

CreativityBench: Evaluating Creative Reasoning via Affordance-Based Tool Repurposing

Cheng Qian^{*1}, Hyeonjeong Ha^{*1}, Jiayu Liu¹, Bingxiang He, Jeonghwan Kim¹, Jiateng Liu¹, Bingxuan Li¹, Aditi Tiwari¹, Dwip Dalal¹, Zhenhailong Wang¹, Xiusi Chen¹, Mahdi Namazifar², Yunzhu Li³, Heng Ji¹

¹UIUC, ²Amazon, ³Columbia

Abstract

Recent advances in large language models (LLMs) have enabled strong performance in reasoning and environment-interaction, yet their ability for creative problem solving remains underexplored. One core form of such creativity is *tool use*: repurposing objects by reasoning about their affordances and attributes rather than relying on canonical usage. To study this capability, we introduce **CreativityBench**, a benchmark for affordance-based creative tool use. It is built on a large-scale affordance knowledge base containing 4K entities and 150K+ annotations linking objects, parts, attributes, and uses, from which we generate 14K grounded tasks requiring non-obvious but physically plausible solutions under constraints. Evaluation across 10 closed and open-source LLMs reveal a consistent gap: while models often select plausible objects, they fail to identify the correct parts, affordances, and underlying physical mechanism. Moreover, general reasoning ability does not reliably transfer to creative affordance discovery, gains from scaling quickly saturate, and Chain-of-Thought yields limited benefit. These findings highlight creative tool use as a missing dimension in current LLMs capabilities, and position CreativityBench as a systematic testbed for advancing creative intelligence in future LLMs and embodied agents.

1 Introduction

The **Triarchic Theory of Intelligence** (Sternberg, 1997) decomposes human intelligence into analytical, practical, and creative components. This framework helps contextualize progress in large language models (LLMs), which has been concentrated in the first two dimensions. Recent LLMs show strong **analytical intelligence**, including logical deduction, mathematical reasoning, and coherent chains of thought, as reflected in standard reasoning and mathematics benchmarks (Hendrycks et al., 2020; Cobbe et al., 2021; Wei et al., 2022). In parallel, they have made progress in **practical intelligence**, acquiring the ability to interact with tools, web browsing, software interaction, and execute long-horizon tasks in simulated or embodied environments (Wei et al., 2025; Mialon et al., 2023; Froger et al., 2025). However, these advances leave a critical dimension underexplored: **creative intelligence**.

Creative intelligence involves **producing novel yet effective solutions under constraints** (Runco & Jaeger, 2012; Sternberg & Lubart, 1999). For LLMs, especially agentic models, this goes beyond creative writing or stylistic generation to creative problem solving in the physical and functional world. Humans show this ability by repurposing available objects in non-obvious ways, for example, using the sharp tip of a key to open a sealed box when scissors are unavailable. Such behavior relies on recognizing that objects can be used beyond their intended use by leveraging their underlying properties. We argue that a core form of creative intelligence is **creative tool use**: the ability to **solve a task by identifying a tool’s underlying affordances and applying them to achieve a goal in a novel way**.

^{*}Equal contribution.

Benchmark	Creative Tool Use	Affordance Grounding	Attribute Grounding	Part-Level Reasoning	Fine-Grained Creativity Levels	Distractors Included
PROST (Aroca-Ouellette et al., 2021)	✗	✓	✓	✗	✗	✓
NEWTON (Wang et al., 2023)	✗	✗	✓	✗	✗	✗
Creation-MM Bench (Tian et al., 2024)	✗	✗	✗	✗	✗	✗
Villager Bench (Dong et al., 2024)	✗	✓	✗	✗	✗	✗
VisEscape (Lim et al., 2025)	✗	✓	✗	✗	✗	✓
PIQA (Bisk et al., 2020)	✓	✓	✓	✗	✗	✓
MacGyver (Tian et al., 2024)	✓	✓	✗	✗	✗	✓
Escape Bench (Qian et al., 2024)	✓	✓	✗	✗	✓	✓
CreativityBench (Ours)	✓	✓	✓	✓	✓	✓

Table 1: For each existing benchmark, the table indicates whether the corresponding dimension is fully addressed (✓), partially addressed (✓), or not addressed (✗).

Under this view, creativity is not random exploration or hallucinated action, but discovering a non-obvious functional connection grounded in physical reality. Tools matter not only for their intended purpose, but for the actions they afford, which arise from their state and physical attributes. Creative reasoning therefore requires divergent thinking while staying grounded in physical constraints, often by recombining existing knowledge and prior tool-use experience into a new solution. For example, in Figure 1, a key is designed for unlocking doors, but its rigidity and sharp tip also allow prying or cutting. It can therefore open a box because its *attribute* (e.g., tip of the key) supports the needed *affordance* (e.g., cutting).

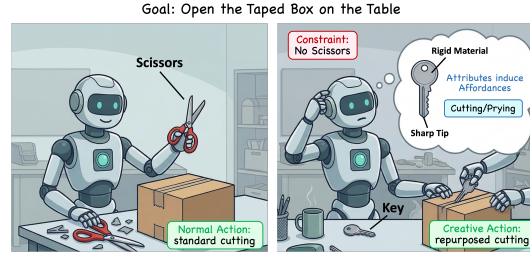


Figure 1: Creative tool use as affordance-based tool repurposing.

Despite its importance, creative tool use remains weakly measured in existing evaluations. Recent benchmarks explore related aspects such as creative investigation, tool use, multimodal interaction, and embodied exploration (Qian et al., 2024; Tian et al., 2024; Fang et al., 2025; Lim et al., 2025; Dong et al., 2024), but largely emphasize planning, execution, or coarse object-level plausibility, rather than whether models can identify concrete, part-level affordances and repurpose tools based on physically usable properties. Our preliminary findings also suggest that this gap is not resolved by simply enforcing more structured reasoning: even when guided to decompose tools into parts, infer physical properties, and reason step by step, strong models show only limited gains (Section 3). These gaps motivate three research questions: i) how to build a scalable and physically grounded affordance knowledge base for training and evaluation, ii) how to design a rigorous benchmark for affordance-based creative tool use beyond object-level reasoning, and iii) how current SOTA models perform on this form of creative intelligence.

To address these questions, we introduce **CreativityBench**, the first large-scale benchmark for systematically evaluating creative tool use through affordance-based creative tool use. It is built on a scalable annotation pipeline that constructs a structured affordance knowledge base (KB) linking objects, parts, attributes, and affordances, grounding judgments in concrete physical properties rather than coarse semantic matching alone. The resulting KB contains 4K entities and 150K+ affordance annotations, which we use to generate 14K diverse, physically grounded tasks by reverse-engineering solution trajectories, ensuring that each task requires non-obvious yet feasible affordance reasoning.

Experiments on ten proprietary and open-source models reveal several key findings. **First, physical grounding is a major bottleneck:** models often identify a plausible object but fail at the part or attribute level, causing performance to drop by over 60%. **Second, strong analytical reasoning does not imply creative affordance discovery:** models that excel on general reasoning like GPT-5.2 can still underperform open-source models on novel tool discovery. **Third, scaling and standard inference-time interventions offer limited benefit:** performance saturates quickly, remains constrained by affordance familiarity, and shows minimal improvement from strategies such as higher sampling temperature, structured Chain-of-Thought, or interactive evaluation. Overall, our contributions are threefold:

- **Affordance KB:** We build a large-scale KB with 4K entities and 150K+ affordances, providing reusable building blocks for grounded task generation, training, and evaluation.
- **CreativityBench:** We introduce an affordance-grounded benchmark for evaluating creative tool use under rigorous and reproducible protocols.

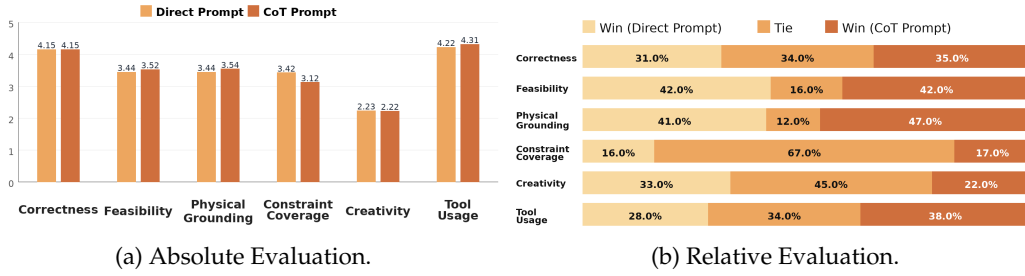


Figure 2: **Preliminary experiment results** between direct prompting and structured affordance-level CoT on creative tool use tasks.

- **Empirical Analysis:** We conduct a systematic study of creative tool use across models, identifying key limits in physical grounding, affordance discovery, and inference scaling.

2 Related Works

Creativity in Language Models. Creativity is a hallmark of human intelligence and has increasingly been studied in LLMs. Recent work has examined creative generation in storytelling and poetry (Akoury et al., 2020; Brown et al., 2020), tool and system design (Qian et al., 2023; Cai et al., 2023; Ha et al., 2025; Wang et al., 2026), scientific ideation and discovery (Si et al., 2024; Wang et al., 2024), and human brainstorming support (Qian et al., 2025). Another line of work evaluates LLM creativity using adapted psychological tests of fluency, originality, and flexibility (Guilford, 1967; Boden, 1998), though these evaluations are often prompt-sensitive and hard to scale. Creativity has also been studied in problem-solving settings, including object repurposing, escape and exploration tasks, and multimodal or embodied environments (Tian et al., 2024; Qian et al., 2024; Fang et al., 2025; Lim et al., 2025; Dong et al., 2024). However, these benchmarks are generally scenario-driven and focus more on planning, reasoning, or multimodal interaction than on the fine-grained mechanism of creative tool use grounded in object affordances.

Affordance and Physical Reasoning. A related line of work studies physical reasoning and affordance knowledge in AI systems. Benchmarks such as PIQA, PROST, and NEWTON evaluate physical commonsense, object attributes, and simple affordances (Bisk et al., 2020; Aroca-Ouellette et al., 2021; Wang et al., 2023), while robotics and embodied-agent research has long modeled affordances as links between perception and action for manipulation and planning (Brohan et al., 2022; 2024; Chu et al., 2019; Montesano et al., 2008; Jamone et al., 2016). More structured representations have also been explored: SYNTHIA, for example, organizes objects into parts and associated affordances (Ha et al., 2025). Yet existing work typically focuses on canonical object use, physical knowledge, or part-affordance associations, rather than on how affordances arise from concrete physical attributes of object parts. Our work positions itself at this intersection by studying creative tool use through affordance-grounded reasoning over entities, parts, attributes, and affordances.

3 Preliminaries: Structured Reasoning Is Not Enough for Creativity

To examine the gap between analytical/practical intelligence and creative intelligence in LLMs, we run a controlled study on 100 creative tool-use tasks sampled from MacGyver (Tian et al., 2024). We compare two prompting strategies: a **direct prompt**, where the model generates a feasible solution under task constraints without prescribed steps, and a **structured affordance-level Chain-of-Thought (CoT)**, where it explicitly lists tools, decomposes parts, infers physical properties, derives affordances, justifies each step, and checks constraints. This comparison tests whether failures in creative tool use stem from missing procedural guidance or from deeper limitations in grounded affordance modeling and recombinational creativity. Detailed prompts and cases are provided in Section B and Figure 11. We use GPT-4.1-mini as the target LLM and GPT-5.2 as the judge model.

We evaluate model outputs along six dimensions: **Correctness**, **Feasibility**, **Physical Grounding**, **Constraint Coverage**, **Tool Usage**, and **Creativity**. As shown in Figure 2, structured CoT yields only modest gains on procedural dimensions in absolute evaluation,

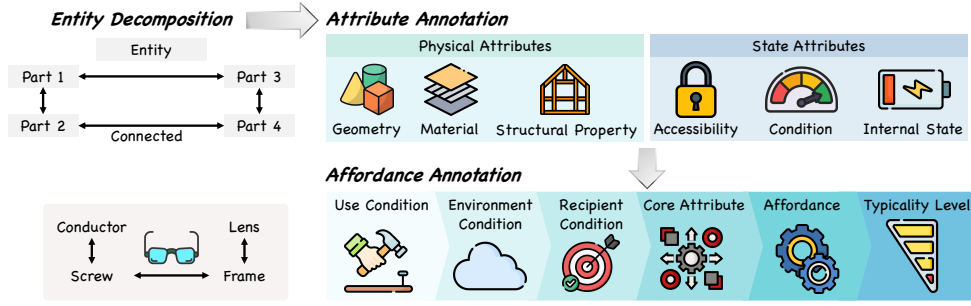


Figure 3: Annotation pipeline for affordance knowledge base construction.

improving Feasibility from 3.44 to 3.52, Physical Grounding from 3.44 to 3.54, and Tool Usage from 4.22 to 4.31; relative evaluation shows a similar pattern, with CoT winning more often on Physical Grounding (47% vs. 41%) and Tool Usage (38% vs. 28%). However, it performs worse on Creative Reasoning, suggesting that structured guidance may improve grounding while constraining divergent thinking. Overall, these results suggest that the main limitation is not missing reasoning structure, but the lack of grounded affordance knowledge that can be flexibly recombined. They also highlight the limits of existing resources such as MacGyver, whose tasks are not explicitly organized around affordance structure and whose evaluation relies heavily on LLM-as-judge scoring. This motivates our next step: building an explicit affordance knowledge base and **CreativityBench**, a fine-grained, part-level, attribute-grounded benchmark for creative tool use.

4 CreativityBench

Creative tool use provides a concrete mechanism for studying creative intelligence in LLMs. A tool is intrinsically defined by its **affordances** (the actions it enables), rather than by its name or intended category. These affordances arise from underlying **attributes**, such as structure, material properties, or accessibility. Creative tool use therefore requires identifying which attributes enable useful affordances and how they can be combined to achieve a goal, rather than matching tasks to tools based on semantic labels. This framing guides our benchmark design. We first build an affordance knowledge base that links entities, parts, attributes, and affordances through a *top-down* annotation pipeline. Formally, for each entity $e \in \mathcal{E}$, the pipeline follows $e \rightarrow P(e) \rightarrow A(p) \rightarrow F(p)$, where $P(e)$ denotes parts of e , $A(p)$ the attributes of each part p , and $F(p)$ the affordances enabled by p . We scale annotation with an LLM-assisted pipeline and generate tasks by reverse-engineering high-level goals from lower-level affordances, as shown in Figure 3.

4.1 Knowledge Base Construction

Our knowledge base is designed to represent affordances as context-dependent, part-level action possibilities rather than fixed object-level labels, following the ecological view of affordances (Gibson, 2014). Accordingly, we organize each entity into parts, attributes, states, and affordances so that the physical basis and applicable conditions of a potential use can be made explicit. This structured representation provides the foundation for grounded task generation, controlled sampling, and fine-grained evaluation.

Entity Decomposition. We start from common entities in eight in-house scenes (e.g., kitchen, living room, bedroom), grounded in ConceptNet 5.5 (Speer et al., 2017). Each entity e is decomposed into non-overlapping parts, $P(e) = \{p_1, p_2, \dots, p_n\}$, under three constraints: **completeness** ($\bigcup_i p_i = e$), **non-overlap** ($p_i \cap p_j = \emptyset$ for $i \neq j$), and **functional granularity**, meaning each part is a structurally or functionally meaningful unit. This part-level decomposition is important because creative tool use often depends on **local structure** rather than an object’s intended global function.

Attributes and Affordances. For each part $p \in P(e)$, we annotate attributes $A(p) = A_p(p) \cup A_s(p)$, where **physical attributes** $A_p(p)$ capture intrinsic properties such as shape, size, material, and mass, while **state attributes** $A_s(p)$ capture interaction-dependent properties such as accessibility, condition, and internal state. We make this distinction because affordances depend not only on what a part physically *is*, but also on what state it is cur-

Category	Statistic	Value	Category	Statistic	Value
Scale	Number of entities	3,816	Annotation counts	Physical attributes annotated	288,318
	Number of parts	26,238		State attributes annotated	124,972
	Total annotations	570,717		Affordances annotated	157,427

Table 2: Overall statistics of the affordance knowledge base containing eight different household scenes, with annotations covering entities, attributes, and affordances.

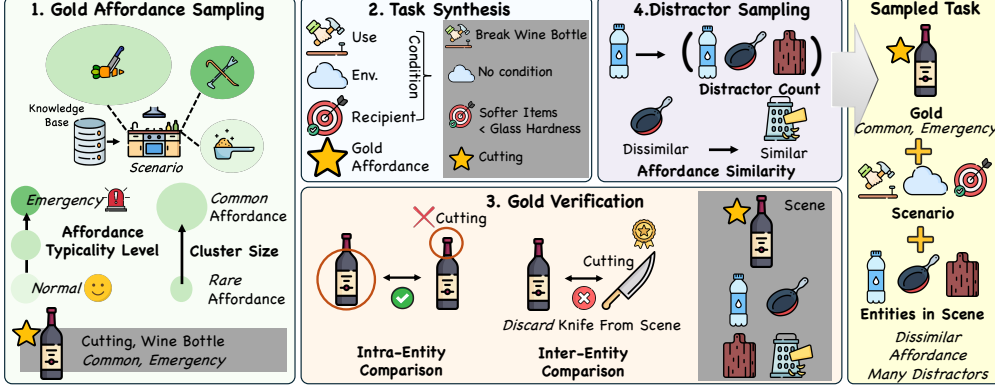


Figure 4: The four-stage pipeline for CreativityBench task sampling.

rently *in*. Different attribute combinations can produce different **variants** of the same entity, which we treat as distinct instances because their induced affordances differ.

We then annotate part-level affordances $F(p) = \{f_1, f_2, \dots, f_m\}$. Each affordance is represented as $f = (a, C_u, C_e, C_r)$, where a is the enabled action, and C_u , C_e , and C_r denote the **use condition**, **environment condition**, and **recipient condition**, respectively. This formulation makes the affordance executable by specifying not only *what* a part can do, but also *when* and *to what* it can apply. We additionally annotate enabling attributes, potential recipients, and failure conditions. Finally, we distinguish **normal affordances**, corresponding to intended or commonly observed functions, from **emergency affordances**, which capture unconventional but plausible repurposing. Each emergency affordance is assigned a typicality level $l \in \{1, \dots, 5\}$ that reflects how natural and practically usable the repurposing is under everyday or emergency conditions. Lower levels correspond to less common uses that may require substantial preparation or transformation of the part, whereas higher levels indicate more natural and readily applicable repurposing.

Knowledge Base Statistics. To scale annotation, we implement the hierarchy **entity** $\rightarrow$ **part** $\rightarrow$ **attributes** $\rightarrow$ **affordances** with GPT-5.2. The statistics of the created KB are summarized in Table 2 showing that it provides sufficient coverage and granularity to support large-scale grounded task construction. More annotation details are provided in Section C.

4.2 Benchmark Task Sampling

Task Description. We construct benchmark tasks from the affordance knowledge base through a **reverse-engineering** process. Instead of starting from a task and searching for a tool, we begin with a known affordance and synthesize a scenario where discovering that affordance is the intended solution. This bottom-up design gives each task a clear ground-truth reasoning path while still requiring the model to infer the relevant affordance from object attributes alone. Let $\mathcal{F} = \bigcup_p F(p)$ denote all annotated affordances. A task is defined as $T = (S, \mathcal{E}_T, g)$, where S is the task scenario description, $\mathcal{E}_T \subset \mathcal{E}$ the entities present in the scene, and $g = (e^*, p^*, f^*)$ the **gold affordance**. The model observes only the entities and their attributes, and must recover the correct entity, part, and affordance.

Construction Stages. The full four-stage task sampling pipeline is illustrated in Figure 4.

- **Gold affordance sampling.** Directly sampling from $\mathcal{F}$ creates strong redundancy, so for each one of the eight scenes s we first cluster all its affordances $\mathcal{F}_s \subset \mathcal{F}$ as $\mathcal{F}_s = \bigcup_{k=1}^{K_s} \mathcal{C}_k$, where each cluster $\mathcal{C}_k$ contains semantically similar affordances. We use cluster size as a **frequency-based commonality** measure: a larger cluster indicates that the corresponding type of affordance occurs more frequently in our affordance KB and is therefore more

common within the KB. We sample f^* while controlling both **cluster size** and **affordance typicality level**, yielding a balanced mixture of common and long-tail affordance types as well as more- and less-natural repurposing cases.

- **Task synthesis.** Given $g = (e^*, p^*, f^*)$ with $f^* = (a, C_u, C_e, C_r)$, we generate a grounded natural-language scenario description S such that $S \Rightarrow (C_u \wedge C_e \wedge C_r) \wedge a$. The scenario description makes the goal and constraints explicit while withholding the affordance itself, so the model must infer it from the attributes of the entities in the scene.
- **Gold verification.** Because creative tasks may admit multiple plausible solutions, we verify **solution uniqueness** through both intra-entity and inter-entity comparison. For the gold entity e^* , we compare f^* against all $\tilde{f} \in F(p)$ for $p \in P(e^*)$; for each distractor candidate $\tilde{e} \in \mathcal{E}_T$, we compare f^* against affordances $\tilde{f} \in F(p)$ for $p \in P(\tilde{e})$. If some alternative affordance is judged preferable, the candidate is rejected or removed. For every retained alternative, we additionally store an attribute-grounded rationale explaining why it is less suitable than the gold solution.
- **Distractor sampling.** After verification, we sample distractors from $\mathcal{E}_T \setminus \{e^*\}$ by controlling **distractor count** and **affordance similarity**, so tasks require fine-grained attribute-level reasoning rather than superficial matching.

CreativityBench Statistics. We scale the full pipeline with GPT-5.2, producing about 14K tasks spanning diverse scenarios, affordance types, and distractor configurations. The dataset is approximately balanced across the main sampling factors, enabling controlled analysis of model behavior. More details are provided in Section [D](#).

5 Experiments

5.1 Experimental Settings

Evaluation Setup. We evaluate a diverse set of open- and closed-source models including GPT ([Singh et al., 2025](#)), Gemini ([Comanici et al., 2025](#)), Qwen ([Yang et al., 2025](#)), Llama ([Grattafiori et al., 2024](#)), and Mistral ([Liu et al., 2026](#)). Later analyses focus mainly on GPT and Qwen. All models are evaluated on the full 14K-task benchmark. In the main setting, each model is given the task description, all entities in the scene, and the attributes of their parts, and must identify the correct entity and part and explain how to use them to solve the task. Additional settings and inference details are provided in Section [E](#).

Metrics. We report two objective selection metrics: **Entity Correct Rate**, the fraction of cases where the correct entity is selected, and **Gold Correct Rate**, the fraction where both the correct entity and part are selected. Their gap captures the difficulty of fine-grained part-level grounding. For gold-correct cases, we further use an LLM judge for a 1-5 Likert scale to evaluate predicted usage on **constraint coverage**, **physical grounding**, **action feasibility**, and **overall correctness**. As each gold affordance includes multiple constraints (see Section [4](#)), we also decompose constraint coverage into **use condition** (C_u), **environment condition** (C_e), and **recipient condition** (C_r). We use Gemini-3.1-Flash-Lite as the judge model for its instruction following and cost-effectiveness. The evaluation is constrained by explicitly defined score criteria and uses the annotated gold solution as a comparison anchor, rather than relying on unconstrained subjective assessment. More details are provided in Section [E](#).

5.2 Main Results

We present the main results in Table [3](#) and summarize the key findings below.

Part-Level Grounding Is The Main Bottleneck. Models can often identify a plausible tool entity, but much less often the correct part. Averaged across models, **Entity Correct Rate** reaches 0.5149, while **Gold Correct Rate** drops to 0.1910, a relative decrease of over 60%; for some models, the gap is as large as 0.35. This indicates current models often know *what object* might help, but fail to localize *which part* enables the target affordance, limiting truly grounded creative tool use.

Model	Tool Usage		Constraint Coverage			Physical Grounding	Action Feasibility	Prediction Correctness
	Gold Correct	Entity Correct	Use (C_u)	Env. (C_e)	Rcpt. (C_r)			
Closed-Source Models								
GPT-5.2	0.1819	0.5210	<u>3.8452</u>	4.2428	<u>4.1458</u>	<u>3.8734</u>	4.3867	<u>3.8696</u>
GPT-5 Mini	0.1687	0.4856	3.8919	<u>4.1947</u>	4.1566	3.8847	<u>4.2607</u>	3.8704
GPT-5 Nano	0.1192	0.5721	2.7535	2.9751	3.0903	2.8513	3.2628	2.8230
Gemini-2.5-Pro	0.1670	0.3552	3.3668	2.9963	2.6502	3.2735	3.6728	3.0495
Gemini-2.5-Flash	0.1532	0.3694	3.7505	3.8405	3.2353	3.6200	3.9844	3.4606
Open-Source Models								
Qwen3-32B	0.2588	0.6246	2.8548	2.8187	2.3335	2.9670	3.2323	2.6451
Qwen3-14B	<u>0.2483</u>	<u>0.6141</u>	2.9569	2.9557	2.2771	3.1194	3.4138	2.7441
Qwen3-4B	0.1882	0.5277	2.5943	2.4151	1.8961	2.8101	2.9881	2.4838
Llama-3-70B	0.2151	0.5532	2.7487	2.6138	1.9710	2.7454	3.4761	2.7180
Ministral-3-14B	0.2091	0.5263	3.0172	2.8073	2.2699	2.8581	3.1819	2.6627
Average	0.1910	0.5149	3.1780	3.1860	2.8026	3.2003	3.5860	3.0327

Table 3: Evaluation summary across models on the judged outputs. Best and second-best results are highlighted in **bold** and underline respectively. For all metrics other than Tool Usage, the maximum score is 5.0. Higher is better for all metrics.

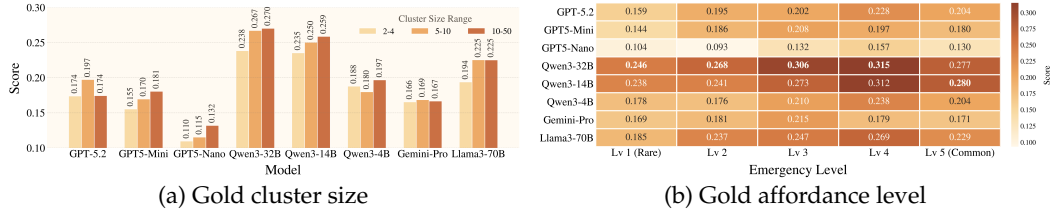


Figure 5: **Impact of gold commonality.** Models perform better when the gold affordance comes from a larger cluster or has a higher affordance level.

Plausible Actions Often Lack Physical Grounding. Predicted actions are often superficially reasonable, but their underlying physical justification is weaker. Across models, **action feasibility** averages 3.5860, compared with 3.2003 for **physical grounding**. A similar pattern appears in constraint coverage: **use** and **environment** conditions score 3.1780 and 3.1860 on average, while **recipient conditions** are lowest at 2.8026. This suggests that models can generate plausible action plans, yet often miss the concrete physical and recipient-side conditions needed for successful execution.

General Reasoning Strength Does Not Guarantee Creative Affordance Discovery. Strong general models do not always perform best on creative tool use. GPT-5.2 and GPT-5-Mini lead on several reasoning-related metrics, but their gold correctness (0.1819) is well below Qwen3-32B (0.2588) and Qwen3-14B (0.2483); Qwen3-32B also achieves the highest entity correctness at 0.6246. This gap suggests a meaningful split between **discovering** non-obvious affordances and **evaluating** them once identified.

Scaling Yields Diminishing Returns. We further validate the judge on 200 responses randomly sampled and balanced across models and evaluation settings. Under the same evaluation instructions, the original judge achieves 90.4% agreement with human judgments, while inter-human agreement is 87.0% and agreement with an additional GPT-5.2 judge is 93.4%. Moreover, Gemini-family models do not receive the highest objective scores, providing no evidence that the selected judge systematically favors its own model family. We therefore treat the judge-based scores as reliable auxiliary diagnostics, while grounding the paper’s primary conclusions in the objective selection metrics.

Reliability of LLM-as-Judge. Performance improves with scale at smaller sizes, but quickly saturates. In the Qwen family, gold correctness rises from 0.1882 (4B) to 0.2483 (14B), but only to 0.2588 at 32B. A similar pattern holds for proprietary models: GPT-5 Nano to GPT-5 Mini improves from 0.1192 to 0.1687, while GPT-5 Mini to GPT-5.2 yields only a small further gain. This suggests that creative tool use is not solved by scale alone, but instead depends on stronger affordance grounding and creative physical reasoning.

5.3 Analysis

We analyze model behavior from three complementary perspectives: **the commonality of the gold solution**, **the severity of distraction**, and **inference-time strategies**. Across these,

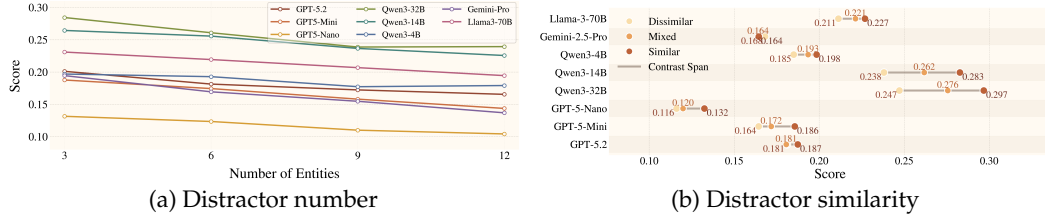


Figure 6: **Impact of distraction severity.** Increasing distractor number consistently reduces accuracy, but distractors with similar affordances often lead to higher performance.

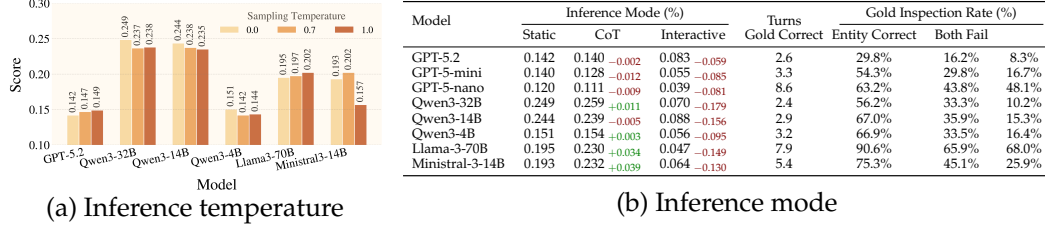


Figure 8: **Impact of inference-time strategies.** Higher sampling temperature generally hurts performance. Interactive evaluation lowers performance across all models relative to static evaluation, while structured CoT provides only limited gains. For interactive mode, we also report the average number of turns and the **Gold Inspection Rate**: the fraction of tasks where the gold entity is inspected before the final answer.

we focus on the **Tool Usage** metrics as they provide the most objective evaluation signal. Detailed setups and extended results are deferred to Section H.

How does Gold Commonality Affect Performance? As shown in Figure 5, performance improves consistently when the gold affordance comes from a **larger cluster** (more common) and when the sampled affordance has a **higher emergency level** (more natural). These two views tell the same story: models are much better at recovering *more familiar* or *more natural* repurposed uses, but struggle on **rare, long-tail affordances**. The key implication is that current failures are not merely about choosing among candidates; they reflect a deeper weakness in **recognizing unconventional yet physically plausible uses of objects** when those uses are less common in data or human intuition.

How Does Distraction Severity Affect Performance? Figure 6 shows that performance generally drops as the number of distractors increases, indicating that a larger candidate set creates a real **comparison burden** for current models. However, distractor *type* matters as much as distractor count: somewhat counterintuitively, tasks with **similar-affordance distractors** often outperform those with dissimilar ones, and the fine-grained results in Figure 7 suggest why. Similar distractors do not simply add confusion; they can also **expose the relevant affordance dimension**, making the correct reasoning space easier to identify. This suggests that the main challenge is not generic distraction alone, but whether the environment helps the model attend to the *right physical attributes* for comparison.

How Do Inference Settings Affect Performance? The results in Figure 8 show that changes in inference-time strategies yield only limited gains. Increasing sampling temperature does not reliably help: smaller models generally worsen, while larger models show at most mild improvement. This is also validated by our empirical finding that higher temperature more often increases **hallucinated exploration** than genuinely better affordance discovery. Changing the evaluation mode is similarly unhelpful. Interactive evaluation substantially reduces performance and, together with the inspection statistics from interactive mode, reveals a clear failure mode: models **explore too little, commit too early, and often fail**

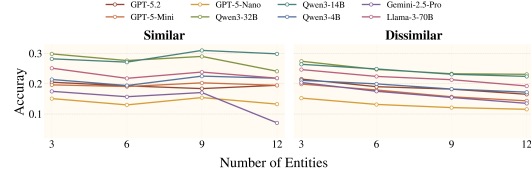


Figure 7: **Fine-grained trends across distractor types.** When distractors share similar affordances with the gold, performance does not monotonically decline with more entities.

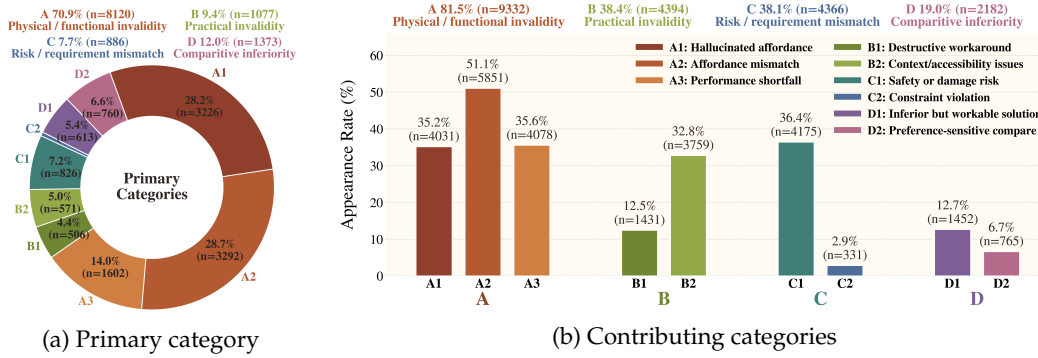


Figure 9: **Model failure attribution categories.** Physical invalidity dominates in both views, with affordance mismatch emerging as the most common fine-grained cause.

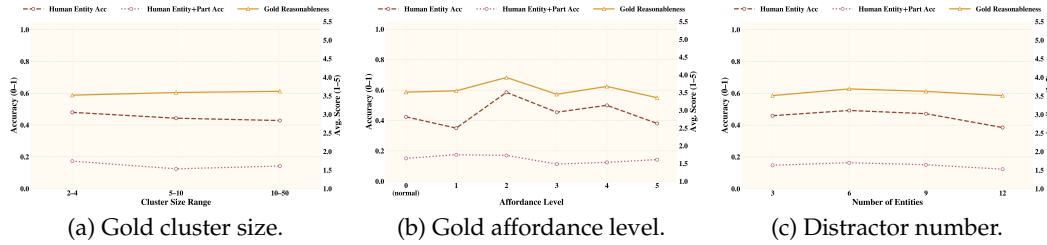


Figure 10: **Human performance results.** Human performance is less sensitive than model performance to gold cluster size, distractor count, and gold affordance level.

to inspect the gold object at all. Structured CoT produces only minor changes, implying that the bottleneck is not the lack of a structured reasoning, but the lack of **grounded comparative search** over candidate tools.

5.4 Attribution Study

To understand *why* model predictions are preferred less often than the gold solution, we conduct an attribution analysis on a random 10% sample of failure cases for each model, using Gemini-3.1-Flash-Lite to categorize the failure reasons. As summarized in Figure 9, the clearest pattern is that **physical invalidity** dominates across models, both as the primary reason and as a frequent contributing factor. In particular, **affordance mismatch** is the most prominent subtype, showing that models often select objects whose geometry, material, or mechanics do not actually support the intended function. This again indicates that the main bottleneck is not simply weak planning, but weak **physical grounding** in the properties that make a candidate tool genuinely usable.

A second recurring pattern is **hallucinated affordances**: models frequently attribute capabilities to objects that they do not realistically have. By comparison, practicality, safety, and other real-world constraints appear as *secondary* factors, while purely comparative inferiority is less common. Taken together, these results suggest that most failures are not near-miss alternatives, but stem from a deeper tendency to reason at the level of *abstract function* without adequately verifying whether the chosen object can physically realize that function. Additional details and model-wise breakdowns are deferred to Section H.

5.5 Human Study

To examine how humans solve these tasks and validate the quality of the annotated affordance KB, we conduct a human study on 100 sampled tasks balanced across key factors, including entity count, affordance similarity, and affordance typicality. We recruit 10 annotators with STEM backgrounds and adopt a two-stage protocol: **problem solving**, in which annotators select the entity and part that best satisfy the task based on the provided attribute

Tool Usage		Gold Solution Persuasiveness	Constraint Coverage			Physical Grounding	Action Feasibility	Human-Judged Creativity
Gold Correct	Entity Correct		Use (C_u)	Env. (C_e)	Rcpt. (C_r)			
0.146	0.450	3.580	3.760	4.380	4.320	3.840	3.800	3.920

Table 4: **Human check results.** Except for human Tool Usage test performance, all other results report gold solution quality manual check on a 5-point scale, where higher is better.

descriptions, and **review**, in which they compare their response with the gold solution and assess whether the KB-supported rationale for preferring the gold answer is convincing. As shown in Table 4, humans achieve **0.146 gold correctness** and **0.450 entity correctness**, both below the best-performing model. We interpret this relatively low task-solving accuracy as being partly associated with the substantial **cognitive load** imposed by the text-based setting, which requires annotators to compare many entities, decomposed parts, and fine-grained physical and state attributes. Consistent with this interpretation, in a preliminary image-based evaluation using the same sampled tasks, human gold correctness increases from **0.146 to 0.357** and entity correctness from **0.450 to 0.677**, whereas the corresponding improvements for GPT-5.2 are substantially smaller.

At the same time, the review-stage results provide complementary evidence for the quality of the annotated gold solutions. More than **90%** of gold-persuasiveness ratings fall between 3 and 5 on the 1–5 scale, indicating that annotators generally find the gold solutions persuasive even when they do not initially identify them. Under a strict exact-score criterion, where agreement requires annotators to assign the identical rating, inter-annotator agreement is **63.0%**. When ratings are grouped into negative (1–2) and positive (3–5) judgments, agreement increases to **93.0%**, suggesting that annotators largely agree on whether a gold solution is acceptable, even when they differ in the strength of their assessment. Furthermore, Figure 10 shows that human performance is less sensitive than model performance to gold cluster size and affordance typicality, while being more affected by the number of distractor entities. This pattern suggests that human difficulty is driven primarily by the overall burden of navigating dense candidate spaces, whereas model performance is more strongly shaped by the familiarity and physical grounding of the target affordance. Full human-study protocols, agreement definitions, and additional analyses are provided in Section H.11.

6 Conclusion and Future Work

In this work, we introduced CreativityBench, a large-scale benchmark for evaluating creative tool repurposing through affordance-based reasoning, along with a structured affordance knowledge base linking entities, parts, attributes, and uses. We show that current LLMs often struggle to ground creative solutions in the correct parts, attributes, and physical mechanisms. Moreover, stronger general reasoning, larger model scale, and standard inference-time strategies do not reliably improve creative affordance discovery. These suggest that creative intelligence is not simply an extension of analytical reasoning or action planning, but a distinct capability requiring grounded comparison, affordance recombination, and attention to physical constraints. We hope CreativityBench will serve as both an evaluation testbed and a resource for training models to reason more creatively and reliably. Future works can extend this framework to multimodal, interactive, and embodied settings, while exploring training objectives that promote grounded exploration and diverse yet physically plausible affordance discovery.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 560000C260018, CapitalOne-Illinois Center for Generative AI Safety, Knowledge Systems, and Cybersecurity (ASKS), U.S. DARPA ECOLE Program No. #HR00112390060, DARPA ITM Program No. FA8650-23-C-7316, DARPA MAGICS Program No. #HR0011263E079, Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE), and NSF NAIRR Pilot Grant. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This research used the Delta and DeltaAI advanced computing and data resources, which are supported by the National Science Foundation (award OAC 2320345 and award OAC 2005572) and the State of Illinois. Delta and DeltaAI are joint efforts of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

References

- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6470–6484, 2020.
- Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina von der Wense. Prost: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4597–4608, 2021.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.
- Margaret A Boden. Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2): 347–356, 1998.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL <https://arxiv.org/abs/2307.15818>, 1:2, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.
- Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4(2):1140–1147, 2019.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Yubo Dong, Xukun Zhu, Zhengzhe Pan, Linchao Zhu, and Yi Yang. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16290–16314, 2024.
- Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, et al. Creation-mmmbench: Assessing context-aware creative intelligence in mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 447–456, 2025.
- Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, et al. Are: Scaling up agent environments and evaluations. *arXiv preprint arXiv:2509.17158*, 2025.
- James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Joy P Guilford. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14, 1967.
- Hyeonjeong Ha, Xiaomeng Jin, Jeonghwan Kim, Jiateng Liu, Zhenhailong Wang, Khanh Duy Nguyen, Ansel Blume, Nanyun Peng, Kai-Wei Chang, and Heng Ji. Synthia: Novel concept design with affordance composition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20939–20958, 2025.
- Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25559–25571, 2025.
- Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangzhi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, et al. How far can unsupervised rlvr scale llm training? *arXiv preprint arXiv:2603.08660*, 2026.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, 2016.
- Seungwon Lim, Sungwoong Kim, Jihwan Yu, Sungjae Lee, Jiwan Chung, and Youngjae Yu. Visescape: A benchmark for evaluating exploration-driven decision-making in virtual escape rooms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16031–16058, 2025.
- Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.
- Li-Chun Lu, Miri Liu, Pin Chun Lu, Yufei Tian, Shao-Hua Sun, and Nanyun Peng. Re-thinking creativity evaluation: A critical analysis of existing creativity evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6329–6352, 2026.

- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.
- Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory-motor coordination to imitation. *Ieee transactions on robotics*, 24(1):15–26, 2008.
- Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. Creator: Tool creation for disentangling abstract and concrete reasoning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6922–6939, 2023.
- Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, et al. Escapebench: Pushing language models to think outside the box. *arXiv e-prints*, pp. arXiv-2412, 2024.
- Cheng Qian, Hongyi Du, Hongru Wang, Xiusi Chen, Yuji Zhang, Avirup Sil, Chengxiang Zhai, Kathleen McKeown, and Heng Ji. Modelingagent: Bridging llms and mathematical modeling for real-world challenges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1599–1633, 2025.
- Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- Robert J Sternberg. The triarchic theory of intelligence. 1997.
- Robert J Sternberg and Todd I Lubart. The concept of creativity: Prospects and paradigms. *Handbook of creativity*, 1(3-15), 1999.
- Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjeh, 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.
- Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, 2024.
- Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9743–9758, 2023.
- Zi-Han Wang, Lam Nguyen, Zhengyang Zhao, Mengyue Yang, Chengwei Qin, Yujiu Yang, and Linyi Yang. Creativebench: Benchmarking and enhancing machine creativity via self-evolving challenges. *arXiv preprint arXiv:2603.11863*, 2026.
- 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.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsec-omp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.