43.97	44.87	**50.45**
LLaMA-3.1-405B	42.19	39.29	38.62	**44.20**
Teacher: Mixtral 8×7B Inst.				
LLaMA-3-8B	43.30	52.01	54.91	**67.86**
LLaMA-3.1-8B	49.33	51.34	53.79	**66.07**
LLaMA-3.1-70B	41.07	43.08	47.10	**50.67**
LLaMA-3.1-405B	41.74	36.83	39.29	**43.30**

Table 1: Effectiveness accuracies for the LLaMA family under two teacher settings. Baseline refers to direct question answering without guidance. HINT provides non-analogical textual hints. SAG uses a single, concept-level analogy. PAG is our method for multi-concept analogies in stepwise form. Best per row in **bold**.

ing later qualitative analysis of analogy strengths and weaknesses.

3.2 Effectiveness Evaluation

The generated analogies are used to assess how analogical scaffolding influences reasoning performance. In this stage, teacher models produce analogies while student models of varying sizes attempt to solve problems under four distinct conditions: (i) **Baseline:** problem presented without any additional context (Zero-Shot); (ii) **HINT:** problem accompanied by a short, non-analogical textual hint; (iii) **SAG:** problem accompanied by a single-concept analogy; (iv) **PAG:** problem accompanied by a multi-step, progressively structured analogy.

4 Experimental Setup and Results

Dataset We evaluate on the GPQA benchmark (Rein et al., 2023) ("Main" subset), a set of 448 graduate-level multiple-choice questions testing deep reasoning in biology, chemistry, and physics. Each question includes four options and requires multi-step conceptual reasoning. All items were reformatted into a unified prompt template. For each question, teacher models produced three conditions, Hint, Single Analogy (SAG), and Progressive Analogy (PAG), each automatically verified for non-revealness before evaluation.

Models Our framework uses three model roles: teacher, student, and LLM-as-Judge (see Figure 1). Teacher models **Mixtral 8×7B Instruct** and **Gemini 2.5 Flash-Lite** generate analogies and hints with `temperature = 0` for reproducibility. Student models solve reasoning tasks under different analogy conditions. To isolate size effects, we use the LLaMA family (`LLaMA-3-8B`, `LLaMA-3.1-8B`, `LLaMA-3.1-70B`, and `LLaMA-3.1-405B`). all tested with identical decoding settings (`temperature = 0`). Finally, the LLM-as-Judge (**GPT-5-Nano**) scores analogies for revealness and quality as described above. All model inferences were conducted via the OpenRouter API under the official provider terms of service.

4.1 Effectiveness Results

Table 1 reports reasoning accuracies for the LLaMA family under two teacher models, Gemini

2.5 Flash-Lite and Mixtral 8×7B Instruct. Across all configurations, Progressive Analogy Generation (PAG) consistently outperforms both SAG and Hint, with accuracy gains of up to +28.6 points over the baseline.

While both teachers enable substantial improvements, Gemini produces slightly higher accuracies and more stable gains across all student scales. For instance, with LLaMA-3-8B as a student, accuracy increases from 42.63 to 71.21 (+28.6) under Gemini compared to 43.30 to 67.86 (+24.6) under Mixtral. A similar trend appears for the 8B and 70B variants (+17.4 vs. +16.7 and +10.9 vs. +9.6), indicating that Gemini's analogies provide clearer conceptual scaffolding. This consistency across teacher architectures indicates that PAG's benefits generalize well, with Gemini emerging as the stronger generator overall.

Performance gains are largest for smaller students and decline with model size (Figure 3). LLaMA-3-8B shows the strongest boost (+28.6), decreasing to +2.0 for 405B. This inverse trend suggests smaller models rely more on external relational structure to compensate for limited abstraction capacity. Notably, under the PAG condition, smaller models even surpass larger ones in absolute accuracy (e.g., 71.2% for 8B vs. 44.2% for 405B), showing that structured analogies can, in some cases, elevate lightweight models beyond much larger counterparts.

To ensure comparability with prior work (e.g., SCUA on GPQA), we replicate the Mixtral–LLaMA-3-8B setting. Accuracy improves from 43.30% to 67.86% with PAG, versus 54.91% with SAG, confirming a clear advantage for progressively structured analogies.

4.2 Quality and Revealness Results

We next assess whether progressively generated analogies are both high-quality and pedagogically sound. Automated judging is conducted on 448 analogies, complemented by human evaluation on a balanced 60-analogy subset. Table 2 (below) reports the aggregated average revealness and quality scores across all subjects (weighted by sample count). The Appendix E also describes the sampling strategy used for selecting the evaluation subset.

PAG achieves higher overall quality (0.91 ± 0.11) than SAG (0.84 ± 0.14), while maintaining similarly low revealness (1.07 ± 0.56 vs. 0.90 ± 0.63). Hints are substantially more revealing (1.92 ± 0.32), confirming that the analogy generation process effectively constrained answer exposure. Quality and revealness are weakly correlated ($|r| < 0.3$, $p > 0.05$), suggesting that quality improvements stemmed from stronger relational mapping rather than leakage.

Type	Reveal Level (↓)	Quality (↑)
PAG Analogy	1.07 ± 0.56	0.91 ± 0.11
SAG Analogy	0.90 ± 0.63	0.84 ± 0.14
Hint Text	1.92 ± 0.32	–

Table 2: Aggregated average revealness and quality scores across all subjects. PAG achieves higher quality than SAG while maintaining low revealness.

Human–LLM Agreement. Human and automated quality scores showed moderately strong correlations (Pearson's r = 0.58–0.70; Spearman's ρ = 0.55–0.67; $p < 0.01$). For revealness, alignment was even higher (r = 0.72; ρ = 0.70; $p < 0.001$)), showing that the LLM-as-Judge reliably mirrors human judgment in detecting answer leakage. These findings confirm that our evaluation framework provides a consistent, scalable, and low-cost proxy for human assessment of analogy quality and pedagogical soundness.

5 Conclusion

This work introduced **Progressive Analogy Generation (PAG)**, a cognitively inspired framework for producing structured, multi-step analogies that enhance scientific reasoning in LLMs. Experiments on **GPQA** show that progressively constructed analogies yield substantial accuracy gains over single-concept analogies and non-analogical hints, with the strongest effects observed in smaller student models. These results indicate that structured analogical scaffolding can compensate for limited internal abstraction, enabling lightweight models to reach or even surpass the performance of much larger ones. Beyond effectiveness, our **LLM-as-Judge** evaluation provides a scalable, interpretable method for assessing analogy quality and revealness. Progressive analogies achieved higher structural and explanatory quality without increasing answer leakage, and human–LLM agreement analyses confirmed the reliability of these automated judgments.

Limitations

This study has several limitations. First, we did not employ very large or frontier teacher models due to high computational cost, which may limit the potential ceiling of analogy quality. Second, our experiments were restricted to the LLaMA family as student models, without testing other architectures or instruction-tuned variants. Third, the evaluation dataset was not perfectly balanced across scientific domains, as some topics contained more questions than others, which may have introduced domain-level bias in the results. Fourth, we did not examine how sensitive the revealness and quality judgments are to the choice of judge model or decoding parameters, and cross-judge agreement was not systematically evaluated.

References

Xi Chen, Qiang Liu, and Yue Zhang. 2023. Storyanalogy: Evaluating analogical reasoning in narrative contexts. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Zhexin Dong, Yitong Xu, Siqi Shen, Yankai Lin, and Maosong Sun. 2023. Analogykb: A large-scale knowledge base of analogical relations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

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

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

Dedre Gentner and Linda Smith. 2012. Analogical reasoning. In Vilayanur S. Ramachandran, editor, *Encyclopedia of Human Behavior (2nd ed.)*, pages 130–136. Academic Press.

Keith J. Holyoak and Paul Thagard. 1995. *Mental Leaps: Analogy in Creative Thought*. MIT Press.

Xiang Li, Haoyu Sun, Wenqi Zhang, Kun Xu, and Minlie Huang. 2024. Boosting scientific concepts understanding: Can analogy from teacher models empower student models? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Nelson F. Liu, Yangyi Ruan, Danqi Chen, Omer Levy, Luke Zettlemoyer, and Samuel R. Bowman. 2024. Lost in the middle: How language models use long contexts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

David Rein, Yuxiao Chen, Nathan Lambert, Paul Christiano, and Sam Ringer. 2023. Gpqa: A benchmark for graduate-level scientific reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13952–13967, Singapore. Association for Computational Linguistics.

Zihan Ye, Haodi Zhang, Zihan Ding, Zhengxuan Du, Shuwen Zhi, and Zhou Yu. 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)*. Association for Computational Linguistics.

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

A Prompt Templates

Listing 1: Concept extraction prompt for Progressive Analogy Generation (PAG).

```
Given a scientific question, you should
extract the key scientific concepts
required to answer it.
Instead of a single concept, produce a
**list of concepts (concept_list)**
arranged in logical order:
- The first item should be the simplest
or most fundamental concept.
- Each following item should build upon
the previous one, leading step by step
to the final, more complex concept
needed to answer the question.
- There should be at most 10 concepts.

This is a scientific question:
{question}
Concepts:
```

Listing 2: Progressive Analogy Generation prompt.

```
You are given an ordered list of
scientific concepts.
Create a progressive sequence of
analogies where each step builds on the
previous ones:

- Begin with a simple analogy for the
first concept.
- For the i-th concept, extend the
analogy you have been building so far.
- Each new analogy must **reuse elements
 from the previous analogies** so the
story grows naturally.
- Do NOT reveal, directly name, or hint
at the actual scientific concepts in the
 analogy.
- Do NOT insert actual numbers,
equations, or the final computed answer.
- Analogies should describe the *idea or
 method*, not the final solution.
- Each analogy should be concise (  100
 tokens). DO NOT OVER PASS 100 TOKENS.

Concept list: {concepts}
Analogy:
```

Listing 3: Hint generation prompt (non-analogical baseline).

```
You are given a science question and an
associated concept.
Write a concise HINT (not an analogy)
that nudges how to think about solving
it.

Rules:
-     40 words.
- No numbers, equations, or final answer
.
- Focus on the method/idea to recall or
apply.
- Avoid revealing the full solution.
- DO NOT REVEAL THE ANSWER IN THE HINT.

concept:
{concept}
Question:
{question}
Hint:
```

Listing 4: Revealness evaluation prompt for LLM-as-Judge.

```
You are judging how a HINT reveals the
answer to a multiple-choice question.

Domains:
- Source domain = analogy/metaphor/story
 in a different everyday context.
- Target domain = the actual subject
domain of the question (e.g., physics,
math, biology).

Revealness scale:
0 = Entirely in source domain (pure
metaphor). No numbers, no option text,
no final answer.
1 = Mix of source + target domain. No
numbers, no option text, no final answer
.
2 = Entirely in target domain (
procedural/explanatory). No numbers, no
option text, no final answer.
3 = Gives the final answer (number or
option) but expressed in the source
domain (metaphor).
4 = Gives the final answer (number or
option) directly in the target domain.

Rules:
-   Final   answer  = any explicit
numeric value from the options, or an
explicit option letter/text.
- If no final answer is given, classify
as 0, 1, or 2 based on domain language.
- If a final answer is given, classify
as 3 (source domain) or 4 (target domain
).

INPUT
options: {options}
hint: {hint_text}

OUTPUT (strict JSON only):
{{
  "level": 0|1|2|3|4,
  "reason": "short explanation"
}}
```

Listing 5: Analogy quality evaluation prompt for LLM-as-Judge.

```
You are an expert judge of analogies.
Evaluate the following analogy according
 to the criteria below.
For each criterion, give a score and a
short justification.

Scoring guide:
- Each dimension is scored 0, 1, or 2.
- 0 = poor, 1 = moderate, 2 = strong.

Criteria:
```

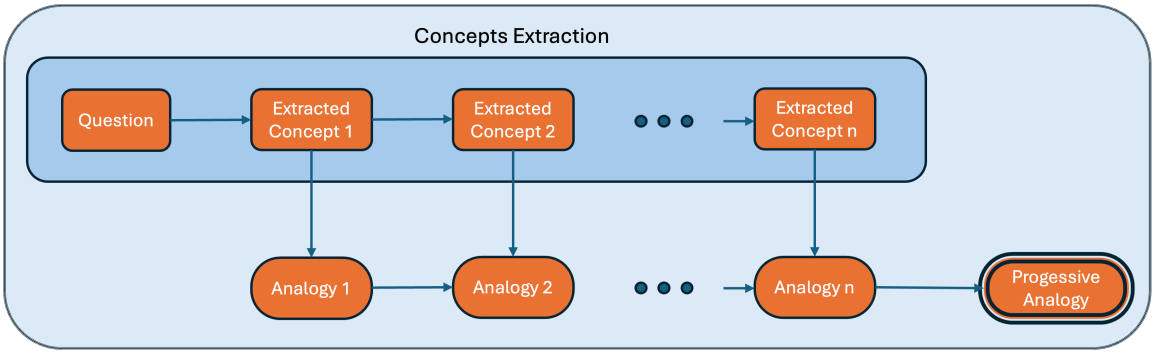


Figure 2: Progressive Analogy Generation (PAG). The teacher model extracts key concepts from a question and sequentially expands an analogy for each concept, forming a coherent multi-step narrative. This progressive process builds structured reasoning support without exposing the final answer.

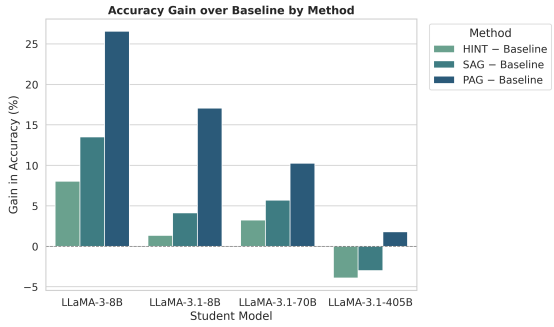


Figure 3: Accuracy gains (%) of Progressive Analogy Generation (PAG) over the baseline across LLaMA student models under two teacher settings. Smaller models achieve the largest improvements, highlighting the greater impact of analogical scaffolding on lower-capacity learners.

.

```
1. Relational Mapping Quality     Do
source and target share deep relational
structure?
2. Systematicity & Coherence     Are
correspondences part of a systematic,
connected structure?
3. Clarity & Interpretability     Is the
 analogy clear and understandable to
humans?
4. Creativity / Cross-Domain Distance
    Does it connect distinct domains
creatively?
5. Plausibility & Faithfulness     Are
the correspondences factually correct
and not misleading?
6. Explanatory Power / Usefulness
Does it help explain the target concept
or support reasoning?

INPUT
options: {options}
Analogy to evaluate: {analogy_text}

OUTPUT (strict JSON only):
{{
  "relational_mapping": {{"score": X, "
  justification": "..."}},
  "systematicity": {{"score": X, "
  justification": "..."}},
  "clarity": {{"score": X, "
  justification": "..."}},
  "creativity": {{"score": X, "
  justification": "..."}},
  "plausibility": {{"score": X, "
  justification": "..."}},
  "explanatory_power": {{"score": X, "
  justification": "..."}},
  "total_score": Y,
  "normalized_score": Z
}}
```

B Examples

B.1 Example of Progressive Analogy Generation (PAG)

Question: Three planets are orbiting the same star, each with orbital periods of 22, 40, and 64 days. They all share the same orbital plane and transit the star. What is the periodicity of a triple transit, meaning when all planets transit simultaneously?

Options: A) 2560 days B) 56320 days C) 28160 days D) 3520 days

Generated Progressive Analogy (PAG):
Concept: Orbital period. On a quiet hill, a lone lantern-walker traces a circle around a crackling campfire. Each full lap returns to the same glow, and we notice the quiet time that passes between visits to that glow.
Concept: Transit timing. As the lantern completes its loop, we watch for a telltale crossing: the beam slips past a fixed mark on a fence. The moment of that crossing becomes a finer clock inside the same rhythm.
Concept: Co-planar transits and alignment. A second lantern joins the stage, sharing the same plane of sight. When both beams strike the same

mark at the same moment, the crowd sees a single line of light crossing the stage.
Concept: Relative phasing of multiple orbits. Now their paths drift and slide relative to each other. The moments when both lines align wander along, like two dancers with different paces but occasionally meeting in rhythm.
Concept: Least common multiple (LCM). We wonder if there is a moment when both lanterns finish whole cycles and land on the same mark together again. Such a shared beat would repeat, offering a clue to predicting the next alignment.
Concept: Prime factorization. To chase that clue, we break each lantern's schedule into its smallest, repeating beats—the building blocks that never change. By listening to those tiny rhythms, we can see how they combine to form a bigger pattern.
Concept: LCM. With three lanterns, each keeping its own steady tempo, we seek the moment when all three cast light on the same mark at once. We search for a single moment where their three tides meet in one flash.
Concept: Triple transit periodicity. Finally, we imagine laying three clear sheets over the same stage, each showing its cycle as marks. We overlay them, then nudge until the shared marks align. From there, the method to find the triple cadence is clear, yet we do not reveal the exact time.

B.2 Example of a HINT

Question: Three planets are orbiting the same star, each with orbital periods of 22, 40, and 64 days. They all share the same orbital plane and transit the star. What is the periodicity of a triple transit, meaning when all planets transit simultaneously?

Options: A) 2560 days B) 56320 days C) 28160 days D) 3520 days

Generated Hint: Think of the three planets like repeating timers that all start together. The "triple transit" happens the next time all three timers ring at the same moment. What single interval would make that happen?

C Complete related work

Analogy and Cognitive Foundations Analogical reasoning is widely recognized as a core mechanism of human cognition, supporting abstraction, learning, and creativity (Gentner, 1983; Holyoak and Thagard, 1995). Cognitive theories such as the Structure-Mapping Theory (Gentner, 1983; Gentner and Markman, 1997) emphasize that effective analogies depend on deep relational correspondences rather than surface similarity. These dimensions, relational mapping, systematicity, and explanatory coherence, form the foundation for evaluating analogical reasoning in both humans and artificial systems. Our work draws directly on these theoretical principles to design cognitively grounded criteria for LLM-based analogy evaluation.

Analogical Reasoning in LLMs Recent research has begun exploring whether LLMs exhibit comparable analogical capabilities. Ye et al. (2024) introduced AnaloBench, a large-scale benchmark testing LLMs' ability to recognize and reason over analogies across varying lengths and abstraction levels, showing that performance degrades as contexts become longer or more abstract. Similarly, Chen et al. (2023) examined narrative analogies and found that even strong models struggle to maintain coherent relational mappings across extended stories. **?** proposed AnaDQ, a dataset for analogical question answering, while Dong et al. (2023) built AnalogyKB, a structured knowledge base of analogical relations for retrieval and reuse. Together, these studies highlight progress in measuring analogical understanding but leave open questions about how analogy structure, progression, and pedagogical utility shape reasoning outcomes, issues our work explicitly targets through progressive analogy generation.

Context, Structure, and Cognitive Load Broader reasoning research shows that model performance is highly sensitive to context organization and input length. The "Lost in the Middle" effect (Liu et al., 2024) revealed that LLMs tend to overlook key information in long sequences, suggesting that concise, well-structured analogical inputs may better support reasoning. From a cognitive perspective, progressive scaffolding has long been shown to facilitate understanding by reducing cognitive load and aligning with hierarchical knowledge representation (Gentner and Smith, 2012). Our progressive analogy generation (PAG) framework operationalizes these insights by structuring analogies into manageable, conceptually coherent steps.

Evaluating Analogy Quality While prior studies have reported qualitative or human-evaluated results for analogical reasoning (Ye et al., 2024),

few have attempted to systematically measure analogy quality. In contrast, we introduce a principled evaluation grounded in cognitive-science theory. Through an automated LLM-as-Judge framework, we operationalize six canonical dimensions of analogy quality, relational mapping, systematicity, clarity, creativity, plausibility, and explanatory power, and pair this with a quantitative revealness metric to detect answer leakage. This dual evaluation connects cognitive theory with scalable, model-based assessment, enabling fair comparison of analogy generation strategies.

D Revealness Regeneration Analysis

Figure 4a–4c illustrate the distribution of revealness scores for progressively generated analogies (PAG), single-step analogies (SAG), and hint texts across Biology, Physics, and Chemistry subjects. Most samples fall within the low-reveal range (scores 2), indicating that the analogy generation process effectively avoids direct answer disclosure. Only a small fraction of analogies exhibit strong revealness (scores 3–4): 1.79% for PAG, 2.68% for SAG, and 0.45% for hint texts. Overall, merely 4.91% of all samples contained any instance of reveal level 3 or 4.

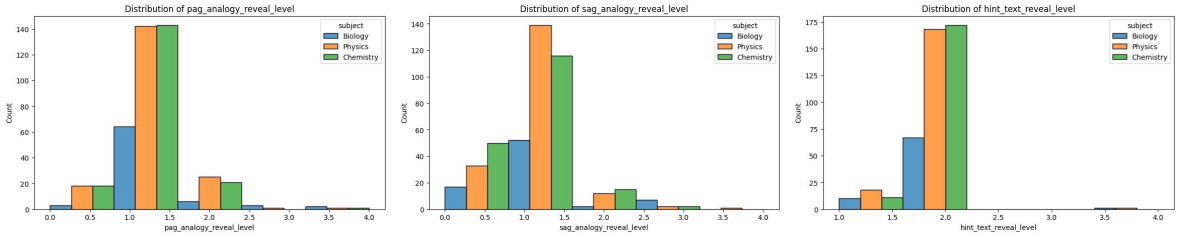
To further reduce revealness, these high-reveal cases were automatically regenerated. For each analogy identified as overly revealing by the LLM-as-Judge, we extracted its associated "reason" explanation. This reason was then prepended to the prompt used for regeneration, serving as an explicit instruction example for the model to avoid such leakage in the new output. All regeneration attempts successfully produced acceptable analogies within three iterations or fewer, ensuring that the corrected outputs maintained pedagogical abstraction while preserving semantic alignment.

E Human Evaluation Details

Sampling Procedure A balanced subset of 60 analogies was selected for human evaluation, stratified by subject domain (Biology, Chemistry, Physics) and generation type (PAG, SAG, Hint) to ensure representative coverage. For each category, examples were randomly drawn from the automatically judged pool, while maintaining proportional distribution across generator models. This ensured that both high- and low-quality analogies were included in the sample.

Annotation Setup and Blinding Three human annotators with graduate-level backgrounds in science and linguistics participated in the evaluation. Annotators were blinded to model identity and to whether an analogy was generated in the progressive (PAG) or single-step (SAG) condition. Each analogy was presented with its corresponding question and choices but without the correct answer, ensuring independent, unbiased judgments of revealness and quality.

Agreement and Validation Inter-annotator reliability was computed using Cohen's and Spearman's , yielding moderate-to-high agreement across annotators (= 0.67 for revealness, = 0.71 for quality). Human–LLM agreement was measured against the averaged human scores, showing consistent patterns with an absolute deviation of less than 0.2 on both metrics. Disagreements were resolved via discussion and re-evaluation in ambiguous cases.



(a) PAG (b) SAG (c) HINT

Figure 4: Distribution of revealness scores across disciplines for PAG, SAG, and Hint conditions. Most analogies remain within low revealness levels (2), indicating effective control of answer leakage.