UNIVERSE OF THOUGHTS: A COMPUTATIONAL FRAMEWORK FOR CREATIVE REASONING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in Large Language Model (LLM) reasoning have improved conventional problem solving, but creative reasoning remains comparatively underexplored. Inspired by cognitive science, we formalize combinational, exploratory, and transformational creativity as executable computational operators over structured problem and solution spaces, specifying how each mode combines, explores, or transforms those spaces, respectively. Combinational reasoning transfers ideas across domains to form unfamiliar combinations; exploratory reasoning searches for new solutions within an existing conceptual space; and transformational reasoning modifies the rules or constraints that define that space. This formalization yields distinct algorithmic procedures, which we directly instantiate in Universe of Thoughts (UoT), an LLM reasoning framework. Existing creativity benchmarks emphasize either open-ended ideation or highly constrained problem solving. We therefore introduce three novel creative-reasoning tasks requiring concrete solutions in low-constraint settings. Across 10 generations per method and task, T-UoT with GPT-4o performs strongest on the low-constraint, high-objective-specificity Bridge and Electricity tasks, while C-UoT shows its strongest relative performance on the low-constraint, lower-objective-specificity Society task. In addition, we evaluate UoT on HypoArena, an independent scientific hypothesis-generation benchmark with 100 tasks across biomedical, machine-learning, and social-science domains. With Qwen3-14B, Exploratory UoT ranks first among seven reasoning methods, achieving a 32.7% pairwise win rate compared with 25.5% for the next-best method. Our results suggest distinct performance patterns across task structures: T-UoT is strongest in low-constraint, high-specificity settings, E-UoT in more constrained, high-specificity settings, and C-UoT in low-constraint, lower-specificity settings.

1 INTRODUCTION

Large Language Models (LLMs) have made rapid progress on complex reasoning, supported by methods such as Chain-of-Thought (CoT) (Wei et al., 2022) and Tree of Thoughts (ToT) (Yao et al., 2023), which structure intermediate reasoning and search (Achiam et al., 2023; Chowdhery et al., 2023). These advances have improved performance on problems with well-defined objectives and verifiable solutions. Creative reasoning remains less explored, particularly for open-ended or ill-defined problems where success requires solutions that are not only valid but also novel and useful (Simon, 1973). Such settings include scientific hypothesis generation, drug discovery, and strategic decision making, where progress may require moving beyond familiar solution patterns (Singh et al., 2024; Gangwal et al., 2024).

Recent work has begun to move LLM reasoning toward open-ended scientific discovery. Agentic systems such as The AI Scientist (Lu et al., 2024) automate broad scientific workflows through iterative idea generation, experimentation, and refinement, while methods such as MOOSE-Chem (Yang et al., 2024) more explicitly structure hypothesis generation around retrieving inspirations and composing them into new scientific ideas. Related work has further connected LLM generation to computational creativity, particularly through analogical or combinational mechanisms that bring together knowledge from different contexts (Gu et al., 2024; Schapiro et al., 2025). These advances demonstrate that structured reasoning can support creative discovery. However, existing approaches

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

typically instantiate a particular discovery workflow or a single creative mechanism; they do not provide a unified computational account of the qualitatively different ways in which creative reasoning can search for new solutions.

We address this gap with *Universe of Thoughts* (UoT), a reasoning framework grounded in Margaret Boden's three modes of creativity: *combinational*, *exploratory*, and *transformational* (Boden, 2004; 2007; 2009). UoT operationalizes these modes as distinct search programs: combinational reasoning transfers useful structure from unfamiliar contexts, exploratory reasoning expands into previously unvisited possibilities while preserving the current problem rules, and transformational reasoning modifies the assumptions that define which solutions can be considered. This unified formulation allows the three modes to be instantiated within the same LLM reasoning framework and, importantly, directly compared to characterize when different forms of creative search succeed or fail.

Evaluating creative reasoning remains difficult because existing benchmarks often fall into two extremes. Open-ended idea-generation tasks permit novelty but make utility difficult to assess, while grounded settings such as scientific discovery provide clearer evaluation but impose stronger constraints that narrow the solution space. We therefore characterize tasks by degree of constraint and objective specificity. We introduce three low-constraint problem-solving tasks that allow both novelty and utility to be evaluated, and additionally use HypoArena (Zhong et al., 2026), which permits multiple plausible hypotheses and pairwise comparison rather than requiring reproduction of a single reference hypothesis. Together, these settings let us study how creative search behavior changes with task structure.

The results reveal distinct performance patterns across task regimes. On the novel creative-reasoning tasks, T-UoT performs strongest on the low-constraint, high-objective-specificity Bridge and Electricity tasks, where flexible solution spaces permit viable rule transformations and clear objectives support their evaluation. C-UoT shows its strongest relative performance on Society, where lower objective specificity broadens analogue retrieval and low constraint facilitates cross-domain mechanism transfer. On HypoArena, E-UoT performs best overall with Qwen3-14B, particularly in the more constrained, higher-specificity biomedical and machine-learning settings, where local search is more strongly guided by the problem context. Our failure analysis localizes an additional bottleneck beyond creative search itself: C-UoT often identifies useful analogies but fails to preserve their mechanisms during final synthesis, while T-UoT can identify meaningful rule transformations whose consequences are not fully carried into the final hypothesis. Targeted interventions that preserve the transferred mechanism or explicitly reason through the consequences of transformed rules substantially improve these failure cases, suggesting that stronger downstream reasoning and integration could allow these creative search processes to realize their potential more reliably.

We summarize our main contributions as follows:

- We formalize and operationalize three cognitive-science creativity modes (combinational, exploratory, and transformational) as executable LLM search procedures in the Universe of Thoughts (UoT) framework.
- We evaluate UoT on novel creative-reasoning tasks in low-constraint settings and on 100 scientific hypothesis-generation tasks from HypoArena.
- We identify task-dependent strengths across creativity modes: transformational reasoning is strongest in low-constraint, high-objective-specificity settings, exploratory reasoning performs best in more constrained, higher-specificity settings, and combinational reasoning shows its strongest relative performance in low-constraint, lower-specificity settings. Ablation and failure analyses further reveal distinct contributions and bottlenecks across the three modes.

2 Related Work

2.1 LLM Reasoning Frameworks

LLM reasoning methods increasingly structure search over intermediate thoughts. Chain-of-Thought (CoT) elicits sequential reasoning (Wei et al., 2022), Tree of Thoughts (ToT) introduces branching search with evaluation and backtracking (Yao et al., 2023), and Graph of Thoughts (GoT) generalizes this process to arbitrary graph structures (Besta et al., 2024). More recent methods enrich these search

procedures through refinement, task-specific reasoning structures, self-reflection, and environment feedback, including Self-Discover (Zhou et al., 2024) and Language Agent Tree Search (Zhou et al., 2023), as well as related refinement-oriented approaches (Shin & Kim, 2025). Although such frameworks can support creative outputs, their mechanisms are primarily designed for generic reasoning, planning, and search rather than explicitly modeling distinct forms of creative reasoning. UoT instead operationalizes creativity-specific search procedures derived from cognitive science.

2.2 Creative Reasoning in LLMs

Prior work has explored individual mechanisms related to creative reasoning, including analogical reasoning (Yu et al., 2023; Yasunaga et al., 2023) and inductive reasoning (Cheng et al., 2024; Wang et al., 2023), as well as agentic systems for scientific discovery such as MOOSE-Chem, The AI Scientist, and AlphaEvolve (Yang et al., 2024; Lu et al., 2024; Novikov et al., 2025). These systems show that iterative retrieval, recombination, refinement, and search can improve candidate solutions and sometimes produce novel outputs. However, they generally do not explicitly specify creativity-oriented search mechanisms for deliberately reaching unfamiliar solutions or changing the space of possibilities being explored.

More recent work explicitly connects LLMs with computational creativity, particularly combinational creativity (Schapiro et al., 2025; Gu et al., 2024). In contrast, UoT provides a unified operationalization of combinational, exploratory, and transformational creativity as distinct search regimes: transferring structure across contexts, exploring new solutions within a fixed space, and modifying the rules that define the space itself. This common formulation enables direct comparison of the three modes and analysis of their distinct strengths and failure modes.

2.3 Benchmarking Creativity in LLMs

Creativity benchmarks span a spectrum from open-ended generation to more constrained problem solving. Open-ended evaluations cover divergent thinking (Hou et al., 2025), story generation (Li et al., 2025), scientific idea generation (Ruan et al., 2026; Qiu et al., 2025; Guo et al., 2025), and creative writing (Sen et al., 2026). These settings allow substantial novelty and diversity, but functional effectiveness can be difficult to assess objectively.

More constrained benchmarks provide clearer evaluation criteria, including MacGyver-style physical problem solving (Tian et al., 2024) and scientific-discovery tasks involving law discovery (Zheng et al., 2026), multimodal discovery (Liu et al., 2026a; Panigrahi et al., 2026), AI model discovery (Wu et al., 2026), end-to-end research (Zhang et al., 2026), and scientific hypothesis generation (Liu et al., 2026b; Lew et al., 2026; Liu et al., 2025; Chen et al., 2025; Zhong et al., 2026). Their stronger grounding makes utility easier to evaluate, but also narrows the admissible solution space.

This leaves an important evaluation gap. Existing benchmarks often either support open-ended idea generation, where utility is difficult to assess concretely, or impose stronger constraints that narrow the admissible solution space. We therefore introduce three low-constraint creative-reasoning tasks that allow both novelty and utility to be evaluated while preserving multiple valid solution strategies.

Because our primary contribution is the reasoning framework rather than benchmark construction, we also evaluate UoT on HypoArena (Zhong et al., 2026), a scientific hypothesis-generation benchmark that permits multiple plausible hypotheses and pairwise comparison rather than requiring reproduction of a single reference hypothesis.

3 Universe of Thoughts: Creative Reasoning as Structured Search

Margaret Boden distinguishes three forms of human creativity: *combinational*, *exploratory*, and *transformational* (Boden, 2004; 2007; 2009). We operationalize these modes as distinct computational search regimes that differ in *what the reasoning process changes* (Figure 1). We follow Boden's original characterizations of the three modes; the source definitions and their correspondence to our computational operators are provided in Appendix A.

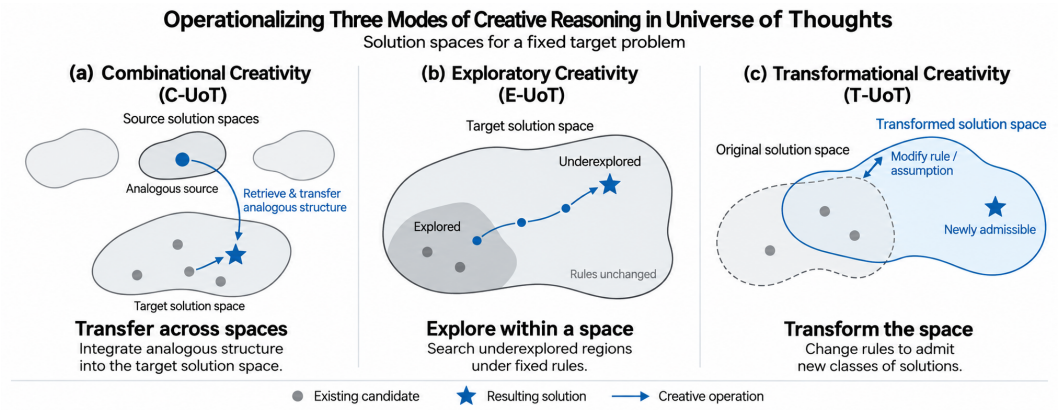


Figure 1: *Three computational regimes of creative search. Combinational reasoning imports transferable structure from a different context; exploratory reasoning expands from existing candidates while preserving the current rules; and transformational reasoning modifies the rules or assumptions that define the admissible solution space.*

Let P denote a target problem, H immutable task requirements, and R mutable rules or assumptions. The admissible solution space is

$$\mathcal{S}(P, H, R) = \{s \mid s \text{ addresses } P \text{ while satisfying } H \text{ and } R\},$$

and $\mathcal{K}_t \subseteq \mathcal{S}(P, H, R)$ denotes the solutions already explored. Under this formulation, combinational reasoning changes the source structure used to construct a solution, exploratory reasoning changes the candidate solution while keeping R fixed, and transformational reasoning changes R itself.

3.1 Combinational UoT

Boden characterizes combinational creativity as producing *unfamiliar combinations of familiar ideas* by forming associations between ideas that were previously only indirectly linked. We interpret this computationally as transferring a useful structural mechanism from a context that is functionally relevant to the target but sufficiently unfamiliar to introduce a new combination.

Combinational UoT restructures direct generation through two intermediate variables, an external analogue a and an abstract mechanism m:

$$\underbrace{p_\theta(s \mid P)}_{\text{direct search}} \quad \Longrightarrow \quad \underbrace{p_\theta(a \mid P)\, p_\theta(m \mid a)\, p_\theta(s \mid P, m)}_{\text{combinational search}}.$$

The product denotes a sequence of conditional generation stages rather than an exact factorization of $p_\theta(s \mid P)$.

The analogue is therefore selected to be both *functionally relevant* to the target and *contextually unfamiliar* relative to previously explored solutions. This differs from Analogical Prompting (Yasunaga et al., 2023), which primarily self-generates relevant examples or knowledge to improve downstream reasoning. C-UoT differs both in *objective* and *structure*: it deliberately searches for analogues from unfamiliar contexts to promote creative recombination, and explicitly decomposes analogical reasoning into *analogue* $\rightarrow$ *mechanism abstraction* $\rightarrow$ *target transfer*.

Operationally, C-UoT proposes distant analogues, solves the source problem, abstracts the transferable mechanism, maps that mechanism into the target context, and evaluates the resulting candidate. This explicit decomposition also makes the reasoning process inspectable, allowing failures to be localized to analogue retrieval, mechanism abstraction, or downstream transfer.

3.2 Exploratory UoT

Boden characterizes exploratory creativity as searching within an existing conceptual space to discover previously unvisited possibilities without changing the rules that define that space. We

interpret this computationally as expanding the set of candidate solutions while preserving the current problem formulation and mutable rules.

Exploratory UoT structures search around an already explored solution $s_k \in \mathcal{K}_t$ while preserving the current rules:

$$\underbrace{p_\theta(s \mid P)}_{\text{direct search}} \quad \Longrightarrow \quad \underbrace{p_\theta(s_k \mid P, \mathcal{K}_t)\, p_\theta(s' \mid P, s_k, R)}_{\text{exploratory search}}, \qquad R \text{ fixed.}$$

The product denotes a sequence of conditional generation stages rather than an exact factorization of $p_\theta(s \mid P)$.

Directly searching for novel solutions in a large conceptual space is difficult. E-UoT therefore operationalizes exploration locally: rather than generating an entirely new solution from scratch, it starts from an existing candidate and searches for neighboring alternatives that have not yet been explored. Repeating this process expands the frontier of $\mathcal{K}_t$ and gradually reaches new regions of the current solution space.

Operationally, E-UoT iteratively proposes such neighboring candidates, evaluates them, and retains promising ones for further expansion. The defining constraint is that the problem formulation and governing rules remain fixed throughout the search, so novelty arises from traversing previously unvisited parts of the existing space rather than importing external structure or transforming the space itself.

3.3 Transformational UoT

Boden characterizes transformational creativity as changing the defining dimensions of a conceptual space so that possibilities excluded by the original space can become reachable. We interpret this computationally as modifying selected mutable rules or assumptions R while preserving the underlying problem and its immutable requirements H.

Transformational UoT searches over alternative rule sets R' while preserving the immutable task requirements H:

$$\underbrace{p_\theta(s \mid P, H, R)}_{\text{direct search}} \quad \Longrightarrow \quad \underbrace{p_\theta(R' \mid P, H, R)\, p_\theta(s \mid P, H, R')}_{\text{transformational search}}.$$

The product denotes a sequence of conditional generation stages rather than an exact factorization of $p_\theta(s \mid P, H, R)$.

Not every change to R constitutes a useful transformation. T-UoT therefore retains only candidate rule changes that preserve the original task and its core difficulty while opening a meaningfully different region of the solution space.

3.4 Three Executable Creative Search Programs

UoT does not claim analogy, iterative search, or rule modification as new primitives. Rather, it operationalizes an established theory of creativity as three executable and empirically testable search programs. Full pseudocode, operator definitions, selection procedures, and hyperparameters are provided in Appendix B and Appendix C.

4 Experimental Setup

We evaluate UoT across task regimes characterized by two properties that directly shape the creative search problem: *degree of constraint* and *objective specificity*. The degree of constraint captures how strongly problem requirements restrict the feasible solution space; stronger constraints narrow the set of admissible solutions and limit the extent to which alternative mechanisms, rules, or solutions can be accommodated. Objective specificity captures how precisely the desired outcome is defined. A more specific objective narrows the search target and provides a clearer criterion for evaluating candidate solutions, whereas a less specific objective permits broader notions of relevance and similarity, expanding the range of potentially useful analogies that can be retrieved from other domains. These

dimensions therefore affect both the space over which creative reasoning operates and how that space can be searched and evaluated.

As illustrated in Fig. 2, we evaluate UoT in two complementary settings spanning these dimensions. HypoArena (Zhong et al., 2026) provides externally defined scientific hypothesis-generation tasks with relatively high degrees of constraint, while spanning higher- and lower-specificity objectives. To complement this regime, we introduce three novel creative-reasoning tasks with substantially lower degrees of constraint and expansive solution spaces: Bridge and Electricity provide relatively specific objectives, while Society evaluates reasoning under a less specific objective. Together, these settings cover distinct regions of the task space and allow us to examine how creative reasoning methods behave as the structure of the underlying search problem changes.

4.1 Novel Creative-Reasoning Tasks

We construct three tasks spanning transportation, energy, and social intervention: *One-Lane Bridge*, which requires designing a traffic-control policy under fixed crossing constraints; *Electricity Tariff*, which requires reducing peak demand without curtailing critical usage; and *Social Cohesion*, which requires increasing cross-group interaction while preserving privacy, inclusivity, and safety. Full task definitions are provided in Appendix D.1.

Following the standard view of creativity as requiring both novelty and usefulness (Runco & Jaeger, 2012), we report *Creativity* as the mean of Novelty and Utility, with Feasibility reported separately as a validity measure. Outputs are canonicalized before evaluation to reduce effects of response length and presentation. Ten PhD students provide human judgments, and Gemini 2.5 Pro provides automated evaluations. Full scoring procedures, novelty computation, prompts, and known-solution baselines are provided in Appendix F.

4.2 HypoArena

We evaluate on a fixed panel of 100 HypoArena tasks from its three single-hypothesis research domains: biomedical science, machine learning, and social science (34/34/32 tasks). Final outputs are canonicalized into their core hypotheses before pairwise evaluation.

We follow the HypoArena evaluation protocol using Gemma-3-12B as the primary judge. Pairwise judgments are aggregated with Bradley–Terry–Davidson (BTD) ratings, and we additionally report pairwise win rate. HypoArena emphasizes contextual grounding and evidential justification while also rewarding non-obvious inferential insight, making it suitable for evaluating creative reasoning under scientific constraints. Additional evaluation details and judge-robustness analyses are provided in the Appendix (Appendix E.3 and Appendix D.7).

4.3 Baselines and Backbones

We compare UoT against general-purpose reasoning baselines spanning direct generation and increasingly structured search: OnePrompt, Chain-of-Thought (CoT) (Wei et al., 2022), Tree of Thoughts (ToT) (Yao et al., 2023), and Enhanced Graph of Thoughts (EGoT) (Shin & Kim, 2025). We focus on matched general-purpose reasoning frameworks rather than specialized scientific-discovery systems, which often incorporate additional retrieval pipelines, tools, or domain-specific assumptions.

For the novel creative-reasoning tasks, GPT-4o is used as the shared backbone for UoT and all matched reasoning baselines; GPT-5 and DeepSeek V3.1 are included only as contextual reference models. For HypoArena, Qwen3-14B is the primary shared backbone, and GLM-4-9B is additionally evaluated to assess backbone sensitivity. Full implementation settings are provided in Appendix E.5.

5 Results

5.1 Novel Creative-Reasoning Task Results

Table 1 summarizes Creativity scores under human and automated evaluation. Among matched GPT-4o reasoning methods, T-UoT performs strongest on Bridge and Electricity, achieving the

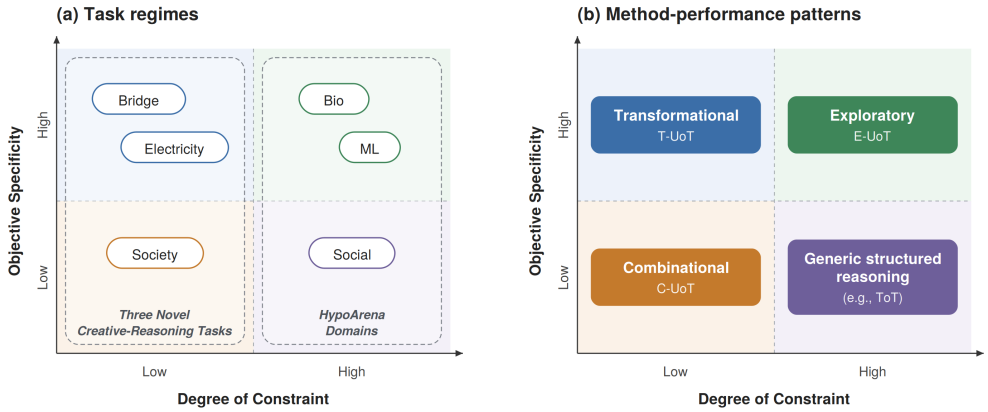


Figure 2: Task regimes and method-performance patterns along degree of constraint and objective specificity. (a) Bridge and Electricity are low-constraint, high-specificity design tasks, while Society is low-constraint and lower-specificity. HypoArena Bio and ML are high-constraint, high-specificity hypothesis-generation tasks, while HypoArena Social is high-constraint and lower-specificity. (b) The observed performance patterns are consistent with the search mechanisms of T-UoT, C-UoT, and E-UoT, while the high-constraint, low-specificity regime shows stronger performance from generic structured reasoning methods such as ToT.

Table 1: Novel creative-reasoning task creativity results. Human and LLM-based creativity scores across the three novel creative-reasoning tasks. Creativity is the mean of Novelty and Utility. Full Novelty, Utility, and Feasibility results are reported in Appendix D.2. Bold indicates the best result among matched GPT-4o reasoning methods; underlined values indicate the best contextual reference model.

	Bridge		Electricity		Society	
Method	**Human C**	**LLM C**	**Human C**	**LLM C**	**Human C**	**LLM C**
OnePrompt	0.480	0.566	0.540	0.508	0.550	0.522
CoT	0.590	0.651	0.560	0.450	0.500	0.586
ToT	0.580	0.579	0.530	0.488	0.530	0.574
EGoT	0.650	0.641	0.500	0.409	**0.560**	0.613
C-UoT	0.580	0.664	0.440	0.512	**0.560**	**0.649**
E-UoT	0.600	0.589	0.430	0.463	0.550	0.545
T-UoT	**0.680**	**0.698**	**0.620**	**0.526**	0.550	0.628
GPT-5	0.620	0.649	0.580	0.504	0.640	0.631
DeepSeek V3.1	0.460	0.458	0.580	0.571	0.720	0.723

highest human-rated Creativity on Bridge (0.680) and Electricity (0.620), as well as the highest LLM-rated scores on Bridge (0.698) and Electricity (0.526).

As shown in Fig. 2, Bridge and Electricity combine relatively low degrees of constraint with high objective specificity. A lower degree of constraint leaves greater flexibility in the rule-generation stage $p_\theta(R' \mid P, H, R)$, since more alternative rule sets R' can modify the search space while still permitting solutions that satisfy the immutable requirements H. High objective specificity, in turn, provides a clearer criterion for evaluating and selecting among the generated rule sets R', helping T-UoT identify transformations that are most promising for the target objective.

C-UoT shows the complementary pattern on Society, where it ties for the highest human-rated Creativity among matched GPT-4o methods (0.560) and achieves the highest LLM-rated score (0.649). As illustrated in Fig. 2, Society combines a low degree of constraint with relatively low objective specificity. Lower objective specificity makes the analogue-retrieval stage $p_\theta(a \mid P)$ less restrictive, allowing a broader range of problems from other domains to be considered functionally relevant to the target. Once a transferable mechanism m is identified, the low degree of constraint

Table 2: Qwen3-14B on HypoArena (100 tasks; Gemma-3-12B judge; canonicalized core hypotheses). Bio 34 / ML 34 / Social 32. Bio/ML/Social/Avg: pooled BTD ratings centered at 1500. WR: pairwise win rate (%). Bold: column best.

Rank	Method	Bio	ML	Social	Avg	WR (%)
1	E-UoT	**1541.8**	**1547.1**	1516.3	**1535.4**	**32.7**
2	ToT	1513.2	1510.4	**1524.5**	1515.8	25.5
3	EGoT	1482.7	1510.6	1521.1	1505.1	25.3
4	CoT	1486.2	1493.3	1501.9	1493.7	18.8
5	C-UoT	1495.1	1477.6	1497.0	1489.6	19.5
6	T-UoT	1490.1	1480.0	1484.6	1484.6	17.3
7	OnePrompt	1490.9	1481.0	1454.5	1475.7	14.3

makes the transfer stage $p_\theta(s \mid P, m)$ less restrictive, since fewer hard requirements must be satisfied when adapting that mechanism to the target problem. Together, these conditions favor both analogue discovery and cross-domain mechanism transfer, consistent with C-UoT's stronger relative performance on Society.

The full Novelty, Utility, and Feasibility breakdown is provided in Appendix D.2. Human and automated evaluations show substantial agreement at the method level (r=0.708; Table 8), supporting the use of automated scores as a complement to human judgments. Representative examples in Appendix D.4 illustrate the distinct search behaviors induced by the three UoT modes.

5.2 HypoArena Results

On `Qwen3-14B` (Table 2), E-UoT achieves the strongest overall performance, ranking first in both biomedical science and machine learning, while ToT performs best in social science. E-UoT reaches an average BTD rating of 1535.4 and a pairwise win rate of 32.7%, compared with 1515.8 and 25.5% for the next-best method, ToT. With GLM-4-9B, OnePrompt ranks first while E-UoT remains the strongest structured method, indicating backbone dependence (Table 10).

As illustrated in Fig. 2, HypoArena tasks have a relatively high degree of constraint, while the biomedical and machine-learning domains additionally provide relatively specific objectives. Both properties make the local expansion stage $p_\theta(s' \mid P, s_k, R)$ more directed. A higher degree of constraint narrows the admissible solution space, reducing the range of viable modifications around the current candidate s_k. High objective specificity provides additional problem context about what constitutes a useful improvement, further guiding the search toward promising neighboring solutions. These conditions are therefore well matched to E-UoT's strategy of iteratively expanding within the existing solution space while keeping the governing rules R fixed.

C-UoT and T-UoT are less reliable on average in these more constrained regimes, as cross-domain transfer and rule transformation must satisfy a narrower set of admissible conditions. Nevertheless, their lower aggregate rankings do not imply that these search modes are uniformly unsuccessful. Under the primary Gemma judge, C-UoT is top-ranked on 15/100 HypoArena tasks (8 uniquely), while T-UoT is top-ranked on 11/100 tasks (6 uniquely). These task-specific advantages also recur across judges: C-UoT and T-UoT outperform ToT on 26 and 24 tasks, respectively, under at least two of three judges, and on 11 tasks each under all three judges (Appendix D.8). Thus, although exploratory search is more reliable on average in the highly constrained HypoArena setting, combinational and transformational search can still uncover stronger hypothesis directions on particular problems.

The main ranking conclusions are not specific to the Gemma judge: re-scoring the same frozen Qwen3-14B outputs with two additional judges preserves the relative advantages of E-UoT over OnePrompt and ToT ($\rho = 0.59$–0.79 across judge pairs), while hard preference reversals remain rare (2.6–5.4%; Appendix D.7).

Operator ablation. To isolate individual creativity-specific operations, we remove the structured cross-domain transfer in C-UoT, the option to switch between mechanisms during E-UoT expansion, and rule transformation in T-UoT while otherwise preserving each mode's search procedure. A

Table 3: Operator ablation on 100 HypoArena tasks. Values are mean pairwise win rates of the full variant against its ablated counterpart (chance $= 50\%$).

Mode	Bio	ML	Social	All
C-UoT	**58.8%**	**57.0%**	**59.0%**	**58.3%**
E-UoT	53.3%	49.2%	55.0%	52.4%
T-UoT	54.1%	53.4%	55.1%	54.2%

position-debiased pairwise judge compares full and ablated outputs on the same 100 HypoArena tasks (Table 3).

The largest ablation effect occurs for C-UoT, whose full variant is preferred in 58.3% of comparisons, showing that structured cross-domain transfer contributes beyond the surrounding pipeline. T-UoT is preferred in 54.2% of comparisons, providing a smaller but consistent indication that explicitly modifying the rule set can expose useful hypothesis directions beyond the same downstream search without transformation. For E-UoT, the ablation removes only mechanism switching while retaining the same ToT-style search controller and local refinement of the current hypothesis. Full E-UoT wins 52.4% of these comparisons, indicating only a modest incremental benefit from mechanism switching. Together with E-UoT's substantially stronger performance than ToT in Table 2, this suggests that its overall advantage is not driven primarily by mechanism switching alone, but by the broader exploratory search procedure, including local refinement.

Failure-mode interventions. Stage-level traces suggest that C-UoT often loses retrieved mechanisms during final synthesis, while T-UoT may fail to propagate the consequences of selected rule transformations (Appendix D.9). On frozen failure cases, preserving the C-UoT analogy mapping improves average rank from $\#4$ to $\#1.5$, while explicitly deriving a mechanistic consequence from the selected T-UoT transformation improves rank from $\#7$ to $\#2.5$. These diagnostic interventions suggest that useful creative structure can be discovered upstream but lost during final synthesis; full intervention details are provided in Appendix D.9. This suggests that the limitation lies partly in the backbone's ability to reason over and integrate the discovered creative structure, and that stronger reasoning capabilities could allow these search modes to realize their benefits more reliably.

6 Conclusions and Future Work

We introduced Universe of Thoughts (UoT), a computational framework that operationalizes combinational, exploratory, and transformational creativity as distinct executable search processes for Large Language Models. Rather than treating creativity as a single undifferentiated capability, UoT separates cross-domain recombination, within-space exploration, and transformation of the rules that define the solution space.

Across complementary creative-reasoning and scientific hypothesis-generation settings, the three search modes exhibit distinct performance patterns depending on task structure. Combinational search performs particularly well when broader objectives permit wider analogue discovery and relatively few constraints facilitate mechanism transfer; exploratory search benefits from more constrained, higher-specificity settings that provide stronger guidance for local improvement; and transformational search performs strongly when flexible solution spaces permit viable rule changes while clear objectives support their evaluation. These results suggest that the effectiveness of different creative search processes depends in part on the structure of the problem, rather than one mode being uniformly strongest across settings.

The current implementation nevertheless depends substantially on the reasoning, search, abstraction, and transfer capabilities of the underlying foundation model. Future work can therefore investigate models trained explicitly to support these creative operations, with the goal of improving their reliability and effectiveness across task regimes. Such models could ultimately be integrated with retrieval, experimentation, and feedback loops toward more capable systems for autonomous scientific discovery.

AI USE STATEMENT

We used generative AI tools to assist with software implementation and debugging, data analysis and interpretation, literature review, preparation and refinement of scientific figures, and feedback on experimental methodology and presentation. Generative AI was also used to edit, reorganize, and refine portions of the manuscript for clarity, structure, grammar, and readability, based on research content, interpretations, and conclusions developed by the authors.

Generative AI was not used to originate the core research question or the central conceptual framework of this work, nor to independently determine the scientific conclusions of the study. Research decisions, experimental configurations, interpretation of the final results, and all claims presented in the paper were reviewed and determined by the authors. AI-generated or AI-assisted code was tested and verified by the authors, and AI-assisted literature discussions and manuscript content were checked against the cited sources.

The authors reviewed all AI-assisted work and take full responsibility for the final content of the paper, including its methods, experiments, analyses, claims, text, code, and figures.

ETHICS STATEMENT

This work studies computational methods for creative reasoning with large language models. The experiments do not involve deployment of generated hypotheses or solutions in real-world scientific, medical, or other high-stakes settings. Generated scientific hypotheses are evaluated only as research artifacts and should not be interpreted as experimentally validated scientific conclusions or recommendations.

Our evaluation uses both human and LLM-based judgments. Human evaluators assess model outputs rather than participating as subjects of behavioral or clinical experimentation. As with any subjective evaluation of creativity, judgments of novelty, usefulness, and hypothesis quality may reflect evaluator preferences or biases. We therefore report aggregate results, compare human and automated evaluations where applicable, and analyze consistency across multiple LLM judges.

Creative reasoning systems may generate plausible but incorrect, unsupported, or potentially misleading ideas. This risk is particularly relevant for scientific hypothesis generation, where novel hypotheses may still lack sufficient grounding or evidential support. We therefore evaluate grounding, feasibility, and evidential support separately and emphasize that generated hypotheses require independent scientific validation before practical use.

The authors are responsible for the design, analysis, interpretation, and reporting of this work and have reviewed the research for potential ethical concerns in accordance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Universe of Thoughts algorithms, including their executable operators, scoring and selection procedures, and default hyperparameters, in Section 3, with Boden's source definitions in Appendix A and full pseudocode, thresholds, and selection rules in Appendices B and C. The experimental setup, benchmark construction, task selection, backbone models, evaluation protocols, and baseline configurations are described in Section 4. Evaluation prompt templates, additional implementation details, known-solution references used for novelty evaluation, and supplementary analyses are provided in the Appendix; generation-operator prompts are released with the code.

To support replication of the reported results, we provide the implementation of UoT together with the experiment configuration files, evaluation scripts, and prompts used in the paper as supplementary material. Where external benchmarks are used, we follow their published evaluation protocols and document any preprocessing or task-selection decisions made in our experiments. Randomness, generation settings, model versions, search budgets, and other experiment-specific parameters are reported where applicable. The novel creative-reasoning task evaluations include both human and automated judgments, and the HypoArena experiments additionally report judge-consistency and operator-ablation analyses to assess the robustness of the main findings.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, 2004.

Margaret A Boden. Creativity in a nutshell. *Think*, 5(15):83–96, 2007.

Margaret A Boden. Computer models of creativity. *Ai Magazine*, 30(3):23–23, 2009.

Tingting Chen, Beibei Lin, Zifeng Yuan, Qiran Zou, Hongyu He, Anirudh Goyal, Yew-Soon Ong, and Dianbo Liu. Hypospace: A diagnostic benchmark for set-valued hypothesis generation under underdetermination and sublinear coverage bounds. *arXiv preprint arXiv:2510.15614*, 2025.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, et al. Inductive or deductive? rethinking the fundamental reasoning abilities of llms. *arXiv preprint arXiv:2408.00114*, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Amit Gangwal, Azim Ansari, Iqrar Ahmad, Abul Kalam Azad, Vinoth Kumarasamy, Vetriselvan Subramaniyan, and Ling Shing Wong. Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Frontiers in pharmacology*, 15: 1331062, 2024.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. Llms can realize combinatorial creativity: generating creative ideas via llms for scientific research. *arXiv preprint arXiv:2412.14141*, 2024.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5888–5899, 2025.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, et al. Creativityprism: A cross-domain evaluation framework for large language model creativity. *arXiv preprint arXiv:2510.20091*, 2025.

Andrew J Lew, Yuan Cao, and Markus J Buehler. Projectionbench: Evaluating scientific hypothesis generation in llms under progressive information disclosure. *arXiv preprint arXiv:2605.30284*, 2026.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. Automated creativity evaluation for large language models: A reference-based approach. In *EMNLP (Findings)*, pp. 21475–21488, 2025.

Fan Liu, Xiaozhao Zeng, and Hao Liu. Towards multimodal data-driven scientific discovery powered by llm agents. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. Hypobench: Towards systematic and principled benchmarking for hypothesis generation. *arXiv preprint arXiv:2504.11524*, 2025.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 2026b.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Siba Smarak Panigrahi, Jovana Videnović, and Maria Brbic. Heurekabench: A benchmarking framework for ai co-scientist. In *International Conference on Learning Representations*, volume 2026, pp. 11029–11061, 2026.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. Ai idea bench 2025: Ai research idea generation benchmark. *arXiv preprint arXiv:2504.14191*, 2025.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Nature communications*, 17(1):3625, 2026.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

Samuel Schapiro, Sumuk Shashidhar, Alexi Gladstone, Jonah Black, Royce Moon, Dilek Hakkani-Tur, and Lav R Varshney. Combinatorial creativity: A new frontier in generalization abilities. *arXiv preprint arXiv:2509.21043*, 2025.

Tan Min Sen, Zachary Choy Kit Chun, Syed Ali Redha Alsagoff, Nadya Yuki Wangsajaya, Banerjee Mohor, Swaagat Bikash Saikia, and Alvin Chan. Automated creativity evaluation of language models across open-ended tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23139–23173, 2026.

SungUk Shin and Youngjoon Kim. Enhancing graph of thought: Enhancing prompts with llm rationales and dynamic temperature control. In *The Thirteenth International Conference on Learning Representations*, 2025.

Herbert A Simon. The structure of ill structured problems. *Artificial intelligence*, 4(3-4):181–201, 1973.

Narotam Singh, Vaibhav Chaudhary, Nimisha Singh, Neha Soni, and Amita Kapoor. Transforming business with generative ai: Research, innovation, market deployment and future shifts in business models. *arXiv preprint arXiv:2411.14437*, 2024.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5303–5324, 2024.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Yang Xiao, et al. Innovatorbench: evaluating agents' ability to conduct innovative ai research. In *International Conference on Learning Representations*, volume 2026, pp. 36144–36185, 2026.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*, 2023.

Junchi Yu, Ran He, and Rex Ying. Thought propagation: An analogical approach to complex reasoning with large language models. *arXiv preprint arXiv:2310.03965*, 2023.

Jintian Zhang, Kewei Xu, Jingsheng Zheng, Zhuoyun Yu, Yuqi Zhu, Yujie Luo, Lanning Wei, Shuofei Qiao, Lun Du, Da Zheng, et al. Innogym: Benchmarking the innovation potential of ai agents. In *International Conference on Learning Representations*, volume 2026, pp. 71349–71379, 2026.

Tianshi Zheng, Kiu Wai Tam, Kim Hue Nam Nguyen, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 99018–99088, 2026.

Tianyun Zhong, Wangyi Jiang, Wei Wang, Xuanang Chen, Yaojie Lu, Shiwei Ye, Yuzhen Shi, Boyu Yang, Jinghang Wang, Han Li, et al. Before the action: Benchmarking llms on prospective hypothesis discovery. *arXiv preprint arXiv:2607.15766*, 2026.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058, 2024.

A CONNECTION TO BODEN'S DEFINITIONS

Our computational formulation follows Boden's distinction among combinational, exploratory, and transformational creativity (Boden, 2004; 2007; 2009). Boden characterizes the three modes as follows (bracketed glosses are added by us):

1. **Combinational creativity.** "Combinational creativity, produces unfamiliar combinations [i.e., solutions] of familiar ideas, and it works by making associations between ideas that were previously only indirectly linked."
2. **Exploratory creativity.** "In exploratory creativity, the person moves through the [solution] space, exploring it to find out what's there (including previously unvisited locations [i.e., new ideas])—and, in the most interesting cases, to discover both the potential and the limits of the space in question."
3. **Transformational creativity.** "In transformational creativity, a [solution] space or style itself is transformed by altering (or dropping) one or more of its defining dimensions [i.e., rules]."

We translate these cognitive descriptions into computational search operations. In C-UoT, an "unfamiliar combination" is operationalized as transferring a functionally relevant mechanism from a contextually different source into the target problem. In E-UoT, movement through the existing conceptual space is operationalized as expanding the currently explored candidate set while keeping the problem rules fixed. In T-UoT, transformation of the conceptual space is operationalized as modifying mutable rules or assumptions while preserving the task's immutable requirements.

Table 4: Correspondence between Boden's creativity modes and UoT operators.

Boden concept	**Computational interpretation**	**UoT operator**
Unfamiliar combinations of familiar ideas	Import structure from another context	analogue → mechanism → transfer
Explore previously unvisited locations in an existing space	Expand the candidate set with rules fixed	SHIFT / DEEPEN
Alter defining dimensions of the space	Modify mutable assumptions or rules	gated rule transformation

We reinterpret these three modes as distinct strategies for a common computational problem:

> *Given a problem and the solutions already explored, what should a reasoning system change in order to reach a useful solution that has not yet been considered?*

This perspective yields three different search regimes. Combinational reasoning changes the *source of structure* used to construct a solution; exploratory reasoning changes the *candidate solution* while preserving the current rules; and transformational reasoning changes the *rules or assumptions* that define which solutions are admissible. These distinctions define the computational structure of Universe of Thoughts (UoT). The staged conditional-generation formalization of each mode appears in Section 3; full pseudocode and selection details are provided in Appendix B and Appendix C.

B UOT PSEUDOCODE

This appendix specifies the executable operator interfaces and the control flow for Combinational, Exploratory, and Transformational UoT (Section 3). The operators name what is generated or evaluated; benchmark-specific prompts instantiate the corresponding representation (policy components for novel creative-reasoning tasks; mechanisms and hypotheses for HypoArena). Soft-gate thresholds default to $\tau_g{=}\tau_i{=}2$.

B.1 EXECUTABLE OPERATOR INTERFACES

All UoT operators are instantiated as structured LLM calls. Across both evaluation settings, the three modes share the same search objects—external structure, candidate solutions under fixed rules, and mutable assumptions—while differing in intermediate representation and selection budgets.

C-UoT. GENERATEANALOGUES proposes source problems that are functionally related to the target but contextually distinct from already explored solutions. SOLVESOURCE obtains a source-side solution or structured account of how the analogue addresses its problem. EXTRACTMECHANISMS removes source-specific surface detail and retains the transferable mechanism. TRANSFERANDSYNTHESIZE instantiates that mechanism in the target problem and produces a candidate solution or hypothesis. Soft quality filters and final ranking then retain promising transfers (Appendix C).

E-UoT. EXPAND generates neighboring candidates while preserving the problem formulation and mutable rules. Each expansion explores either a different direction by changing the current mechanism (SHIFT) or an underexplored consequence of the current mechanism (DEEPEN). Candidates are evaluated and retained by the shared search controller ($\mathrm{LLM}_{\mathrm{vote}}$ / beam retention), matching the Tree-of-Thoughts control structure used by the ToT baseline. FINALIZE converts the final beam into submitted outputs.

T-UoT. EXTRACTRULES identifies mutable assumptions and separates them from immutable task requirements. GENERATETRANSFORMATIONS proposes modifications to those assumptions (e.g., remove, relax, invert, add, or replace). A retained transformation must preserve the original task goal and core difficulty while opening a meaningfully different solution region. REFRAME constructs the working problem representation under the selected transformed rules, after which CREATIVESEARCH runs downstream solution search in that reframed space.

B.2 ALGORITHMS

Algorithm 1: Combinational UoT (HypoArena: $R_{\max}$=5 with soft gate; novel tasks: final F/U/N ranking)

Input: Target problem P_t; known/explored solutions $\mathcal{K}_t$
Output: Ranked candidates $\mathcal{Y}_C$

$\mathcal{Y} \leftarrow \emptyset$;
for $r = 1, \ldots, R_{\max}$ **do**
  $\mathcal{A} \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{GenerateAnalogues} \mid P_t, \mathcal{K}_t \cup \mathcal{Y};\ \text{far})$;
  foreach $a \in \mathcal{A}$ **do**
    $\mathcal{S}_a \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{SolveSource} \mid a)$;
    $\mathcal{Z}_a \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{ExtractMechanisms} \mid a, \mathcal{S}_a)$;
    $y \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{TransferAndSynthesize} \mid P_t, \mathcal{Z}_a)$;
    $(g, i) \leftarrow \mathrm{LLM}_{\mathrm{score}}(y \mid P_t;\ \text{grounding/insight} \in \{1..5\})$;
    `// HypoArena soft gate; novel creative-reasoning tasks defer filtering to final F/U/N ranking` **if** $g \geq \tau_g \wedge i \geq \tau_i$ **then**
      $\mathcal{Y} \leftarrow \mathcal{Y} \cup \{y\}$;
foreach $y \in \mathcal{Y}$ **do**
  $\mathrm{Score}(y) \leftarrow \mathrm{LLM}_{\mathrm{score}}(y \mid P_t, \mathcal{K}_t;\ \text{pair axes or F/U/N})$;
$\mathcal{Y}_C \leftarrow \mathrm{TopK}(\mathcal{Y};\ \mathrm{Score})$;
return $\mathcal{Y}_C$;

Algorithm 2: Exploratory UoT (defaults: T=4, k=5, b=$|\mathcal{Y}_E|$)

Input: Problem P; initial beam $\mathcal{B}^{(0)}$; depth T; per-node budget k; beam width b
Output: Finalized candidates $\mathcal{Y}_E$

for $t = 0, \ldots, T-1$ **do**
  $\widetilde{\mathcal{B}}^{(t+1)} \leftarrow \emptyset$;
  foreach $s \in \mathcal{B}^{(t)}$ **do**
    $\mathcal{C} \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{Expand} \mid P, s; |\mathcal{C}| \leq k)$;
    // Each c carries $\mathrm{op}(c) \in \{\textsc{Shift}, \textsc{Deepen}\}$ $\widetilde{\mathcal{B}}^{(t+1)} \leftarrow \widetilde{\mathcal{B}}^{(t+1)} \cup \mathcal{C}$;
  $\mathcal{I} \leftarrow \mathrm{LLM}_{\mathrm{vote}}(\widetilde{\mathcal{B}}^{(t+1)}; |\mathcal{I}|=b)$;
  $\mathcal{B}^{(t+1)} \leftarrow \{\widetilde{\mathcal{B}}_i^{(t+1)} : i \in \mathcal{I}\}$;
$\mathcal{Y}_E \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{Finalize} \mid P, \mathcal{B}^{(T)})$;
return $\mathcal{Y}_E$;

Algorithm 3: Transformational UoT (paper: k=5; S_{select} as below)

Input: Problem P; immutable constraints H
Output: Candidates $\mathcal{Y}_T$

$R \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{ExtractRules} \mid P, H)$;
$\mathcal{R}' \leftarrow \mathrm{LLM}_{\mathrm{propose}}(\textsc{GenerateTransformations} \mid P, H, R; |\mathcal{R}'|=k)$;
$\mathcal{R}^* \leftarrow \emptyset$;
foreach $R' \in \mathcal{R}'$ **do**
  $(b_{\mathrm{goal}}, b_{\mathrm{diff}}, b_{\mathrm{space}}) \leftarrow \mathrm{LLM}_{\mathrm{score}}(P, H, R, R'; \{0,1\}^3)$;
  // b_{goal}: goal preserved; b_{diff}: difficulty preserved; b_{space}: solution space opened **if** $b_{\mathrm{goal}} \wedge b_{\mathrm{diff}} \wedge b_{\mathrm{space}}$ **then**
    $\mathcal{R}^* \leftarrow \mathcal{R}^* \cup \{R'\}$;
$R^* \leftarrow S_{\mathrm{select}}(\mathcal{R}^*)$;
$P' \leftarrow \textsc{Reframe}(P, R^*)$;
$\mathcal{Y}_T \leftarrow \textsc{CreativeSearch}(P')$;
return $\mathcal{Y}_T$;

C Scoring and Selection Protocol

Throughout UoT, $\mathrm{LLM}_{\mathrm{propose}}$, $\mathrm{LLM}_{\mathrm{score}}$, and $\mathrm{LLM}_{\mathrm{vote}}$ denote structured backbone calls that return candidate lists, task-specific scores, or discrete selection indices. Lightweight deterministic filters are applied where specified.

These interfaces implement the operational criteria of each search regime: functional relevance and unfamiliarity for combinational transfer, candidate selection during exploratory expansion, and goal/difficulty preservation together with solution-space opening for transformational search. Exact generation-operator prompts are released with the code; Appendix F provides the novel creative-reasoning task evaluation prompts used for scoring.

C.1 Novel Creative-Reasoning Task Configuration

C-UoT. Final candidates are ranked by feasibility, utility, and novelty for a total score of 0–100: feasibility and utility contribute 0–30 each and novelty contributes 0–40. Defaults are 10 analogues, 5 solutions per analogue, and 20 combinational candidates before TopK selection. Soft grounding/insight gates used in HypoArena are deferred; filtering is performed by the final F/U/N ranking.

E-UoT. E-UoT does not use a continuous creativity score inside the search loop. $\mathrm{LLM}_{\mathrm{vote}}$ instead returns exactly b candidate indices, using the same voting interface as the matched ToT baseline. The default configuration is $T = 4$, $k = 5$, and $b = |\mathcal{Y}_E|$.

T-UoT. Each transformation is evaluated using the Boolean gate $(b_{\text{goal}}, b_{\text{diff}}, b_{\text{space}})$. S_{select} is set-valued over gated transforms, and CREATIVESEARCH runs E-UoT under each retained transform.

C.2 HYPOARENA CONFIGURATION

C-UoT. Each transferred candidate is evaluated during search on $(g, i) \in \{1, \ldots, 5\}^2$, corresponding to grounding and insight. Candidates are retained when $g \geq 2$ and $i \geq 2$, with at most $R_{\max} = 5$ proposal rounds and typically three analogues per target tension.

Final ranking uses the HypoArena-pair criteria on a 0–100 scale: grounding, insight, justification, and redundancy. Grounding acts as a hard gate with threshold $\tau = 50$. Among candidates passing the gate, ranking is

$$\text{mean}(g, i, j) - \lambda \left(\frac{\text{redundancy}}{100} \right), \qquad \lambda = 1.5.$$

Mechanism-diverse TopK selection produces the final submitted set.

E-UoT. The exploratory controller matches the novel creative-reasoning task defaults ($T = 4$, $k = 5$, $b = |\mathcal{Y}_E|$), with LLM_{vote} returning exactly b indices as in the matched ToT baseline.

T-UoT. Each transformation is evaluated with the same Boolean gate $(b_{\text{goal}}, b_{\text{diff}}, b_{\text{space}})$. S_{select} follows the paper configuration: an analogue-probe score argmax in which one analogue is generated under each surviving transformed observation, its structural richness is scored programmatically, and

$$R^* = \arg\max_{R'} \text{ProbeScore}(R').$$

The default alternative is an LLM index vote over the gated set. CREATIVESEARCH is Alg. 1 on the reframed observation.

D SUPPLEMENTARY RESULTS

D.1 NOVEL CREATIVE-REASONING TASKS

Table 5 gives the full objective, context, and constraint specifications for the three novel creative-reasoning tasks summarized in Section 4.

Table 5: Novel creative-reasoning tasks

Task	Description
Task 1: One-Lane Bridge	**Objective:** Design a traffic-control policy that minimizes average vehicle delay. **Context:** A single-lane bridge with known traffic patterns, fixed crossing times, and direction-switching penalties. **Constraints:** No new infrastructure; traffic must remain unidirectional at all times.
Task 2: Electricity Tariff	**Objective:** Design a tariff and demand-response program that reduces daily peak load. **Context:** A residential feeder experiences predictable periods of high demand. **Constraints:** No new physical capacity; critical medical loads must never be curtailed.
Task 3: Social Cohesion	**Objective:** Design an intervention that strengthens cross-group cohesion, measured by new ties, mixing, and mutual aid. **Context:** A community contains distinct social groups and a fixed intervention period. **Constraints:** The intervention must protect privacy, remain inclusive, and maintain a safe environment.

D.2 Novel Creative-Reasoning Task Full Results

We generate 10 independent solutions for each method and report the mean score across generations. Tables 6 and 7 report the full Novelty, Utility, and Feasibility decomposition corresponding to the Creativity summary in Table 1. Creativity is the mean of Novelty and Utility.

As contextual reference points, GPT-5 and DeepSeek V3.1 are evaluated under a single task prompt, allowing each model to rely on its own internal reasoning capabilities. These are not matched backbone comparisons. GPT-4o-based T-UoT achieves higher Creativity than GPT-5 in four of six task–evaluator comparisons and than DeepSeek V3.1 in three of six; the advantage is concentrated on Bridge and Electricity, while the frontier models lead on the more ambiguous Society task.

Table 6: Human evaluation results across all creative reasoning tasks. C=Creativity, N=Novelty, U=Utility, and F=Feasibility. Bold: column best; underline: column second-best.

Method	Bridge Task				Electricity Task				Society Task			
	C	N	U	F	C	N	U	F	C	N	U	F
OnePrompt	0.480	0.450	0.500	0.820	0.540	0.480	0.610	0.870	0.550	0.490	0.610	**0.810**
CoT	0.590	0.610	0.570	0.750	0.560	0.520	0.600	**0.920**	0.500	0.430	0.580	**0.810**
ToT	0.580	0.600	0.560	0.680	0.530	0.540	0.520	0.760	0.530	0.440	0.610	**0.810**
EGoT	0.650	0.650	0.650	0.810	0.500	0.440	0.550	0.880	0.560	0.480	0.640	0.800
C-UoT	0.580	0.550	0.610	0.820	0.440	0.440	0.440	0.870	0.560	0.500	0.620	**0.810**
E-UoT	0.600	0.600	0.610	0.790	0.430	0.380	0.470	0.740	0.550	0.480	0.620	0.730
T-UoT	**0.680**	**0.690**	**0.670**	0.770	**0.620**	**0.610**	**0.640**	0.800	0.550	0.490	0.600	0.730
GPT-5	0.620	0.610	0.640	0.840	0.580	0.600	0.560	0.770	0.640	0.680	0.600	0.770
DeepSeek V3.1	0.460	0.460	0.470	**0.910**	0.580	0.580	0.580	0.750	**0.720**	**0.690**	**0.740**	**0.810**

Table 7: LLM evaluation results across all creative reasoning tasks. C=Creativity, N=Novelty, U=Utility, and F=Feasibility. Bold: column best; underline: column second-best.

Method	Bridge Task				Electricity Task				Society Task			
	C	N	U	F	C	N	U	F	C	N	U	F
OnePrompt	0.566	0.732	0.400	0.925	0.508	0.587	0.430	0.680	0.522	0.675	0.370	0.730
CoT	0.651	0.742	0.560	0.980	0.450	0.541	0.360	0.675	0.586	0.732	0.440	0.560
ToT	0.579	0.799	0.360	0.900	0.488	**0.657**	0.320	0.610	0.574	0.727	0.420	0.495
EGoT	0.641	0.643	0.640	**1.000**	0.409	0.428	0.390	**0.825**	0.613	0.775	0.450	0.525
C-UoT	0.664	0.728	0.600	**1.000**	0.512	0.563	0.460	0.675	0.649	0.799	0.500	0.530
E-UoT	0.589	0.778	0.400	1.000	0.463	0.587	0.340	0.755	0.545	0.810	0.280	0.575
T-UoT	**0.698**	**0.807**	0.590	0.930	0.526	0.641	0.410	0.665	0.628	**0.846**	0.410	0.575
GPT-5	0.649	0.649	**0.650**	**1.000**	0.504	0.497	0.510	0.815	0.631	0.652	0.610	**0.870**
DeepSeek V3.1	0.458	0.476	0.440	0.950	**0.571**	0.563	**0.580**	0.770	**0.723**	0.767	**0.680**	0.825

D.3 Human–LLM Judge Agreement

To assess the reliability of automated evaluation, we compare Gemini 2.5 Pro scores with human judgments. At the method level ($n = 27$ method–task cells), Creativity shows strong agreement ($r = 0.708$, $p < 10^{-4}$), while Utility shows moderate agreement ($r = 0.481$). Agreement is weaker for Novelty and Feasibility. Instance-level correlations are lower, suggesting that the LLM judge is more reliable for comparing aggregate method behavior than for assigning scores to individual solutions.

D.4 Sample Solutions for Bridge Task

Sample solutions for the Bridge Task are provided in Table 9 for illustration.

D.5 Novel Creative-Reasoning Task Sensitivity Analysis

Figures 3a and 3b present sensitivity of C-UoT's performance on the Bridge Task with respect to two key parameters: the number of analogous problems and the number of solutions generated per problem.

Figure 3a shows that performance follows a U-shaped curve as the number of analogous problems increases. We hypothesize this indicates three distinct regimes: a small number of analogies may yield a few high-quality, easily combinable thoughts; a moderate number can introduce conceptual noise that hinders effective synthesis; and a very large number provides a sufficiently rich and diverse pool of concepts where the benefits of flexibility overcome the noise.

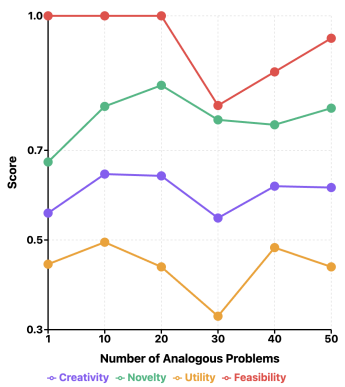
Figure 3b shows that increasing the number of solutions per analogous problem, while exhibiting some fluctuation, generally leads to improved performance. This suggests that a larger pool of "conceptual primitives" or thoughts derived from each analogy provides more raw material for synthesizing a high-quality final solution.

Table 8: Pearson correlation between human and Gemini 2.5 Pro scores on the novel creative-reasoning tasks (identical Utility/Feasibility prompts). Method level: one cell per method×task (n=27). Instance level: paired individual solutions (n=267). Bold: strongest correlation in each row block.

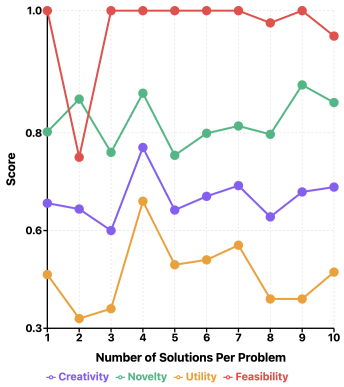
Metric	r	p	n
Method level			
Creativity	**0.708**	3.6×10^{-5}	27
Utility	0.481	1.1×10^{-2}	27
Novelty	0.224	2.6×10^{-1}	27
Feasibility	0.039	8.5×10^{-1}	27
Instance level			
Creativity	**0.342**	9.6×10^{-9}	267
Utility	0.249	3.9×10^{-5}	267
Novelty	0.168	5.8×10^{-3}	267
Feasibility	0.067	2.7×10^{-1}	267

Table 9: Sample solutions for the Bridge task

Method	**Solution**
C-UoT	The core functional mechanism is a cooperative negotiation system. Its essential components are a driver interface for signaling crossing intent (input) and a central allocation system. The system aggregates these individual intents to form a comprehensive demand profile, and based on this, allocates specific crossing times (output) back to the drivers. This interaction, inspired by negotiation protocols in distributed systems, ensures fair and efficient resource use.
E-UoT	Based on a feedback-based mechanism, this solution implements a dynamic scheduling model that uses real-time traffic data (input) to continuously adjust vehicle crossing schedules. This interaction dynamically allocates crossing windows based on current conditions (output), which serves to optimize overall flow and reduce vehicle idle times.
T-UoT	This solution models the crossing scenario using game theory, treating vehicles as rational players aiming to minimize their individual travel time. The system analyzes the strategic interactions between all vehicles and calculates a Nash Equilibrium—a state where no single vehicle can benefit by unilaterally changing its crossing strategy. The resulting equilibrium strategy is then implemented as the global crossing schedule, ensuring a stable and efficient outcome for the entire system.



(a) Effect of number of analogous problems



(b) Effect of number of solutions generated

Figure 3: *Sensitivity analysis for UoT*

D.6 GLM-4-9B HypoArena Results

Table 10 reports the same 100-task HypoArena panel with GLM-4-9B as the generation backbone. Unlike the Qwen3-14B setting (Table 2), OnePrompt ranks first overall (Avg. BTD 1560.4; WR 46.3%), while E-UoT remains the strongest structured search method (Avg. BTD 1542.1; WR 40.7%), ahead of ToT, CoT, EGoT, C-UoT, and T-UoT. This reversal relative to Qwen indicates that which creativity mode is most effective can depend on backbone capability, even when the task panel and judge are held fixed.

D.7 Judge Consistency on Frozen Qwen-100 Hypotheses

To assess whether the HypoArena results depend on a particular LLM judge, we re-score the same frozen Qwen3-14B hypotheses with Mistral-Small-24B and Prometheus-2 while holding all generated outputs fixed. Table 11 reports agreement in pairwise outcomes, hard preference reversals, and Spearman correlation over case–method win shares. Table 12 further measures consistency in the relative advantage of UoT variants over key baselines on a per-case basis.

Across judge pairs, case–method rankings show moderate agreement ($\rho = 0.587$–0.616), while hard A↔B reversals remain rare (2.6–5.4%). More importantly, the relative advantage of E-UoT over OnePrompt and ToT is consistently preserved across judges ($\rho = 0.59$–0.79), with similarly moderate consistency for C-UoT and T-UoT relative to ToT. These results indicate that exact scores vary across judges, but the main comparative trends are not driven by a single evaluator.

Table 10: GLM-4-9B on HypoArena (100 tasks; Gemma-3-12B judge; canonicalized core hypotheses). Bio 34 / ML 34 / Social 32. Bio/ML/Social/Avg: pooled BTD ratings centered at 1500. WR: pairwise win rate (%). Bold: column best.

Rank	Method	Bio	ML	Social	Avg	WR (%)
1	OnePrompt	**1577.9**	1537.0	**1564.6**	**1560.4**	**46.3**
2	E-UoT	1558.8	**1545.4**	1522.4	1542.1	40.7
3	ToT	1504.0	1528.2	1516.2	1516.1	34.0
4	CoT	1506.2	1475.3	1518.1	1500.0	30.2
5	EGoT	1438.8	1520.3	1442.1	1467.1	25.2
6	C-UoT	1444.6	1454.9	1483.5	1460.7	21.7
7	T-UoT	1469.7	1439.0	1453.0	1453.6	19.0

Table 11: Judge consistency on frozen Qwen-100 hypotheses (Gemma-3-12B, Mistral-Small-24B, Prometheus-2). Spearman ρ over case$\times$method win-shares (n=700). Exact 3-way agreement and hard A$\leftrightarrow$B flip over n=2100 matched pairs. Bold: column best (lowest flip).

Judge pair	Spearman ρ	Exact 3-way	Hard A$\leftrightarrow$B flip
Gemma $\leftrightarrow$ Mistral	**0.616**	**0.604**	**2.6%**
Gemma $\leftrightarrow$ Prometheus	0.587	0.478	4.4%
Mistral $\leftrightarrow$ Prometheus	0.604	0.520	5.4%

Table 12: Spearman ρ on per-case method contrasts across judges (G=Gemma, M=Mistral, P=Prometheus). Bold: row best.

Contrast	G$\leftrightarrow$M	G$\leftrightarrow$P	M$\leftrightarrow$P
Δ(E $-$ OnePrompt)	0.71	0.71	**0.79**
Δ(E $-$ ToT)	0.65	0.59	**0.69**
Δ(C $-$ ToT)	**0.67**	0.49	0.52
Δ(T $-$ ToT)	**0.63**	0.51	0.57

D.8 Qualitative HypoArena Examples and Non-Redundant Successes

Table 13 presents illustrative high-performing cases for each UoT mode. For visualization, we select one case per mode in which that mode ranks first among the seven methods with a substantial margin over the next-best method. These examples are not used as evidence of aggregate performance; rather, they illustrate the qualitatively different reasoning operations induced by combinational, exploratory, and transformational UoT. We compare the canonicalized core hypothesis produced by each UoT mode with the ToT hypothesis from the same frozen Qwen3-14B generations.

The examples illustrate distinct search behaviors. C-UoT introduces structure from an external analogy, E-UoT develops a new hypothesis while retaining the original conceptual framing, and T-UoT changes the explanatory assumptions that define the hypothesis space. These qualitative differences complement the aggregate results by showing that the UoT variants do not merely produce alternative phrasings of similar hypotheses, but induce different forms of creative search.

Non-redundant C-UoT and T-UoT successes. Although C-UoT and T-UoT rank below ToT on average, they exhibit nontrivial pockets of strong performance (Table 14). Under the primary Gemma judge, C-UoT is top-ranked on 15/100 tasks (8 uniquely), while T-UoT is top-ranked on 11/100 tasks (6 uniquely). These task-specific advantages also recur across evaluators: C-UoT and T-UoT outperform ToT on 26 and 24 tasks, respectively, under at least two of three judges, and on 11 tasks each under all three judges. Thus, their weaker aggregate rankings do not imply that the corresponding search modes are uniformly unproductive; rather, they appear to access useful hypothesis directions on a subset of problems.

Table 13: Illustrative high-performing HypoArena cases for each UoT mode (Qwen3-14B; Gemma-3-12B judge). Each row shows a case in which the listed mode ranks first among seven methods with a substantial margin over the next-best method. Hypotheses are abbreviated to their canonicalized core claims.

Mode	Domain	Gap to #2	Task (brief)	UoT idea	ToT idea
C-UoT	ML	$+0.33$ vs E-UoT	Why entropy-minimization test-time adaptation fails for vision–language models under shift	Standard Shannon entropy assumes static class uniformity and clashes with pretraining-induced confidence asymmetries; use class-specific Tsallis entropy shapes tuned to those biases	Reweight entropy with class-specific "niche factors" so adaptation favors views that keep the correct class despite imperfect top-1 scores
E-UoT	Social	$+0.27$ vs EGoT	What sustains adherence to digital fitness coaches (live, avatar, or agent)	Retention depends on *dynamic* alignment: anthropomorphic design plus real-time shifts in emotional vs. technical support as user needs and adherence phases change	Engagement is a static balance of instrumental utility and parasocial bonds, mediated by anthropomorphism, authenticity, and support type
T-UoT	Social	$+0.15$ vs One-Prompt	Why holiday e-commerce returns spike beyond ordinary product dissatisfaction	Surges arise when shoppers privilege local cues (gift urgency, immediate satisfaction) over systemic factors (return windows, recipient mismatch, delayed evaluation)—heuristic misalignment, not quality failure	Returns are driven by social-identity pressure and temporal dissonance between purchase intent and post-holiday reevaluation

Table 14: Task-level success frequency on Qwen3-14B HypoArena. Top-ranked counts use the primary Gemma judge; cross-judge columns report tasks on which the method outperforms ToT under multiple judges.

Method	Top-1	Sole Top-1	$>$ ToT, $\geq$2/3 judges	$>$ ToT, 3/3 judges
C-UoT	15	8	26	11
T-UoT	11	6	24	11

D.9 Failure-Mode Diagnosis and Interventions

Stage-level diagnosis. The C-UoT trace analysis suggests that failures often occur after analogy retrieval rather than because no meaningful analogy is found (Table 15). An analogue and mapping are selected on essentially all tasks, and mapped-mechanism specificity is nearly identical when C-UoT does and does not beat ToT. In contrast, preservation of mapped mechanism content into the final hypothesis is higher when C-UoT beats ToT, with strong transfer more frequent in wins than non-wins. This pattern motivates the downstream-synthesis intervention below.

For T-UoT, transformation execution alone does not distinguish success (Table 16): rule changes are applied in both winning and losing cases, and lexical preservation into the final hypothesis is weak in both groups. Among the 12 largest applied-transform failures, 7/12 involved reasonable transformations that were not converted into competitive hypotheses, while 5/12 involved weak, misaimed, or derailing transformations. This suggests a dual bottleneck for T-UoT: downstream consequence reasoning is often weak, while the transformation gate itself can also admit poor rule changes.

To test whether the diagnosed C-UoT and T-UoT bottlenecks are actionable, we perform targeted interventions on frozen failure cases while keeping the upstream creative search fixed (Table 17). For C-UoT, we preserve the retrieved analogy and mapping and modify only the final synthesis using a human-written rewrite; this intervention tests whether the frozen intermediate analogy can support a stronger final hypothesis when downstream synthesis is improved. For T-UoT, we keep the selected transformation fixed but require the model to explicitly derive one mechanistic consequence of that transformation before producing the final hypothesis.

We report

$$\Delta = P(\text{intervened} \succ \text{original}) - 0.5,$$

averaged over four random seeds, together with the resulting rank among the intervened method and five fixed baselines. Positive Δ indicates that correcting the diagnosed downstream stage improves the final hypothesis without changing the upstream creative operation.

Table 15: C-UoT stage-level diagnosis relative to ToT on Qwen-100. Wins denote tasks where C-UoT is preferred to ToT by the primary judge.

Signal	C-UoT $>$ ToT	C-UoT $\not>$ ToT
Analogy selected	100%	100%
Mapped-mechanism specificity	0.97	0.97
Mapped $\rightarrow$ final preservation	0.53	0.43
Strong transfer	52%	32%

Table 16: T-UoT stage-level diagnosis relative to ToT on Qwen-100.

Signal	T-UoT $>$ ToT	T-UoT $\not>$ ToT
Gate passed	100%	100%
Transform applied	75%	83%
Rules $\rightarrow$ framing	0.75	0.83
Rules $\rightarrow$ final	0.13	0.16

Table 17: Targeted failure-mode interventions on frozen C-UoT and T-UoT cases. Δ denotes $P(\text{intervened} \succ \text{original}) - 0.5$ averaged over four seeds. Rank is computed among the intervened method and five fixed baselines; lower is better.

Mode	**Domain**	**Task summary**	**Intervened step**	**Mean Δ**	**Rank**	**What changed (before $\rightarrow$ after)**
C-UoT	Social	Why do vending-machine choices become habitually locked onto familiar snacks despite healthier alternatives?	Final synthesis only (mapping frozen). Human rewrites the final from the frozen positive-feedback / path-dependence mapping; no new analogy search.	$+0.28$	$\#4 \rightarrow \#1.5$ (#1 on 3/4 seeds)	**Before:** generic "habit via cues/placement" without the mapped feedback loop or a discriminating test. **After:** keeps the frozen local positive-feedback story (repeated choice reinforces favorites and suppresses alternatives) and adds the mapped prediction—disrupting cues should raise switching, whereas merely placing healthier items near favorites should not. Mapping-token preserve rises $0.20 \rightarrow 0.33$.
T-UoT	Bio	Why are ponerine ants much richer in some regions—diversification-rate differences, occupancy time, or dispersal source–sink structure?	Post-transform synthesis. Transform rules stay frozen (relax fossil/dispersal assumptions); model must state one mechanistic consequence of those rules, then write the final.	$+0.25$	$\#7 \rightarrow \#2.5$	**Before:** retreats to data-limitation language (incomplete fossils, uneven sampling, barriers) and never uses the transform. **After:** applies the consequence of the relaxed rules—dispersal/colonization *sequence* imposes geographic constraints, so richness tracks early vs. late colonization timing rather than present-day diversification rates.

Both interventions improve the corresponding failure cases. For C-UoT, preserving and explicitly realizing the frozen analogy mapping raises the hypothesis from an average rank of #4 to #1.5. For T-UoT, forcing the model to derive a mechanistic consequence of the selected transformation improves the average rank from #7 to #2.5. These interventions are diagnostic rather than complete repairs, but they show that the creative operation itself can contain useful structure that is lost during downstream synthesis. This supports our failure analysis: a key limitation of C-UoT and T-UoT is not always failure to discover an analogy or transformation, but failure to faithfully integrate that creative move into the final grounded hypothesis.

Table 18: Inference cost vs. HypoArena quality (Qwen3-14B). Tokens/task: mean total tokens on a fixed 10-task cost-audit subset (4 Bio / 3 ML / 3 Social). Relative: vs. OnePrompt on the same subset. Avg. BTD and WR (%): full 100-task panel (Table 2). Bold: column best (lowest cost / highest quality).

Method	Tokens/task	Rel. cost	Avg. BTD	WR (%)
OnePrompt	3.7k	**1.0**×	1475.7	14.3
CoT	**3.5k**	0.9×	1493.7	18.8
ToT	27.1k	7.4×	1515.8	25.5
EGoT	101.6k	27.6×	1505.1	25.3
C-UoT	134.3k	36.5×	1489.6	19.5
E-UoT	36.7k	10.0×	**1535.4**	**32.7**
T-UoT	193.3k	52.5×	1484.6	17.3

D.10 Inference Cost versus Quality

A natural concern is that UoT gains might simply reflect larger inference budgets. Table 18 reports mean generation tokens per task on a fixed 10-task cost-audit subset (4 biomedical / 3 machine learning / 3 social science; Qwen3-14B), alongside the full 100-task BTD ratings and win rates from Table 2. Agentic methods are substantially more expensive than OnePrompt/CoT, as expected. However, quality does not increase monotonically with tokens: E-UoT achieves the best Avg. BTD (1535.4) at $10.0\times$ OnePrompt cost ($\approx 1.35\times$ ToT), whereas EGoT, C-UoT, and T-UoT require 27.6–$52.5\times$ OnePrompt cost (2.8–$5.3\times$ E-UoT) yet achieve lower BTD than E-UoT. Thus, on this backbone, search structure—not token count alone—appears to drive the HypoArena ranking.

E Experimental Design

E.1 Novel Creative-Reasoning Task Construction and Validation

Although the novel creative-reasoning tasks do not rely on external domain simulators (e.g., physics engines), they were constructed and validated through a human-in-the-loop process to ensure logical consistency and accessibility. We first generated ten candidate open-ended tasks, then manually solved each candidate to verify that the problem statement was coherent and that the solution space was reachable. From this pool we selected three diverse tasks (One-Lane Bridge, Electricity Tariff, and Social Cohesion) and refined their rules to balance assessable constraints against sufficient open-endedness for creative reasoning.

Our goal was to isolate creative reasoning rather than domain-expertise retrieval. Heavy domain-specific constraints that require complex engineering simulations would risk confounding the two. By restricting the tasks to self-contained logic settings validated by human solvers, solution validity is defined by explicit task constraints and objectives, which supports reproducible evaluation.

E.2 Benchmark Selection Rationale

As discussed in Section 2.3, existing creativity benchmarks tend to emphasize either highly open-ended generation or more tightly constrained problem solving. Open-ended ideation tasks permit diverse outputs but often make functional success difficult to assess objectively, whereas benchmarks such as MacGyver (Tian et al., 2024) and scientific hypothesis-generation tasks provide clearer problem definitions but impose stronger constraints on the admissible solution space. Our novel creative-reasoning tasks and HypoArena evaluation are chosen as complementary settings that jointly cover clear objectives with expansive solution spaces and scientifically grounded multi-hypothesis comparison.

E.3 HypoArena Protocol

HypoArena provides scientific hypothesis-generation cases grounded in source documents. We restrict evaluation to its three single-hypothesis research domains—biomedical science, machine learning, and social science. The remaining domains—financial analysis, IT operations, and safety

investigation—require multi-hypothesis submissions and are therefore outside the scope of our single-hypothesis generation setting. From the eligible research-domain pool, we evaluate a fixed panel of 100 tasks, stratified approximately evenly across domains (34 biomedical, 34 machine learning, and 32 social science). The same panel is held fixed across all methods and backbone variants.

Because different reasoning methods can produce outputs with substantially different levels of detail, we first canonicalize each final response into its main core hypothesis before pairwise evaluation. This ensures that comparisons focus on the underlying scientific idea rather than rewarding longer or more elaborated explanations.

HypoArena is not designed as a direct metric of creativity; its primary rubric dimensions emphasize Contextual Grounding and Evidential Justification. At the same time, the Inferential Insight dimension explicitly asks whether a hypothesis synthesizes context observations into a *non-obvious* but still context-constrained explanatory or mechanistic claim. Because judges therefore partially reward non-obvious scientific insight when it remains well supported, creative hypotheses can still receive high scores.

For each task, an LLM judge compares every pair of methods using bidirectional prompts and a five-level preference scale ranging from a clear preference for one hypothesis to a clear preference for the other, with ties permitted. Judgments follow the HypoArena rubric, emphasizing context grounding and analytical quality rather than stylistic fluency or verbosity.

We use Gemma-3-12B as the primary judge. Pairwise outcomes are aggregated with the Bradley–Terry–Davidson (BTD) model, initialized at 1500, and we additionally report pairwise win rate. BTD ratings are computed separately within each domain, and the overall score is the mean of the biomedical, machine-learning, and social-science ratings. Judge-consistency analyses under additional evaluators are reported in Appendix D.7.

E.4 Baselines

Our primary goal is to evaluate whether explicitly modeling creative reasoning provides benefits beyond general-purpose reasoning and search strategies. We compare UoT against the following established reasoning frameworks across both evaluation settings:

- *OnePrompt*: The task description is provided directly to the model without an explicit reasoning scaffold.
- *Chain-of-Thought (CoT)* (Wei et al., 2022): The model generates a sequential step-by-step reasoning trajectory.
- *Tree of Thoughts (ToT)* (Yao et al., 2023): The model explores multiple branching reasoning trajectories through iterative generation, evaluation, and selection.
- *Enhanced Graph of Thoughts (EGoT)* (Shin & Kim, 2025): A graph-based reasoning framework that supports generation, aggregation, and refinement of intermediate thoughts.

These baselines span increasingly structured forms of general-purpose reasoning, from direct generation to sequential, tree-based, and graph-based search. This allows us to test whether explicitly modeling creative reasoning provides benefits beyond increasing the complexity of generic reasoning search.

We intentionally focus on general-purpose reasoning frameworks rather than domain-specific scientific-discovery agents. UoT is designed as a domain-agnostic creative reasoning framework, and our HypoArena experiment tests whether the same creative reasoning procedures transfer to scientific hypothesis generation under matched reasoning conditions, rather than whether UoT outperforms specialized discovery systems that rely on additional retrieval pipelines, tools, or domain-specific assumptions.

E.5 Implementation Details

Novel creative-reasoning tasks. We use GPT-4o as the backbone for UoT and all reasoning baselines, holding the underlying model fixed so that comparisons more directly reflect differences among reasoning frameworks. We additionally report GPT-5 and DeepSeek V3.1 as contextual reference models; these are not treated as matched baselines because their internal reasoning procedures and computational budgets are not directly comparable.

HypoArena. Our primary experiments use Qwen3-14B as the shared backbone for UoT and all baseline reasoning methods. We use the same task inputs, backbone, and shared decoding settings where applicable. To assess whether the observed behavior depends on a particular backbone, we additionally evaluate selected settings with GLM-4-9B. Unless otherwise stated, HypoArena results refer to the Qwen3-14B experiments. We use open-weight models for HypoArena rather than GPT-4o because the 100-task, multi-method setup requires a large number of generations and pairwise judgments, making proprietary API evaluation prohibitively expensive under our budget.

F Novel Creative-Reasoning Task Evaluation Metrics and Prompts

Our three-task evaluation framework is guided by the classic definition of creativity (Runco & Jaeger, 2012), which comprises two primary components: Utility and Novelty. We also introduce Feasibility as a critical third metric to ensure solutions are valid and practical by adhering to all problem constraints.

A key challenge in evaluating LLM-generated outputs is their varying granularity. One model might offer a concise core idea, while another provides a detailed plan with sub-solutions. To ensure a fair comparison focused on the generated core creative insight, we introduce a Solution Canonicalization step. Before scoring, each generated solution is processed by an LLM (Gemini 2.5 Pro, temperature 0) to extract its single, core conceptual idea. This pre-processing step standardizes all outputs to a consistent format, enabling an equitable assessment of creativity. Following canonicalization, each solution is assessed using the metrics below:

- *Utility*: This metric measures how effectively the solution solves the problem. It is scored on a continuous scale from 0 to 1. An LLM evaluator places the solution on a 0–10 rubric relative to known solutions (and an optional optimal-solution upper reference), then rescales the score to $[0, 1]$.
- *Novelty*: To measure conceptual novelty, we first extract a short conceptual fingerprint ("theme") of the canonicalized solution with Gemini 2.5 Pro. We then embed that theme with `gemini-embedding-001` and compute its cosine distance from the embeddings of known solution concepts. The novelty score is $1 - \max(\text{similarity})$ to known concepts, normalized by the maximum known-to-optimal concept distance so that the optimal reference provides a practical upper bound on the scale.
- *Feasibility*: This metric is scored from 0 to 1 and assesses whether a solution is plausible and meets the rules of the problem. The LLM evaluator is provided with a pre-defined checklist of all explicit constraints and scores the solution based on adherence to those constraints.

The final *Creativity Score* is the mean of the Utility and Novelty scores. Feasibility is reported separately as a validity check.

F.1 Known Solutions for Novelty and Utility

For each novel creative-reasoning task, we maintain a fixed set of *known solutions*: author-curated conventional baselines that represent standard approaches to the problem, not outputs of any evaluated model. Novelty compares each candidate against short conceptual fingerprints of these baselines (`known_solutions_concept`); Utility uses the corresponding fuller descriptions as calibration anchors. The baselines used in our experiments are:

Bridge. (1) Fixed time blocks (alternate direction on a timer). (2) Queue-based switching (switch when the active queue is empty). (3) Batch processing (serve a fixed batch size before switching).

Electricity. (1) Time-of-use pricing with static peak/off-peak blocks. (2) Critical peak pricing on declared stress events. (3) Peak-time rebates for voluntary load reduction.

Society. (1) Structured mixers that force cross-group interaction. (2) Cross-group micro-grants requiring multi-group teams. (3) Randomized one-on-one "coffee chat" pairing. (4) A shared resource library that creates interdependence.

Human and automated evaluations use the same Utility and Feasibility prompts. Automated evaluation uses Gemini 2.5 Pro with temperature set to 0. The prompts below are the templates used in our implementation; placeholders such as `[SOLUTION]`, `[TASK_DESCRIPTION]`, `[KNOWN_SOLUTIONS]`, `[OPTIMAL_SOLUTION]`, and `[CHECK_POINTS]` are filled per task and candidate.

F.2 Canonicalization Prompt

```
You are a solution processing assistant. Your task is to identify
all individual solutions from the text provided and canonicalize
each one into a concise paragraph that functions as an operational
blueprint.

Text to Process:
[SOLUTION]

Instructions:
1. Scan the text and identify the boundaries of each distinct
   solution.
2. For each solution, author a self-contained paragraph describing
   its core functional mechanism. This paragraph must identify the
   essential components, their primary inputs and outputs, and
   explain how they interact to produce the final result.
3. If an inspirational domain is mentioned, include it for context.
4. The level of detail must be sufficient to make the process
   unambiguous and reproducible in principle, but should not be a
   granular, step-by-step implementation guide. Omit low-level
   details like specific parameters or mathematical formulas.

Return your response as a single, valid JSON array of strings,
where each string is a descriptive paragraph.
```

F.3 Feasibility Prompt

```
Evaluate the feasibility of the following solution against these
check points:

Solution: [SOLUTION]

Check Points:
[CHECK_POINTS]

Return a JSON object with two fields:
- "score": a number between 0.0 and 1.0 representing the
  feasibility score, where:
  - 0.0 = None of the check points are met
  - 1.0 = All check points are fully met
  - Values in between represent partial fulfillment
- "reasoning": a 1-2 sentence explanation of your evaluation

Please score it on the assumption that the underlying technology
and strategy are sound.
```

F.4 Utility Prompt

```
You are an expert systems evaluator. Your task is to assess the
utility of the "Current Solution" by placing it on a percentile
scale relative to the provided "Known Solutions."
```

```
Optimal Solution Reference (Score 10 - Maximum Utility):
[OPTIMAL_SOLUTION]

Task Description:
[TASK_DESCRIPTION]

Known Solutions (Consider these a representative sample of
standard, conventional approaches):
[KNOWN_SOLUTIONS]

Current Solution to Evaluate:
[SOLUTION]

Instructions:
1. Analyze the Current Solution's effectiveness at achieving the
   primary objective in the Task Description.
2. Compare this effectiveness to the quality of the Known Solutions.
3. Place the Current Solution on a scale from 0 to 10 based on the
   rubric below. A score of 5 means it is a minor, incremental
   improvement over the average known solution.

Evaluate the effectiveness of the core idea itself, not the quality
of its description or implementation plan.

Scoring Rubric (Absolute Improvement Scale):
* 0-2 (Incremental): A minor or trivial improvement over the
  baseline. The core strategy is still conventional and offers only
  marginal gains.
* 3-5 (Competent): A solid, well-defined solution that is clearly
  more effective than the baseline. It represents a competent, but
  not necessarily innovative, approach.
* 6-8 (Significant): A significant conceptual or practical
  improvement. The strategy is advanced, highly effective, and
  demonstrates a sophisticated understanding of the problem's
  dynamics.
* 9-10 (Transformative): A state-of-the-art or paradigm-shifting
  approach. The solution reframes the problem or introduces a
  mechanism that is fundamentally superior to conventional
  strategies.

Return a single JSON object with two fields:
- "score": A single integer from 0 to 10 representing the placement.
- "reasoning": A 1-2 sentence justification for the score,
  explicitly referencing the rubric and comparing the Current
  Solution to the Known Solutions.
```

F.5 Novelty Theme-Extraction Prompt

```
Deconstruct the following solution into its core operational logic
to create a "conceptual fingerprint." Your goal is to create a
single, dense phrase that describes the solution's fundamental
system of operation, ignoring prose and justifications.

To do this, you will:
1. Identify the Core Mechanism: The central process or tool
   (e.g., a timer, an auction, a sensor).
2. Identify the Actors & Interaction: The key components and what
   they do (e.g., vehicles submit bids).
```

```
3. Identify the Governing Principle: The underlying rule that
   dictates the outcome (e.g., fixed intervals, willingness to pay).
4. Synthesize: Combine these elements into a single, logical phrase
   that describes the system.

Crucial Rule: Do NOT summarize into high-level keywords. The output
must describe the unique process or system rule.

Solution to Analyze:
[SOLUTION]

Return format:
"A system where [actors] [interaction] based on the principle of
[governing principle]."
```

After theme extraction, Novelty is computed with `gemini-embedding-001` embeddings as described above (cosine distance to known concepts, normalized by the known-to-optimal concept distance).