Assessing the Effect of Cross-Domain Mapping on Creativity in Humans and Large Language Models

Qiawen Ella Liu[1], Marina Dubova[2], Henry Conklin[1], Takumi Harada[1,3], and Thomas L. Griffiths[1]

[1]Princeton University

[2]Santa Fe Institute

[3]Toyota Motor North America, Inc., Plano, Texas

Abstract

Creativity is the ability to come up with novel ideas, a capacity crucial for human development and flourishing. Are large language models (LLMs) creative in the same way humans are, and can the same interventions increase creativity in both? We study a promising but largely untested intervention for creativity: forcing creators to draw an analogy from a random, remote source domain ("cross-domain mapping"). Human participants and LLMs generated novel designs for ten daily products (e.g., backpack, TV) under two prompts: (i) cross-domain mapping, which required drawing inspiration from a randomly assigned source (e.g., octopus, cactus, GPS), and (ii) user need, which required proposing innovations targeting unmet user needs. We show that humans reliably benefit from randomly assigned cross-domain mappings, while LLMs, on average, generate more original ideas than humans and do not show a statistically significant effect of cross-domain mappings. However, in both systems, the impact of cross-domain mapping increases when the inspiration source becomes more semantically distant from the target, and above a distance threshold, cross-domain mapping benefits the most capable LLMs too. Humans and LLMs differed in how they used the source concept: Humans tended to transfer surface features of the source, whereas LLMs transferred structural and functional properties. Our results highlight the role of remote association in creative ideation and systematic differences in how humans and LLMs respond to the same intervention for creativity.

Keywords: creativity, analogy, large language models, product design

Coming back from a hike in 1941, Swiss engineer George de Mestral noticed burrs stuck

---

Marina Dubova and Henry Conklin contributed equally to this work.

The authors declare no competing interests. We thank Toyota Motor North America, Inc. for their support of this research. We also thank Pavel Chvykov, Matthew Henderson, and the Diverse Intelligence Summer Institute for many thoughtful discussions that inspired this work.

Correspondence concerning this article should be addressed to Qiawen Ella Liu, Princeton University. E-mail: ella.qiawen.liu@princeton.edu

stubbornly to his dog's fur. Upon close examination of these burrs with a microscope, he found numerous tiny hooks catching on loops – a structure that inspired his invention of Velcro, the fastener now securing everything from toddlers' shoes to spacesuits (Stephens, 2007). Similar cross-domain transfers appear throughout the history of invention: Early aircraft were inspired by the structure of bird wings (Lilienthal, 1889), and neonatal incubators were inspired by temperature-regulation techniques from poultry brooders (Budin, 1907). In each case, invention comes not from nothing but from mapping existing concepts across domains.

Drawing inspiration across domains may appear more like a game of chance than a systematic method for inducing creativity, given it depends on the serendipitous combination of a specific hike with a specific dog at a specific place. In the general case, what makes a cross-domain mapping successful, and can it be engineered as a reliable intervention to enhance creativity? Existing research suggests humans might benefit from cross-domain inspirations (Chan et al., 2011; Dahl & Moreau, 2002; Wilson, Rosen, Nelson, & Yen, 2010). The reason is that humans are notoriously prone to fixation: once a problem is framed in a familiar way, people tend to reuse known solutions within a single domain rather than scanning for distant alternatives across domains (Basalla, 1988; German & Barrett, 2005; Jansson & Smith, 1991). Even when people try to overcome fixation by actively considering alternatives, they can still be constrained by knowledge fragmentation (Simon, 1955), characterized by the classic 'unknown unknown' problem (Firestein, 2012).

Large language models (LLMs) present a valuable target for investigating whether cross-domain mapping is useful for creativity in general, or whether it is helpful for idea-generation specifically in human minds. LLMs are trained on massive, cross-disciplinary corpora that might allow them to generate remote associations without falling prey to the same cognitive bottlenecks. While a central theme of human creativity research is developing scalable methods that help innovators overcome both fixation and unknown unknowns, it remains an open question whether interventions designed for humans will prove equally effective for systems with fundamentally different capacity for storing, retrieving, and recombining conceptual knowledge.

Here, we test cross-domain mapping as a practical intervention for idea generation by both humans and LLMs. We prompt both to generate novel product ideas either by explicitly drawing inspiration from a distant source domain or by proposing incremental improvements based on existing user needs. Human raters then assessed the ideas along four dimensions commonly associated with creative products: originality (our focal dimension), feasibility, utility, and overall investment-worthiness. Evaluating these dimensions separately, rather than collapsing them into a single creativity score, lets us examine how cross-domain mapping affects different components of creative evaluation and, as we will show, the same intervention can move these components in opposite directions. We ask whether cross-domain mapping reliably boosts originality of ideas, whether its effects differ between humans and LLMs, and what kinds of these mappings are generative in both systems.

We show that cross-domain mapping significantly boosts the originality of ideas for humans but not LLMs, and that LLMs produce more original ideas than humans on average. We also show that a random cross-domain mapping does not necessarily result in creative ideas; rather, ideas are seen as more original when the inspiration is drawn from a more distant domain – and once that distance is sufficiently large, the most capable LLMs benefit from cross-domain mapping too. Finally, we reveal a qualitative difference in the information the two systems map from an inspiration source: humans tend to transfer surface features of a semantic concept, while LLMs tend to transfer a concept's underlying structures and functions. Overall, our results highlight the

generative role of remote association and systematic differences in how humans and LLMs engage in creative generation.

Creativity as the Ability to Bridge Remote Associations

Creativity researchers have long recognized that creative ideas often arise by linking concepts that are *weakly associated*, traversing longer distances in semantic space or recombining elements drawn from distinct semantic domains (Kenett, 2018; Koestler, 1964; Mednick, 1962; Youn, Strumsky, Bettencourt, & Lobo, 2015). In these associative accounts of creativity, generating novel ideas depends on how semantic memory is organized. Semantic retrieval is not uniform, but organized into *associative hierarchies*, such that individuals with steep hierarchies prioritize dominant, conventional associates, whereas those with flatter hierarchies more readily connect distant associates (Mednick, 1962). Consistent with this view, highly creative individuals tend to navigate semantic space more expansively, retrieving more distantly related concepts (Beaty, Zeitlen, Baker, & Kenett, 2021; Gray et al., 2019; Kenett, Anaki, & Faust, 2014), shifting flexibly across semantic subcategories (Zhang et al., 2023), and producing larger associative leaps both with and without explicit goals (see Beaty & Kenett, 2023, for a review).

Despite broad agreement that creativity benefits from bridging remote associations, interventions that systematically induce such connections remain relatively rare (Linsey et al., 2010; Zhan, Yao, & Li, 2023). While some interventions are effective at inducing creative ideas in certain contexts (e.g., Zhan et al., 2023), they usually focus on a subset of possible targets and inspirations, offering little coverage of the vast combinatorial space of potential mappings. We use cross-domain mapping as a systematic intervention to "flatten" the associative hierarchy, forcing creators to leap beyond a target domain's well-trodden semantic neighborhood. Since the space of potentially useful inspirations is vast, we create a broad set of cross-domain mappings by randomly selecting pairs of source domains. In doing so, we hope to intentionally recreate the kind of serendipity that has been shown to be so important in the history of discovery (James, 1880; Souriau, 1881).

Analogy as the Mechanism for Transferring Relational Structure Across Domains

Structure-Mapping Theory (Gentner, 1983) formalizes analogy: the cognitive mechanism underlying cross-domain transfer of knowledge. An analogy can support systematic inference by aligning relational structure between a source and a target – importing a coherent system of relations that makes previously overlooked aspects of the target problem visible (Gentner, 1983; Holyoak & Thagard, 1995). In particular, aligning with and projecting from the source structure can highlight implicit limitations in the target (constraints) as well as suggest new possibilities for intervention (affordances). For example, by mapping the *circulatory system* onto *urban traffic flow*, a planner might notice the need for one-way "valves" to prevent backflow, or see an opportunity to build lots of small neighborhood routes that deliver traffic deep into side streets, the way capillaries carry blood into tissue.

However, not all mappings are equally generative. Structure-mapping theory distinguishes *surface* mappings, which transfer perceptual attributes (e.g., the sun and an atom's nucleus are both round and bright), from *structural* mappings, which transfer systems of interconnected relations (e.g., a central body attracts smaller orbiting bodies) (Gentner & Markman, 1997; Holyoak & Koh, 1987). Developmental work documents a *relational shift*: children initially map shared surface attributes, and only later develop a preference for systems of mutually constraining relations over

isolated features (Clement & Gentner, 1991; Gentner, 1988; Gentner & Toupin, 1986; Markman & Gentner, 1993; Rattermann & Gentner, 1998). Yet successful structural mapping remains hard even for adults, who are readily distracted by surface similarities. For example, when people are asked to recall a story similar to a military battle they just read, they tend to retrieve stories with surface-level matches (e.g., other battles with soldiers and tanks) rather than structural matches (e.g., a political or scientific story that shares the same underlying problem-solving logic), even though they rate the structural matches as more sound (Gentner, Rattermann, & Forbus, 1993; Gick & Holyoak, 1983; Holyoak & Koh, 1987; Ross, 1987).

Overcoming this surface bias appears to require substantial domain expertise: For example, asked to complete the analogy "a floating balloon is like __ because __," novices typically retrieve things that share a surface feature with the balloon, e.g., a kite, a leaf, other objects "in the air." Expert geoscientists, by contrast, overwhelmingly retrieved analogies grounded in the mechanism that makes a balloon rise, like other systems in which something less dense rises through a denser medium, such as warm water rising in a cold sea (Goldwater, Gentner, LaDue, & Libarkin, 2021). In addition to domain expertise, the capacity for identifying the more abstract kind of relational similarities is supported by the distributional properties of natural language (Christie & Gentner, 2014; Gentner & Christie, 2010; Loewenstein & Gentner, 2005): shared labels and relational terms recur across contexts, implicitly linking otherwise distinct domains and making higher-order structure available for comparison and transfer.

LLMs as Creative Systems That Can Generate Cross-Domain Associations at Scale

Across creative writing, design, and hypothesis generation tasks, LLMs reliably produce large numbers of novel ideas and often emulate or outperform human participants (Bellemare-Pepin et al., 2026; Cropley, 2023; Koivisto & Grassini, 2023; Si, Yang, & Hashimoto, 2024). Recent evidence further shows that LLMs can perform analogical reasoning that rivals human performance (Ding, Srinivasan, MacNeil, & Chan, 2023; Liu, Marjieh, Zhu, Goldberg, & Griffiths, 2026; Webb, Holyoak, & Lu, 2023) and flexibly recombine knowledge to generate novel solutions (Gu, Wang, Zhang, & Li, 2024; Mehrotra, Parab, & Gulwani, 2024). While the most creative humans still outperform current models in the diversity of the ideas they produce (Koivisto & Grassini, 2023; Meincke, Nave, & Terwiesch, 2025; Wenger & Kenett, 2025), evidence nevertheless suggests contemporary LLMs have substantial creative capacity.

This raises a question about whether a cross-domain mapping intervention will have the same effect on LLMs as on humans. The scaffolds that support analogical creativity in humans (e.g., extensive domain knowledge and the abstract relations encoded in language) are precisely what LLMs learn to model over pre-training. By virtue of being trained on massive, cross-disciplinary corpora, LLMs may better traverse semantic space without the bottlenecks that constrain human analogical reasoning. On the other hand, if LLMs already draw broadly across domains by default, an explicit cross-domain prompt may add little on top of what they produce under an ordinary creative framing, yielding a diminished effect relative to humans. Exploring these possibilities requires directly comparing how humans and LLMs respond to the same intervention, which is the aim of the present study.

The Current Study

We elicited ideas for novel product designs from both humans and LLMs under two matched prompting conditions. In the *cross-domain mapping* condition, participants (human or LLM) were given a target product (e.g., *backpack*) and a randomly assigned inspiration source (e.g., *cactus*), and instructed to translate a property of the source into a novel feature of the target. In the *user-need* condition, the same targets were presented without an inspiration source; participants instead identified an unmet user need and proposed a feature addressing it. Ideas from both sources were then rated by a separate pool of human judges on originality, feasibility, utility, and investment-worthiness. We tested seven LLMs spanning major model families (`o3`, `Claude-Sonnet-4`, `gpt-4o`, `gemini-2.5-pro`, `deepseek-v3`, `Llama-3.3-70b`, `Qwen2.5-72B`). Full materials, procedures, and analytic details are reported in the Method section.

Method

Eliciting Novel Ideas From Humans and LLMs

Participants

We recruited online English-speaking adults from Prolific, retaining a final sample of $N = 140$ participants after screening[1] (56 male, 83 female, 1 non-binary). Participants' mean age was 43.24 years ($SD = 13$, range = 21–77).

Target Products

We selected ten everyday consumer products spanning diverse categories: *smartphone, TV* (electronics), *sofa, desk* (furniture), *car* (vehicle), *refrigerator* (appliances), *backpack* (accessories), *yoga mat* (sports equipment), *chef's knife* (kitchen tools), and *sneakers* (clothing).

Inspiration Sources

We also selected 26 inspiration sources across various domains: animals (octopus, chameleon), plants (bamboo, cactus), natural systems (coral reef, beehive), technological systems (hydroelectric dam, GPS navigation), physical phenomena (water cycle, tornado formation), abstract concepts (storytelling), social/organizational systems (symphony orchestra, hospital, government), and food (pickle, sandwich). Specifically, the ten target products could also serve as inspiration sources, provided they were not paired with themselves.

Design and Procedure

Participants were randomly assigned to one of two conditions and completed 10 design challenges, generating one novel feature for each product. In the *Cross-Domain Mapping* condition, each target product (*e.g., toothbrush*) was paired with a randomly sampled inspiration source (*e.g., mattress*). Participants first described a property of the source (e.g., *toothbrush bristles reach tiny spaces*) and then proposed a product feature that translated that property to the target (*a mattress layer of ultrafine bristles that gently brushes away dead skin while you sleep*). From the 10 targets and 26 inspiration sources, we sampled 100 target–source pairs (see Supplemental Material, Figures

[1] We manually reviewed all responses for AI-like phrasing and atypical response consistency, and cross-referencing with AI-detection software.

S5–S6, for the full set of combinations and their mean originality ratings for humans and LLMs). Each of the 100 target–source pairs was completed by 7 participants, yielding 700 human-generated cross-domain ideas. In the *User-Need* condition, participants saw the same target products (e.g., mattress) without a source; instead, they identified an unmet user need (*e.g., partners disturb each other by moving during sleep*) and proposed an innovative solution (*e.g., separate mattress zones that absorb motion and automatically adjust firmness*). Each product was received by 70 participants, yielding 700 human-generated user-need ideas.

Rephrasing and Format Normalization

To control for superficial differences in style, framing, and length between ideas generated by humans and LLMs, we normalize each idea by asking the same LLM (`o3`) to rephrase the original content into a short (1–2 sentence), reader-friendly product description. Rephrased outputs omit both the cross-domain analogy framing and explicit user-need framing, focusing only on what the product is and does.[2] A pilot study that did not apply this rephrasing step reproduces the model-level ordering and the distance–originality relationship reported in the main analyses (see Supplemental Material, Section S6).

Evaluating Ideas

Participants

We recruited $N = 1002$ online English-speaking adults from Prolific (459 male, 528 female, 11 non-binary, 4 preferred not to say). Participants' mean age was 41.9 years ($SD = 12.9$, range = 20–80). This study was preregistered at Open Science Framework (link).

Material

We split all rephrased human- and LLM-generated ideas ($n = 2800$) into 100 lists, each containing approximately 30 ideas sampled evenly across generation source (human vs. LLM), condition (cross-domain vs. user-need), and target product. Each idea was rated by approximately 10 independent participants. Each list also included four catch trials to assess attention and understanding of the rating scales. 108 (about 10%) participants who failed more than 50% of these checks were excluded.

Procedure

Participants were instructed to imagine themselves as judges evaluating startup pitches on a show like *Shark Tank*. They rated each idea on four dimensions using 5-point Likert scales with clearly labeled endpoints, with the order of these ratings randomized between participants: *Originality*: the extent to which the idea was novel or unexpected. *Feasibility*: the degree to which the idea could plausibly be implemented with current technology. *Utility*: the extent to which the idea would provide meaningful real-world benefits to users. *Investment-Worthiness*: an overall judgment of whether the idea was strong enough to merit investment. Participants were instructed

[2] For example, the idea "A cactus has spines that collect moisture from the air and tissue that stores water. Like a cactus, this backpack uses micro-spines to condense moisture as you walk and channels it into an internal reservoir, providing water and cooling during long trips." is rephrased as "This backpack collects moisture from the air as you move and channels it into an internal reservoir. The collected water can be used for hydration and provides gentle cooling during extended use."

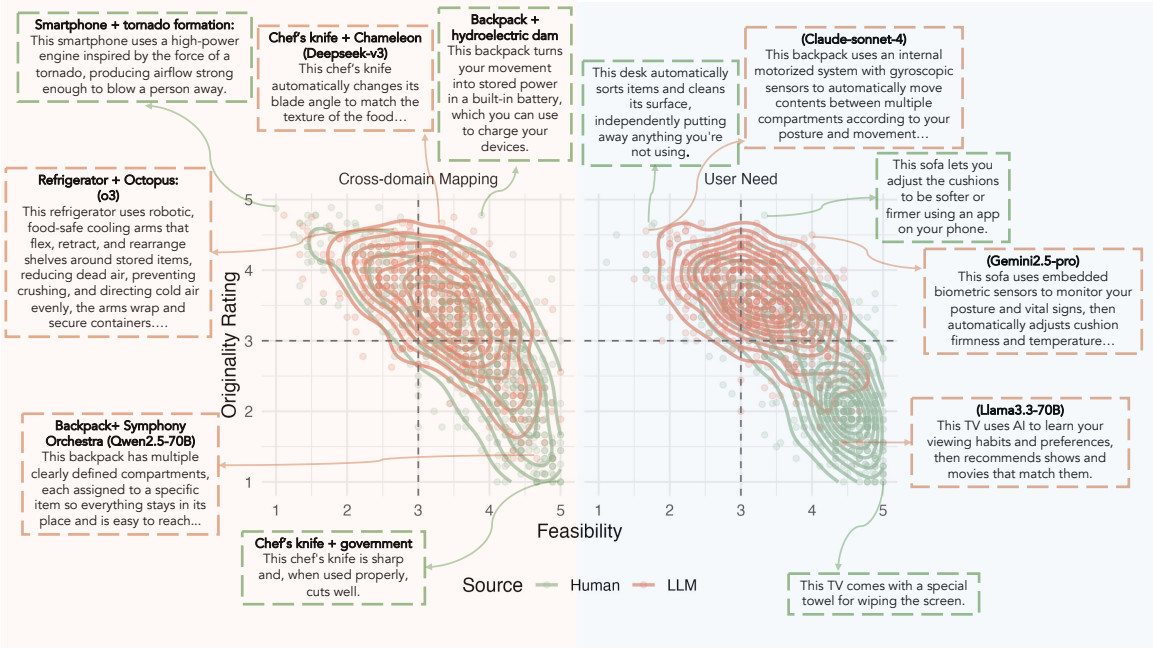

Figure 1

Cross-domain mapping increases originality for humans but yields weaker benefits for LLMs. Ideas (points) are presented along originality (vertical axis) and feasibility (horizontal axis), with density contours showing distributions for human (green) and LLM (orange) outputs. Cross-domain prompting (left) produces more highly original human ideas relative to the user-need prompt (right).

to rely on their intuitive judgments and to evaluate each idea independently. Ideas were presented in randomized order within each list.

Results

Originality, Feasibility, Utility, and Investment-Worthiness

Across ideas, originality shows a strong trade-off with feasibility ($r = -0.74$, $p < .001$, as shown in Figure 1), and is weakly negatively correlated with utility ($r = -0.08$, $p < .001$). More original ideas are only modestly more likely to be judged as worth investing in ($r = 0.23$, $p < .001$). Investment-worthiness is driven primarily by perceived utility ($r = 0.71$, $p < .001$), far more than by originality or feasibility ($r = .11$, $p < .001$). The same qualitative relationships hold when considering human-generated ideas or LLM-generated ideas alone (see Supplemental Material, Figure S1).

Humans and LLMs Differ Across All Four Dimensions

We assessed each of the four rating dimensions using a linear mixed-effects model predicting originality ratings from each source of ideas, with random intercepts for raters (since each rater rated multiple ideas) and ideas (since each idea was rated by multiple raters). Across both conditions, LLM-generated ideas were rated as more original, more useful, and more worthy of investment

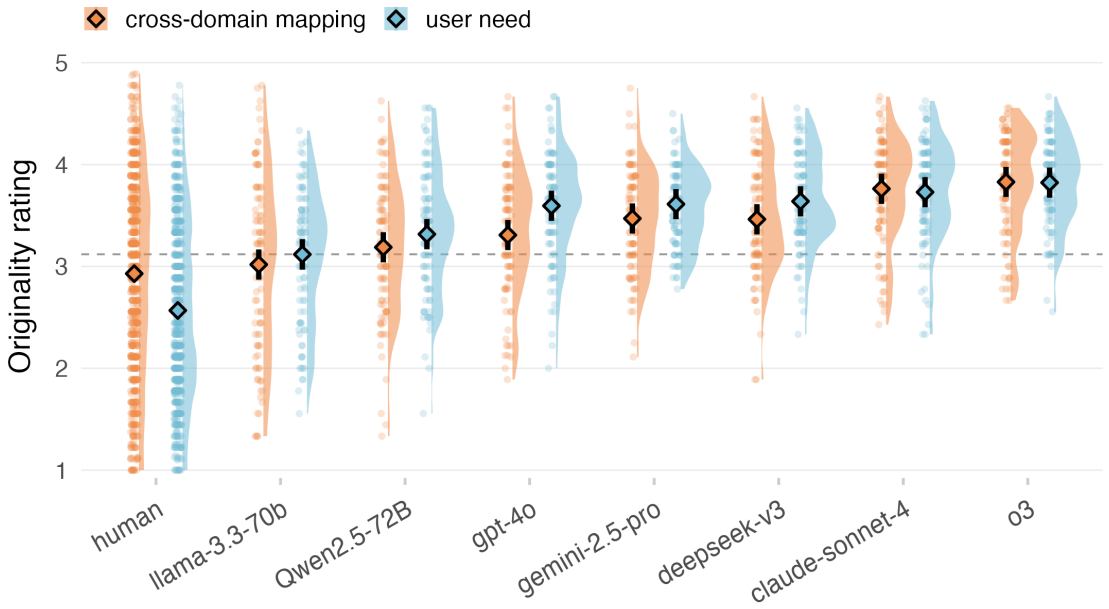

Figure 2

Originality ratings by model and condition. Distributions show originality ratings for ideas generated by different models under each prompting conditions. Diamonds indicate estimated marginal means, with error bars showing 95% confidence intervals.

than humans, while human-generated ideas were rated as more feasible. We unpack the results on originality in the subsequent section and report full distributions and statistics for the other three dimensions in the Supplemental Material (Figures S2–S4).

Meaningful Differences in Originality Between Humans and LLMs, and Among LLMs

Across both conditions, LLM-generated ideas were rated as more original than their human counterparts (see Figure 2). Although cross-domain prompting increased originality for humans relative to the user-need condition, human ideas in the cross-domain mapping condition were still rated as less original than LLM-generated ideas overall ($M = 2.93$, 95% CI $[2.87, 2.99]$). Most LLMs were rated as statistically significantly more original than humans after correction for multiple comparisons (Holm-adjusted $p < .01$), with the exception of `Llama-3.3-70b`, which did not reliably outperform humans ($M = 3.02$, 95% CI $[2.87, 3.17]$). Among LLMs, originality ratings differed significantly: `o3` and `Claude-Sonnet-4` achieved the highest ratings ($M = 3.83$, 95% CI $[3.68, 3.98]$ and $M = 3.76$, 95% CI $[3.61, 3.91]$, respectively), significantly exceeding humans and other LLMs (all Holm-adjusted $p < .001$). `o3` and `Claude-Sonnet-4` did not differ reliably from each other, together forming a top-tier group that we revisit separately below, where we examine how the semantic distance between target and source impact originality.

A similar pattern held in the user-need condition. Human ideas again showed the lowest originality ratings ($M = 2.57$, 95% CI $[2.50, 2.63]$), with all LLMs rated as more original than humans (all Holm-adjusted $p < .001$). However, unlike in the cross-domain mapping condition,

differences among LLMs were smaller. Descriptively, `o3` and `Claude-Sonnet-4` continued to receive the highest originality ratings; however, their advantage was statistically significant only relative to `Llama-3.3-70b` and `Qwen2.5-72B` (all Holm-adjusted $p < .001$), and not relative to `gpt-4o`, `gemini-2.5-pro`, or `deepseek-v3` after correction. The narrower spread among LLMs suggests that frontier models perform similarly when generating ideas from user needs, with clear differences emerging only under the more challenging cross-domain mapping task.

Cross-Domain Mappings Increase Originality for Humans, but Not for LLMs

We examined whether cross-domain prompting increases originality by fitting a linear mixed-effects model predicting originality ratings from idea source (human vs. LLM), condition (cross-domain mapping vs. user need), and their interaction, with random intercepts for participants and ideas. Aggregating across ideas, cross-domain mapping led to significantly higher originality ratings overall ($b = 0.36$, $SE = 0.04$, $t = 9.05$, $p < .001$). However, the augmentation effect of cross-domain mapping was primarily driven by human responses, as we also found a significant interaction between condition and idea source ($b = 0.48$, $SE = 0.06$, $t = 8.42$, $p < .001$). Specifically, for humans, cross-domain prompts reliably increased originality relative to the user-need condition, whereas LLM originality showed little evidence of a comparable benefit (see also Figure 1 and Figure 2). However, this null effect for LLMs on average masks important heterogeneity across combinations and models. As we show in the next section, cross-domain mapping *does* boost originality for the most capable LLMs when the inspiration source is sufficiently semantically distant from the target product.

The augmentation effect of cross-domain mappings for humans is even more pronounced in the upper tail of the originality distribution. When we focus on the top 10% original ideas of each idea pool (human/LLM) ($n = 96$), human ideas are slightly *more original* than LLM ideas ($b = -0.07$, $SE = 0.03$, $t = -2.19$, $p = .031$). For example, when asked to map *tornado formation* onto a novel feature of smartphones, one human envisioned a phone that generates airflow strong enough to blow a person away (Figure 1, upper left corner); another imagined a smartphone which "contains powerful internal fans that, when you press a button after misplacing it, spin to create a cyclone that guides the phone back to you." Inspired by *GPS navigation*, someone proposed a shoe that "shows your GPS route in bright colors, uses a light-up arrow on the top to point the way, and displays street names, road signs, and lights around the shoe." Across all ideas, these human ideas were rated as the most original (4.88–4.89 out of 5).

What Combinations Tend to Yield More Original Ideas

Not all target–inspiration combinations were equally effective at inducing original ideas. Some pairings consistently yielded higher originality ratings across both humans and LLMs (see Figures S5 and S6). For instance, combining *octopus* with *car* produced original ideas from both humans and LLMs (e.g., human: "This car moves on eight tentacle-like appendages instead of wheels, letting it drive, climb, sink, and float"; `o3`: "like octopus, this car deploys a flexible skin embedded with micro-vacuum pads that [...] create temporary adhesion, allowing self-parking on steep walls"). In contrast, pairing *sneaker* with *sofa* resulted in mundane ideas that transferred cushioning features (e.g., human: "This sneaker has extremely cushioned, plush soles designed specifically for comfort"; `Qwen2.5-72B`: "this sneaker uses a multi-layered cushioning system ..."). The correlation between average originality ratings for the same combinations produced by humans

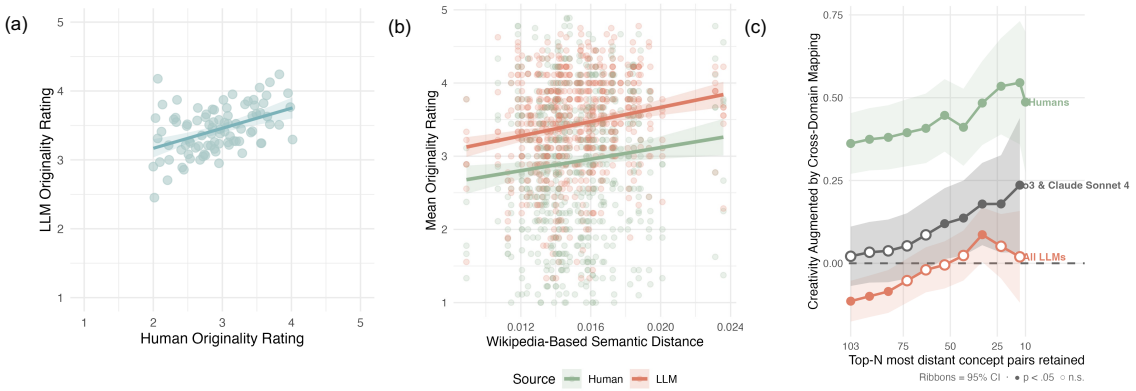

Figure 3

Originality varies by cross-domain combinations and semantic distance: (a) originality ratings for human vs LLM ideas given the same target-source combinations are significantly correlated. (b) Mean originality ratings plotted against Wikipedia-based semantic distance between source and target concepts by source (human vs LLM). (c) Estimated benefit of cross-domain mapping over the user-need baseline as progressively closer target–source pairs are excluded; the leftmost point includes all pairs, and moving rightward restricts the analysis to increasingly distant subsets. Ribbons show 95% CIs; filled points indicate $p < .05$*, open points n.s.*

versus LLMs was significantly positive ($r = 0.44$, $p < .001$, see also Figure 3a), suggesting that certain combinations are inherently more conducive to original thinking.

To understand more systematically what kind of target–source combinations tend to yield more original ideas, we quantified the semantic distance between each pair using general knowledge associated with the corresponding concepts in Wikipedia. For each concept, we retrieved up to 50 sentences from its corresponding Wikipedia article (e.g., the "Octopus" article for *octopus*) X and embedded these sentences using a pretrained sentence-transformer model θ (`all-mpnet-base-v2`). We use the density estimation technique from Conklin (2025), which takes the embeddings $Z = \theta(X)$ and computes a soft quantization $\hat{Z}$ for each token by computing the cosine similarity between each embedding and n uniformly distributed points on the surface of the unit sphere S and normalising. The result is a categorical distribution $\hat{Z}$ over the discrete points in S that describes the model's representation space:

$$\underset{1\times n}{P(\hat{Z})} \propto \sum \text{softmax}\left(\underset{bs\times h}{\frac{Z}{|Z|}} \cdot \underset{h\times n}{\frac{S}{|S|}} \right) \tag{1}$$

By conditioning these estimates $P(\hat{Z}|X = x)$, we get a distribution for each concept x describing how it is represented in the model's embedding space. Semantic distance between a target i and its inspiration j was computed as the Jensen–Shannon divergence between their distributions $D_{KL}(P(\hat{Z}|\text{concept}_i)||P(\hat{Z}|\text{concept}_j))$ at each layer, then aggregated. Across ideas, combinations that paired more semantically distant concepts tended to receive higher originality ratings (human $\beta = .1$, $t = 2.721$, $p < .01$; LLMs $\beta = .18$, $t = 4.81$, $p < .001$) (see Figure 3b).

Given that semantic distance predicts originality, a natural question is whether the null effect of cross-domain mapping on LLM still holds once we account for the distance of the pairing. We

progressively excluded low-distance combinations and compared originality in the cross-domain mapping condition against the user-need baseline for three groups: humans, all LLMs, and the two highest-performing LLMs (`o3` and `Claude-Sonnet-4`) (see Figure 3c). Across all three groups, progressively restricting the analysis to more distant pairings gradually increased the benefit of cross-domain mapping. For humans, the effect was already large and significant when all pairings were included, and grew larger still as closer pairings were excluded. Pooled across LLMs, cross-domain mapping initially *underperformed* the user-need condition, but the gap narrowed as closer pairings were excluded and the difference between two conditions was no longer significant once the closest 20% of pairs were excluded. For `o3` and `Claude-Sonnet-4`, when all pairings were included, there was no significant effect for cross-domain mapping ($b = 0.02$, $p = .64$). However, as lower-distance pairings were excluded, the effect grew in magnitude and became statistically significant: restricting to the top 50% most distant pairings yielded a significant effect ($b = 0.12$, $p = .029$), and restricting to the top 10% produced a substantially larger effect ($b = 0.24$, $p = .024$). Together, these results suggest that the benefit of cross-domain mapping scales with semantic distance for both humans and LLMs, but the distance required before a benefit appears differs considerably across humans, LLMs on average, and the best LLMs. One interpretation is that LLMs have flatter associative hierarchies than humans by default, which we revisit in the Discussion.

Originality Comes With a Price for Humans, but Less So With LLMs

The preceding analyses establish that LLMs produce more original ideas than humans on average, but human ideas at the upper tail—particularly under cross-domain mapping—can match or even exceed LLM-generated ideas on originality. Here, we note a trade-off affecting humans but not LLMs: when humans reach the frontier of originality, their ideas come at a cost to other dimensions in ways that LLM ideas do not.

We identified the top 10% most original ideas generated by humans and LLMs respectively. We then compared their ratings across all four dimensions using linear mixed-effects models with random intercepts for raters and ideas. The results are shown in Figure 4a, where positive coefficients for a dimension indicate that LLM ideas were rated higher along that dimension than human ideas.

The results reveal an interesting asymmetry between the two prompting conditions. In the *user-need* condition, the top 10% of human ideas were rated as less original than their LLM counterparts ($b = 0.32$, $SE = 0.05$, $p < .001$), yet matched LLM ideas on utility ($b = 0.14$, $p = .20$), feasibility ($b = -0.06$, $p = .63$), and investment-worthiness ($b = 0.12$, $p = .28$). In other words, when humans focus on addressing user needs, their most original ideas were less novel than the LLMs' but just as useful, feasible, and worthy of investment.

Under *cross-domain mapping*, this pattern flipped. The top 10% of human ideas now matched LLMs on originality ($b = 0.02$, $p = .63$), but were rated as substantially less feasible ($b = 0.26$, $p = .045$), less worthy of investment ($b = 0.60$, $SE = 0.12$, $p < .001$), and most prominently, less useful ($b = 0.81$, $SE = 0.15$, $p < .001$). Some of these highly original but impractical human ideas include a desk inspired by a *pickle* ("This desk is green, gives off a vinegar smell, and releases juice when you squeeze it," originality = 4.78/5, utility = 1.22/5) or a TV inspired by a *sandwich* ("This TV has a bread-shaped frame with layered sections that look like sandwich fillings..."; originality = 4.75/5, utility = 1.13/5).

The most original human ideas come at a sharp cost to utility, while equally original LLM ideas do not. Why does cross-domain inspiration lead humans to trade away utility? Humans and

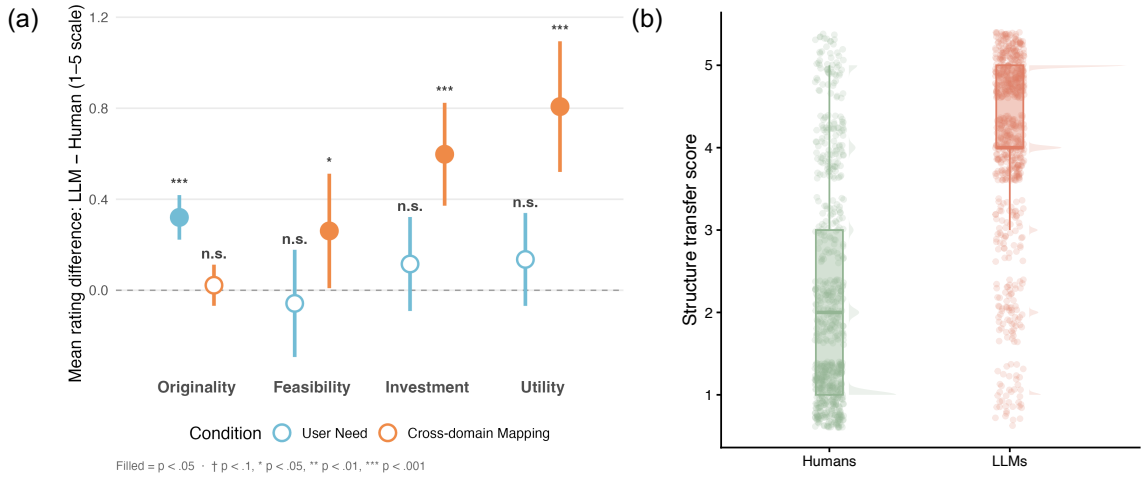

Figure 4

Originality–utility trade-off and surface-vs-structure transfer. (a) Mean rating differences (LLM − human, 1–5 scale) on the top 10% most original ideas in each pool, by condition. Points are coefficients from linear mixed-effects models (positive = LLM rated higher); filled = $p < .05$; error bars = 95% CIs. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, $^{\dagger}p < .10$. (b) Manually labeled structure-transfer scores for all 1400 cross-domain ideas (1 = pure surface transfer, 5 = pure structural transfer).

LLMs may be drawing on source domains in fundamentally different ways and transferring different kinds of information from source to target. We investigate this possibility next.

Humans Transfer Surface Features; LLMs Transfer Structures and Functions

A long-standing distinction in the analogy literature holds that mappings vary in depth: one can import a source's surface appearance, or one can import its underlying structure (Gentner & Toupin, 1986). This distinction offers a natural explanation for why cross-domain inspiration might lead humans, but not LLMs, to sacrifice utility for originality: the two systems may favor different kinds of transfer.

As a first, indirect test, we looked for a lexical signature of surface-based mapping using the Lancaster Sensorimotor Norms (Lynott, Connell, Brysbaert, Brand, & Carney, 2020), which quantify how strongly each English word evokes perceptual experience (e.g., visual, auditory, haptic) and bodily action. For each content word in an idea, we retrieved its maximum sensorimotor strength across the eleven perceptual and action dimensions in the norms, and averaged these values across all matched content words to obtain an idea-level sensorimotor score. In the cross-domain mapping condition, human ideas drew more on words that evoke sensory and motor experience than LLM ideas (M = 3.11 vs. 3.07, $p < .001$). Crucially, this is not a general human preference for perceptual vocabulary, because in the user-need condition, where no source domain is provided, the direction reverses and human ideas rely on *less* sensory and action-laden language than LLM ideas (M = 3.02 vs. 3.08, $p < .001$; see also Supplemental Material, Figure S7).

To directly measure the extent to which humans and LLMs transfer structure in cross-domain mappings, the first author manually coded all 1400 cross-domain ideas on a 1–5 scale,

from 1 indicating pure surface transfer (importing a visual or physical attribute without its function) to 5 indicating pure structural transfer (importing a causal mechanism or organizational principle without surface resemblance). LLM ideas clustered near the structural end of the scale, while human ideas were concentrated at the surface end of the scale, with a minority reaching higher structural scores, suggesting that deep analogical mapping is possible but effortful for humans (Figure 4b). For example, inspired by pickles, many human participants proposed a car with "pickle-green paint with grooves and bumps," or inspired by cacti, a backpack with thorny textures to deter theft. In contrast, LLM cross-domain mappings more often align and project mechanisms or functions across domains: the same example of a pickle-inspired car leads to a self-healing, anti-corrosion mechanism (micro-brine capsules that seal scratches), while the cactus becomes a water-harvesting system (micro-spine condensation fins channeling moisture into a drinkable reservoir). It's unlikely that LLMs don't know pickles are typically green and dimpled while cacti are spiky, but they differ from humans in what is treated as generative during analogical transfer.

This surface-vs-structure distinction offers an explanation for why human ideas in the cross-domain mapping condition tend to be less useful. We fit linear mixed-effects models predicting each dimension from the structure-transfer score, separately for human and LLM ideas (see also Supplemental Material, Figure S8). Structuralness positively predicted originality (human $b = 0.06$, $p = .026$; LLM $b = 0.14$, $p < .001$) and, with the largest magnitude, utility (human $b = 0.21$, $p < .001$; LLM $b = 0.06$, $p < .001$), while negatively predicted feasibility (human $b = -0.06$, $p = .019$; LLM $b = -0.15$, $p < .001$).

Discussion

Creativity is often associated with sudden and serendipitous cross-domain inspirations. The central question motivating this work was whether such cross-domain mappings can be induced at scale, and whether cross-domain mappings affect creative insights differently for humans and LLMs. We found that (1) LLMs generated ideas that were perceived as more original on average, regardless of whether cross-domain mappings were explicitly prompted; (2) on average, cross-domain mapping prompts increased the perceived originality of human but not LLM ideas; (3) across both systems, greater semantic distance between target and source domains predicted higher perceived originality, and once that distance was sufficiently large, the most capable LLMs also benefited from cross-domain mapping; and (4) humans and LLMs used source domains in qualitatively different ways, with consequences for the utility of the resulting ideas.

Semantic Distance as a Driver of Originality

We found cross-domain mapping interventions increased originality for people but not for LLMs. One interpretation is that without an explicit invitation to map across domains, human participants may remain anchored to incremental improvements that the product already affords, which is exactly the kind of fixation cross-domain prompts are designed to disrupt. LLMs may be less constrained in this way: trained on cross-disciplinary corpora, they may already draw broadly across domains under a user-need framing, leaving less room for improvement when given an explicit cross-domain prompt. We cannot test this possibility, because our study only looks at the behaviors that humans and LLMs ended up producing, not at the representations that drive them.

The null effect for LLMs is conditional, not absolute. Originality scaled with the semantic distance between target and inspiration for both systems, and once that distance was large enough

(e.g., car and octopus, backpack and cactus), the most capable LLMs benefited from cross-domain prompts as well. We interpret the human–LLM asymmetry to cross-domain intervention as more of a difference in sensitivity to prompting than an indication of whether cross-domain mappings can be generative in principle.

Originality, Feasibility, and Utility

Across ideas, originality showed a strong negative relationship with feasibility and a weaker but still negative relationship with utility, especially among human-generated ideas. At face value, this might suggest that pushing ideas toward high originality necessarily produces impractical ideas with unclear utility. Some of the most original human ideas we collected illustrate the concern: a desk that smells of vinegar and oozes juice when squeezed or a TV with a sandwich-shaped frame layered to mimic fillings. These ideas are vivid and unexpected, but it is hard to see what problem they solve.

But this does not mean every impractical-looking idea should be dismissed. Judgments of feasibility and utility are necessarily made relative to a currently imaginable problem/solution space (Shtulman, 2023). Ideas that sit outside that space may look both impossible and pointless, not necessarily because they violate principles of physics or markets, but because the mechanisms by which they could be realized are not yet representable. Many historically transformative ideas initially suffered from this problem. For example, early skepticism about AI-generated arts often rested on the assumption that a creative machine would entail a human-like robot arms manipulating brushes (McCorduck, 1991), rather than statistical models operating over latent representations—an implementation path that was not merely unlikely, but effectively *unthinkable* at the time. Once a workable path appears, judgments of feasibility can shift rapidly; and until adoption, utility is necessarily speculative.

Embodiment, Language, and the Structuralness of Analogy

We observed a clear qualitative difference between humans and LLMs in what they transferred from a source given a cross-domain mapping: humans tended to transfer surface features, while LLMs tended to transfer relational structure and function. This difference speaks to the broader question motivating this work: which principles of creativity generalize across intelligent systems, and which are tied to the particular constraints of being human?

A simple explanation is that humans participants put limited effort into the task, while LLMs, by design, give their best response to the questions they are posed. We cannot objectively measure the effort put in by participants, but we can instruct LLMs to try less. In a preliminary investigation, when we instructed several LLMs to respond as a lazy human who doesn't think too much, their outputs grew shorter but remained predominantly structural. When we instead instructed them to respond as an artist attentive to perceptual features, outputs shifted toward surface mappings, and pickle-green paint cars are proposed. The gap, then, may be less about effort than about which features each system is steered to attend to.

A deeper explanation is that the surface bias documented in decades of analogy research (Gentner et al., 1993; Holyoak & Koh, 1987; Ross, 1987) reflects the conditions under which human cognition operates. For humans, many concepts are anchored more deeply in embodied perception than in abstract description: a pickle is something we have seen, held, and tasted far more often than we have read about its fermentation chemistry. Perceptual features are immediate

and cheap to retrieve, while the relational structure that facilitates useful inference requires effortful abstraction and structural alignment. LLMs, by contrast, never experience a pickle perceptually. They encounter the concept primarily through text in which relational and functional properties (*how* they are produced, *why* they preserve well, etc.) are richly described, while perceptual features may be comparatively underspecified. This account is consistent with evidence that people who lack direct perceptual access to a given modality shift toward relational and functional content. Congenitally blind adults, for instance, list reliably fewer perceptual features of concrete objects (e.g., that a pencil is cylindrical) and more contextual, taxonomic, and functional ones (e.g., that a pencil is used with paper) than sighted adults (Kim, Elli, & Bedny, 2019; Lenci, Baroni, Cazzolli, & Marotta, 2013; Mamus, Özyürek, Ortega, & Majid, 2025).

Limitations and Future Directions

Our design captures one-shot idea generation, whereas real-world creativity is inherently iterative: ideas are criticized, combined, revised, and gradually made workable. Future work should therefore study iterative workflows which could instantiate explicit mechanisms that humans naturally use in creative practice—e.g., reflection and adversarial critique that could be implemented either within models (self-critique and refinement loops), in multi-agent settings, or in human–AI collaborations. Second, our originality, feasibility, and utility ratings come from lay judges, which is appropriate for consumer products but leaves unanswered how domain experts or markets would adjudicate the same ideas. Third, we observe what each system produced, not the candidate mappings it considered and discarded; our finding that LLMs default to structural transfer is therefore a claim about outputs, not about the underlying generative process. Probing that process through e.g., sampling distributions of completions, reasoning traces, or internal representations, is a natural next step we leave to future work.

Data Availability

This research is approved by the Princeton University Institutional Review Board (Protocol No. 17087). Data, materials, and preregistrations are available through the Open Science Framework at OSF (https://osf.io/xp4ed/).

References

Basalla, G. (1988). *The evolution of technology*. Cambridge University Press.

Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, *27*(7), 671–683. doi: 10.1016/j.tics.2023.04.004

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, *41*, 100859.

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., ... Jerbi, K. (2026). Divergent creativity in humans and large language models. *Scientific Reports*, *16*(1), 1279.

Budin, P. (1907). *The nursling: the feeding and hygiene of premature & full-term infants*. Caxton Publishing Company.

Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., & Kotovsky, K. (2011). On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples.

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive science*, *38*(2), 383–397.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive science*, *15*(1), 89–132.

Conklin, H. C. (2025). *Information structure in mappings: an approach to learning, representation and generalisation* (Unpublished doctoral dissertation). The University of Edinburgh.

Cropley, D. (2023). Is artificial intelligence more creative than humans?: ChatGPT and the divergent association task. *Learning Letters*, *2*, 13–13.

Dahl, D. W., & Moreau, P. (2002). The influence and value of analogical thinking during new product ideation. *Journal of marketing research*, *39*(1), 47–60.

Ding, Z., Srinivasan, A., MacNeil, S., & Chan, J. (2023). Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In *Proceedings of the 15th conference on creativity and cognition* (pp. 489–505).

Firestein, S. (2012). *Ignorance: How it drives science*. Oxford University Press.

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

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child development*, 47–59.

Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. *Language and Cognition*, *2*(2), 261–283.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, *52*(1), 45.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*(4), 524–575.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive science*, *10*(3), 277–300.

German, T. P., & Barrett, H. C. (2005). Functional fixedness in a technologically sparse culture. *Psychological Science*, *16*(1), 1–5.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1–38.

Goldwater, M. B., Gentner, D., LaDue, N. D., & Libarkin, J. C. (2021). Analogy generation in science experts and novices. *Cognitive Science*, *45*(9), e13036.

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., ... Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist*, *74*, 539–554.

Gu, T., Wang, J., Zhang, Z., & Li, H. (2024). LLMs can realize combinatorial creativity: generating creative ideas via LLMs for scientific research. *arXiv preprint arXiv:2412.14141*.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & cognition*, *15*(4), 332–340.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. MIT Press.

James, W. (1880, October). Great men, great thoughts, and the environment. *The Atlantic Monthly*, *46*, 441–459.

Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, *12*(1), 3–11.

Kenett, Y. N. (2018). Going the extra creative mile: The role of semantic distance in creativity—theory, research, and measurement. *The Cambridge handbook of the neuroscience of creativity*, *3*, 233–48.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*, 407.

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23), 11213–11222. doi: 10.1073/pnas.1900952116

Koestler, A. (1964). *The act of creation*. London: Hutchinson.

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, *13*(1), 13601.

Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, *45*(4), 1218–1233. doi: 10.3758/s13428-013-0323-4

Lilienthal, O. (1889). *Birdflight as the basis of aviation*. London: Longmans, Green and Co.

Linsey, J. S., Tseng, I., Fu, K., Cagan, J., Wood, K. L., & Schunn, C. (2010). A study of design fixation, its mitigation and perception in engineering design faculty. *Journal of Mechanical Design*, *132*(4), 041003.

Liu, Q. E., Marjieh, R., Zhu, J.-Q., Goldberg, A. E., & Griffiths, T. L. (2026). Parallelograms strike back: Llms generate better analogies than people. *arXiv preprint arXiv:2603.19066*.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive psychology*, *50*(4), 315–353.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020, June). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. Retrieved 2021-03-23, from `https://doi.org/10.3758/s13428-019-01316-z` doi: 10.3758/s13428-019-01316-z

Mamus, E., Özyürek, A., Ortega, G., & Majid, A. (2025). Gestural and verbal evidence of conceptual representation differences in blind and sighted individuals. *Cognitive Science*, *49*(10), e70125. doi: 10.1111/cogs.70125

Markman, A. B., & Gentner, D. (1993, October). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467. Retrieved 2021-07-16, from `https://www.sciencedirect.com/science/article/pii/S001002858371011X` doi: 10.1006/cogp.1993.1011

McCorduck, P. (1991). *Aaron's code: meta-art, artificial intelligence, and the work of harold cohen*. Macmillan.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*, 220–232.

Mehrotra, P., Parab, A., & Gulwani, S. (2024). Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*.

Meincke, L., Nave, G., & Terwiesch, C. (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour*, *9*, 1107–1109.

Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive development*, *13*(4), 453–478.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629–639.

Shtulman, A. (2023). *Learning to imagine: The science of discovering new possibilities*. Harvard University Press.

Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Souriau, P. (1881). *Théorie de l'invention*. Hachette.

Stephens, T. (2007). *How a Swiss invention hooked the world.* (SWI swissinfo.ch. Available at `https://www.swissinfo.ch/eng/how-a-swiss-invention-hooked-the-world/`. Accessed 13 May 2026.)

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wenger, E., & Kenett, Y. (2025). We're different, we're the same: Creative homogeneity across LLMs. *arXiv preprint arXiv:2501.19361*.

Wilson, J. O., Rosen, D., Nelson, B. A., & Yen, J. (2010). The effects of biological examples in idea generation. *Design Studies*, *31*(2), 169–186.

Youn, H., Strumsky, D., Bettencourt, L., & Lobo, J. (2015). Invention as a combinatorial process: evidence from us patents. *Journal of the Royal Society interface*, *12*(106).

Zhan, Z., Yao, X., & Li, T. (2023). Effects of association interventions on students' creative thinking, aptitude, empathy, and design scheme in a steam course: considering remote and close association. *International Journal of Technology and Design Education*, *33*(5), 1773–1795.

Zhang, J., Zhuang, K., Sun, J., Liu, C., Fan, L., Wang, X., ... Qiu, J. (2023). Retrieval flexibility links to creativity: evidence from computational linguistic measure. *Cerebral Cortex*, *33*(8), 4964–4976.