MEASURING COMBINATORIAL CREATIVITY IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Combinatorial creativity, the process of combining familiar ideas in unfamiliar ways, is one of the foremost engines of scientific, technological, and artistic progress. While large language models (LLMs) can fluently produce novel text relative to their training data, their capacity to combine information in novel *and* useful ways, in order to invent new concepts, remains unclear. This work introduces **KOMBINE**, a combinatorial creativity benchmark tasking models to discover surprising, original, and useful associations, analogies, and blends between unrelated entities, and to use those analogies and blends to *invent novel concepts by re-combining the parts of existing ones*. In a large-scale study across 35 LLMs, we find that: **(1)** LLMs are 10–15 percentage points better at discovering associations and analogies than blends, and **(2)** exhibit an *abstraction bottleneck* on blending, in which even frontier LLMs struggle to find a shared abstraction between inputs and instantiate a coherent blended space. **(3)** LLMs also exhibit "inventive homogeneity," in which one in five concepts invented for a given item reuses properties invented by another model. Moreover, **(4)** models are $14\times$ more likely to reuse other invented properties on blends than analogies, and **(5)** models from the same provider are $2.3\times$ more likely to reuse properties than those from different providers. Our results reveal creative strengths in LLMs' capacity for association and analogy, point to an abstraction bottleneck on blending, and show how output homogeneity constrains the conceptual diversity of inventions. In summary, we provide a tool that can be used to assess progress towards more creative LLMs.

1 INTRODUCTION

Einstein famously remarked that "Combinatory play seems to be the essential feature in productive thought," (Hadamard, 1954, p. 142) referring to the cognitive processes he believed underpinned creative insight in mathematics and the sciences. Indeed, there is a rich body of psychological literature modeling creativity as a combinatorial process over representations and abstractions (Koestler, 1964; Simonton, 2004; 2021; 2010; Thagard, 2010), and a growing body of work in the science of science showing that highly impactful scholarly work tends to comprise more distant combinations of ideas, methods, and concepts (Shi and Evans, 2023; Fang and Evans, 2025). This form of creativity elicited by making unfamiliar combinations of familiar ideas has been defined by Boden (2004) as *combinatorial (or combinational) creativity*, and in recent years interest in evaluating and enhancing this ability among LLMs, especially for scientific applications, has steadily grown (Nagarajan et al., 2025; Gu et al., 2025; Artiles et al., 2026; Shen et al., 2026; Sanyal et al., 2025).

In the cognitive science literature, three main forms of combinatorial creativity are studied, as pictured in Figure 1. First, **associative creativity** describes the ability to make distant connections between otherwise unrelated ideas (Mednick, 1962), and has been studied extensively in both humans and LLMs (Schapiro et al., 2026a) through tests like the Divergent Association Task (DAT) (Olson et al., 2021), its variants like the Conditional DAT (Nakajima et al., 2026), and the Remote Associates Test (RAT) (Mednick, 1962), and a recent associative creativity benchmark by Wadhwa et al. (2026). Second, **analogy** involves using knowledge from a familiar domain to infer or generate knowledge about a different, but somehow connected domain (Gentner, 1983; Petersen et al., 2026; Liu et al., 2026b; Holyoak, 2025). Recent work has explored using analogical reasoning to improve creativity in LLMs for scientific discovery (Shen et al., 2026) and technological innovation (Liu et al., 2026a). Lastly, conceptual **blending** (Fauconnier and Turner, 2008; Eppe et al., 2018) involves selectively projecting structure from two input spaces into a new blended space with emergent properties of its

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

Table 1: **Forms of combinatorial creativity assessed in KOMBINE.** Each task assesses an ability recognized in cognitive science as integral to combinatorial creativity (Boden, 2004; Simonton, 2023)

Task (§3)	Creativity Mechanism(s) & Source(s)	Example
Association	Remote association (Mednick, 1962)	*Einstein* → *violin* → *Mozart*
Analogy	Analogy (Gentner, 1983; Rumelhart and Abrahamson, 1973) and metaphor (Lakoff and Johnson, 2008)	*immune system* : *body* :: *national defense* : *nation*
Blending	Conceptual blending (Fauconnier and Turner, 2008); conceptual combination (Thagard, 1984)	*organism* + *machine* → *cyborg*

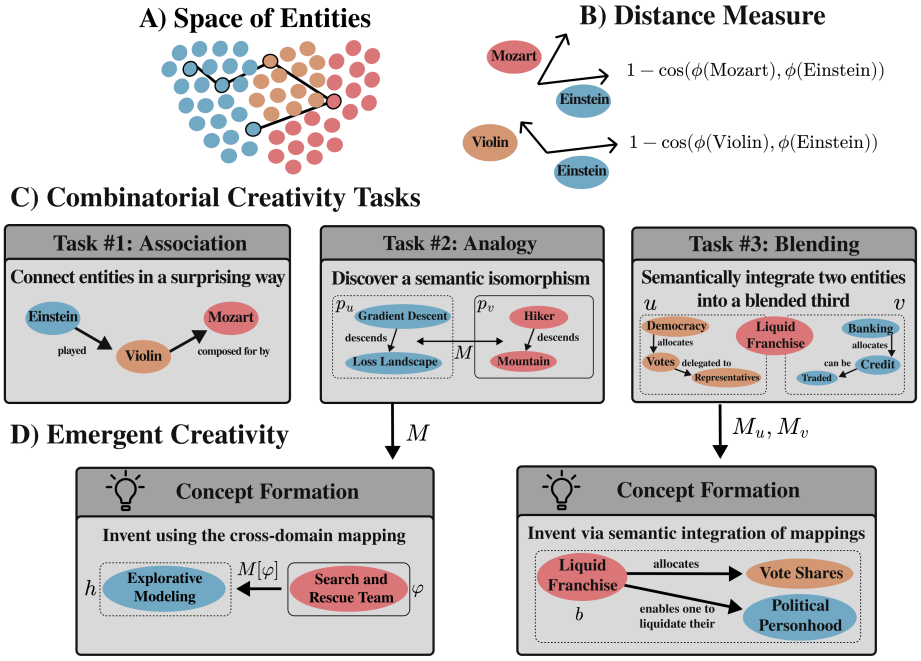


Figure 1: **An overview of the KOMBINE benchmark**. **(A)** Combinatorial creativity involves combining entities in original, useful, and surprising ways. **(B)** The semantic distance between entities is commonly used to calculate surprise and originality. **(C)** Combinatorial creativity largely falls into one of three types: (1) *Association*, (2) *Analogy*, and (3) *Blending*. **(D)** Analogies and blends support original concept invention via structure mapping and semantic integration.

own, an ability implicated in the formation of new concepts (Thagard, 1984; Marchiori and Scharp, 2024).

In this work, we introduce **KOMBINE**, a combinatorial creativity benchmark for LLMs tasking models to *discover* associations, analogies, and blends between random entity pairs with no known relationship (e.g., `a blue whale` and `a mattress`). Then, using these discovered relationships, models are tasked to invent novel concepts by re-combining the parts of existing ones, employing the structure mappings afforded by analogies and blends. In a large-scale study across 35 LLMs, including the latest frontier systems such as GPT-6 Astra and Claude Fable 5.1, we make five key findings: (1) Models are, on average, 10–15 percentage points better at discovering associations and analogies than blends. (2) Models exhibit an *abstraction bottleneck* on blending, in which they struggle to find shared abstractions between inputs in order to instantiate a coherent blended space. (3) Models also exhibit *inventive homogeneity*, in which one in five concepts invented for a given item reuses two or more properties invented by another model. (4) Further analysis reveals that inventive homogeneity is more pronounced among models from the same provider, which are $2.3\times$ more likely to reuse properties than those from different providers. (5) Moreover, models are $14\times$ more likely to reuse other invented properties on blends than analogies. In summary, KOMBINE offers an open-ended framework for tracking progress towards LLMs, and AI systems more broadly, capable of the combinatorial creativity that underlies scientific discovery and technological innovation.

2 Background and Related Work

Before introducing our benchmark, we start by providing background on creativity, describing how creativity relates to forms of generalization, and discussing related work.

Creativity and its Types Though creativity can refer to a person, process, product, or press (environment) (Rhodes, 1961), in the study of computationally creative systems, it is most common to adopt the product or process view (Varshney, 2019; Green et al., 2024), as we do here. In this case, the creative process involves *the generation of novel, useful, and surprising products* (Simonton, 2010; 2021; Boden, 2004; Varshney, 2019; Schapiro et al., 2025). Boden (2004) famously distinguishes between three types of creative processes: combinatorial, exploratory, and transformational creativity. The first models the creation of new artifacts as combinations of existing elements in a space of possible components, and is the focus of the present study. Consider recipe design (Varshney et al., 2019; Mizrahi et al., 2026), for example, where new recipes are generated by taking combinations of existing ingredients in varying proportions. Detailed descriptions of exploratory, transformational creativity, and other types of creativity (e.g., divergent, convergent thinking), are deferred to Appendix A. Next, we detail various forms of combinatorial creativity recognized in psychological literature.

Combinatorial Creativity Although Boden (2004) was the first to formally define combinatorial (or combinational) creativity as *the combination of familiar ideas in unfamiliar ways*, the study of this ability traces back to at least Hadamard (1954), who provided a survey of introspective accounts from famous mathematicians, scientists, and even musical composers, in which creative ideation was described as a combinatorial process. Mednick (1962) argued that creative thinking involved associating or combining mental representations, with more distant associations correlated with more creative artifacts, and designed the Remote Association Test (RAT) to assess this ability. Since then, empirical studies have found associative capacity to underpin human creativity, especially in scientific domains (Simonton, 2004; Beaty and Kenett, 2023; Benedek et al., 2023; Thagard, 2010; Simonton, 2021; Koestler, 1964). Hereafter, we refer to this ability as **associative creativity**. Subsequent work provided more structured accounts of how information can be combined across concepts and domains—such as through analogy (Rumelhart and Abrahamson, 1973; Petersen et al., 2026; Liu et al., 2026b), bisociation (Koestler, 1964), or blending (Fauconnier and Turner, 2008)—emphasizing the pivotal role of cross-domain mappings in scientific discovery and technological innovation (Gentner et al., 1997; Thagard, 1988; Holyoak and Thagard, 1994). In particular, Gentner's (1983) structure mapping theory formalized **analogical creativity** as systematic, structure-preserving alignment between concepts (e.g., `nucleus` is to `atom` as `sun` is to `solar system`) in ways which enable one to project structure from a source domain to infer or generate knowledge about a target domain. Similarly, **blending creativity** involves cross-domain structure mapping, and further defines selective mapping of two (or more) input spaces onto an additional space–referred to as *the blend*–which has its own emergent properties realizable by "running" or "elaborating" the blend (Fauconnier and Turner, 2008; Eppe et al., 2018). Both analogy and blending are implicated in the creation of new concepts via novel abstractions (Marchiori and Scharp, 2024; Fauconnier and Turner, 2008; Gentner and Smith, 2013; Mitchell, 2021).

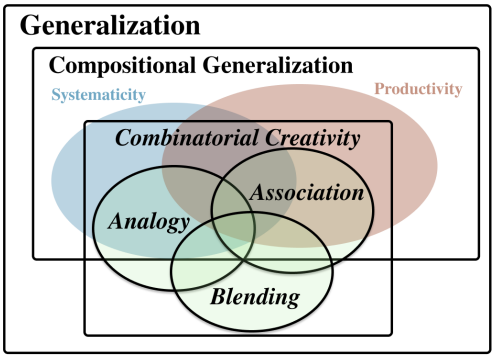


Figure 2: **Conceptual relationships between combinatorial creativity and compositional generalization.** Combinatorial creativity overlaps significantly with compositional generalization, though the two abilities are not identical. Analogy involves structure mapping between domains, which draws upon the systematicity to combine known parts and rules; association involves connecting entities in distant ways, often making longer connections between concepts than those made in training; while blending overlaps the least, involving abstraction ability not typically associated with compositionality.

Combinatorial Creativity and Compositional Generalization Among the five types of generalization studied in NLP research (Hupkes et al., 2023), *combinatorial creativity* most closely resembles

compositional generalization. Broadly, compositionality is a linguistic principle that the meaning of a complex expression is a function of the meaning of its parts and the way they are combined (Kim and Linzen, 2020; Fodor and Pylyshyn, 1988), and compositional generalization is typically divided into one of five types: systematicity, productivity, substitutivity, localism, and overgeneralization (Sinha et al., 2024; Hupkes et al., 2020). As shown in Figure 2, combinatorial creativity and compositional generalization are largely overlapping abilities, with association overlapping mostly with productivity (length generalization), analogy overlapping mostly with systematicity (the ability to recombine known parts and rules), and blending overlapping the least, as it involves abstraction and emergent elaboration not typically associated with compositional generalization. More broadly, combinatorial creativity, association, analogy, and blending are not fully subsumed by compositional generalization because creativity involves combining ideas in *unfamiliar and novel ways*, requiring open-ended evaluation uncharacteristic of compositional generalization tasks (cf. Lake and Baroni (2017) and Kim and Linzen (2020)).

Related Work Although there is no LLM benchmark for combinatorial creativity, previous works have each studied association, analogy, and blending separately. For example, the most closely related associative creativity benchmark over real-world entities is CREATE (Wadhwa et al., 2026), which assesses the ability to produce diverse and specific connections between entities with known relationships. However, CREATE does not score output originality—a necessary property of creative products (Maher, 2010; Boden, 2004; Simonton, 2023)—and indexes more heavily on closed-ended recognition than open-ended discovery ability due to pre-filtering entities with known connectivity. Psychometric tests of associative creativity, like the Divergent Association Test (DAT) and its variants (Olson et al., 2021; Nakajima et al., 2026; Qiu and Hu, 2025; Bellemare-Pepin et al., 2026), have also been studied extensively in LLMs, but often show poor external validity in predicting creative achievement on real-world tasks like scientific ideation, creative writing, and divergent thinking, casting doubt on their efficacy as psychometric instruments for LLMs (Schapiro et al., 2026a). Similarly, Nagarajan et al. (2025) studied the effect of pre-training objectives on creativity through synthetic algorithmic tasks loosely inspired by combinatorial creativity, but such tasks do not constitute a benchmark for LLMs. Next, there are numerous closed-ended tasks or benchmarks assessing analogical *reasoning*, the ability for LLMs to recognize or complete known analogies (Chen et al., 2022; Lewis and Mitchell, 2024; Wijesiriwardene et al., 2023), yet comparatively few large-scale studies of analogical *creativity*, the ability for LLMs to discover original analogies and employ them to meet task demands. The most closely related works are small-scale studies examining whether LLMs can generate original analogical mappings to enhance performance on design tasks (Liu et al., 2026a; Kokate and Onkar, 2026) or scientific ideation (Shen et al., 2026). Lastly, while no large-scale conceptual blending benchmark for LLMs exists, Wang et al. (2023) and Song et al. (2025) have studied or proposed methods to enhance conceptual blending ability at smaller scale. Most similar to our work, Song et al. (2025) study conceptual combination of noun phrases, finding that LLMs (as of 2025) struggle to generate noun phrases with emergent properties (such the "white color" emerging when `peeled` + `apple` are combined into "`peeled apple`").

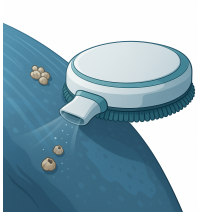
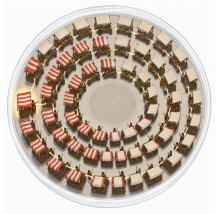
3 KOMBINE: AN OPEN-ENDED COMBINATORIAL CREATIVITY BENCHMARK

Here, we introduce a benchmark for combinatorial creativity that assesses its predominant forms—**association**, **analogy**, and **blending**—in an open-ended way. Unlike prior benchmarks which assess closed-ended recognition of known combinations, each task on KOMBINE involves discovering an association, analogy, or blend between a random pair of input entities u, v with no known relationship (Table 2). This makes our benchmark more representative of real-world discovery conditions in which novel connections relative to existing human knowledge are the primary desideratum (Swanson, 1986).[1] Using these discovered relationships, models are then tasked to *invent novel concepts* by re-combining the parts of existing ones, employing the structure mappings afforded by analogies and blends. We start by providing shared notation and setup, then formally define each task and its scoring, and lastly detail our experimental setup.

Preliminaries Combinatorial creativity occurs over a space of entities $\mathcal{C}$ and relations $\mathcal{R}$. An entity-relation-entity *triple* is a tuple $t = (c, r, c') \in \mathcal{C} \times \mathcal{R} \times \mathcal{C}$ whose truth or falsity is determined by a judge $F : \mathcal{C} \times \mathcal{R} \times \mathcal{C} \to \{0, 1\}$. Consistent with a large body of literature in psychometrics

[1] Though see the distinction between historical vs. personal creativity acknowledged in Section 5.

Table 2: **Example outputs on each task**. Colors indicate the domain of each input (u, v), the **invented entity**, and **emergent structure** of the blend. Images are for illustration purposes only.

Task (model)	Example output	Invention illustration
(a) Association (claude-opus-5)	**Inputs** (u, v): *Radio* → *The Milky Way*. **Path** (p): (Radio, aired hit, Vincent (song)) → (Vincent (song), tribute to, Vincent van Gogh) → (Vincent van Gogh, painted, The Starry Night) → (The Starry Night, swirls depict, The Milky Way).	N/A
(b) Analogy (claude-fable-5)	**Inputs** (u, v): *The blue whale* :: *The mattress*. **Path a** (p_u): (The blue whale, covered by, epidermis), (epidermis, hosts, barnacles), (barnacles, feed on, plankton). **Path b** (p_v): (The mattress, covered by, ticking), (ticking, hosts, dust mites), (dust mites, feed on, skin flakes). **Source** (φ): vacuum cleaner (from The mattress, no counterpart in The blue whale). **Invention** $(h \leftarrow M[\varphi])$: **whale groomer drone**. **Source to Target Mapping** $(M^{u,v})$: (vacuum cleaner, removes, dust mites) ↦ (whale groomer drone, removes, barnacles) (vacuum cleaner, cleans, ticking) ↦ (whale groomer drone, cleans, epidermis) (vacuum cleaner, maintains, The mattress) ↦ (whale groomer drone, maintains, The blue whale)	
(c) Blending (claude-opus-4.5)	**Inputs** (u, v): *Adam Smith* + *Bacteria*. **Generic Space** (g): autonomous agents whose local signals generate collective order without central control. **Invention** (b): **market colony**. **Blended Space** (c'), each row tagged by its projection: (market colony, coordinates through, local signals) [uv] (market colony, produces, emergent efficiency) [uv] (market colony, specializes via, division of labor) [u] (market colony, adapts through, replication) [v] (market colony, evolves its, exchange protocols) [*emergent*] (market colony, spreads, practices contagiously) [*emergent*]	

and computational creativity that assesses semantic divergence via the embedding distance between elements (Olson et al., 2021; Johnson et al., 2023; Varshney et al., 2020), we define a text embedding function ϕ and let $d^{\phi}_{\cos} : \mathcal{C} \times \mathcal{C} \to [0, 1]$ denote its induced cosine distance measure. For an added element e of a response to inputs $u, v \in \mathcal{C}$ (i.e., any $e \in \mathcal{R} \cup \mathcal{C} \setminus \{u, v\}$), $\rho(e)$ is its mean cosine distance to its m nearest neighbors among the distinct added elements that all models produced for the same item, so that an element is original when nothing semantically close to it was produced by any model—including that same model's previous items for association, so originality implicitly captures diversity (Jiang et al., 2026; Zhang et al., 2025).[2] An artifact's originality is then the mean of ρ over all its added elements. Each task is scored according to utility, surprise, and originality (novelty), which are established criteria for creative products in the psychology literature (Maher, 2010; Simonton, 2004; Boden, 2004). On all three tasks, utility is a binary indicator of whether the task-specific constraints are met, consistent with psychometric practice that scores this dimension as either satisfied or not (Maher, 2010; Nakajima et al., 2026). On analogy and blending, we also assess the ***emergent creativity*** of invented concepts according to their originality O, coherence and feasibility V, and semantic integration quality Q. Both V and Q are assessed using LLM judges whose prompts are given in Appendix D. We now formally define each task and its scoring.

Task #1 (Association): The association task assesses the ability to connect two distant entities otherwise believed to be unrelated, by discovering a surprising, original, and factual path of triples between them. On association, models are tasked to provide as many distinct paths as possible, and each dimension is averaged across all paths.

Association Task

Given inputs u, v, an *association* is a path of triples $p = \big((u, r_1, c_1), \ldots, (c_{n-1}, r_n, v)\big)$ connecting them ($c_0 = u$, $c_n = v$). It is *useful* $U_{\text{as}}(p) := \prod_{t \in p} F(t) \in \{0, 1\}$ only when every triple is factual, its *surprise* $S_{\text{as}}(p) := \frac{1}{n-1} \sum_i d^{\phi}_{\cos}(c_{i-1}, c_i)$ is the average semantic distance between consecutive entities in the path, and its *originality* $O_{\text{as}}(p) := \frac{1}{|E_p|} \sum_{e \in E_p} \rho(e)$ is the mean distance of its added elements E_p from the response pool.

[2]By default, we use `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) for ϕ, but Appendix E reports robustness across varying embedding models. Moreover, originality uses $m = 5$ nearest neighbors.

Task #2 (Analogy): The analogy task involves discovering relational structure shared by the two unrelated inputs, and then using this mapping to invent a new concept relevant to one of the inputs. Formally, a pair of length-n paths $p_u = \big((u, r_1, a_1), \ldots, (a_{n-1}, r_n, a_n)\big)$ and $p_v = \big((v, q_1, b_1), \ldots, (b_{n-1}, q_n, b_n)\big)$ ($a_0 = u$, $b_0 = v$) that share a relation sequence ($r_i \equiv q_i$) define a *structure mapping* $M^{u,v} : \mathcal{C} \to \mathcal{C}$ from each entity of p_u to its counterpart in p_v, $M(a_i) := b_i$, which extends to triples and to sets or paths of triples T according to $M[(c, r, c')] := (M(c),\ r,\ M(c'))$ and $M[T] := \{M[t] : t \in T\}$. Consistent with structure mapping theory, the mapping is semantically valid only if every triple in each path is factual and the two paths satisfy relational similarity.

Analogy Task

Given inputs u, v, an *analogy* $a = (p_u \cup p_v, h)$ consists of a pair of paths p_u, p_v that define a structure mapping M, and a concept $h := M[\Phi]$ invented by projecting a source domain entity φ and its structure $\Phi := \{(\varphi, r_j, s_j)\}_{j=1}^{k}$ through the mapping. Its *utility* $U_{\text{an}}(a) := \prod_{i=1}^{n} F(p_u[i])F(p_v[i]) \cdot \prod_{i=1}^{n} \mathbb{I}[r_i \equiv q_i] \in \{0, 1\}$ holds when the mapping M is semantically valid, its *surprise* $S_{\text{an}}(a) := \frac{1}{n+1} \sum_i d^{\phi}_{\cos}(a_i, b_i)$ is the average distance between entities in corresponding roles of each path, and its *originality* $O_{\text{an}}(a)$ is defined as for association over its added elements E_a. Lastly, the ***emergent creativity*** of h is scored by the invented concept's originality $O(h)$, utility $V(h) \in \{0, 1\}$, and semantic integration quality $Q(h) \in \{0, 1\}$, which indicates whether the mapping was actually used to construct h.

Task #3 (Blending): Lastly, the blending task involves discovering abstract structure both inputs instantiate (referred to by Fauconnier and Turner (2008) as a *generic space*) and fusing them into a new space which inherits structure from both inputs. Blends allow one to "elaborate" in this blended space to describe emergent properties, such as a market colony that "evolves its exchange protocols" (Table 2). Following Fauconnier and Turner (2008), a blend is *single-scope* when both inputs contribute projected structure, *double-scope* when both intersect on at least one shared property, and *double-scope emergent* when a double-scope blend also elicits an emergent property.

Blending Task

Given inputs u, v, a *blend* $a = (g, c')$ consists of a generic space g that both inputs instantiate and a blended space describing an invented concept $c' := M_u[\Phi_u] \cup M_v[\Phi_v] \cup \Delta$, into which mappings $M^{u,c'}, M^{v,c'}$ selectively project triples $\Phi_u := \{(u, r_j, s_j)\}_j$ and $\Phi_v := \{(v, q_j, t_j)\}_j$ true of each input, and where $\Delta := \{t \in c' : u, v \not\models t\}$[3] is emergent structure true of the blend but of neither input. Its *utility* $U_{\text{bl}}(g) := G(g) \in \{0, 1\}$ is scored by a judge G of whether the generic space g fits both inputs, its *surprise* $S_{\text{bl}}(g) := \frac{1}{2}\big(d^{\phi}_{\cos}(u, g) + d^{\phi}_{\cos}(v, g)\big)$ is the mean distance of g from the two inputs, and its *originality* $O_{\text{bl}}(g) := \rho(g)$ is the distance of g from the generic spaces in other responses to the same item.[4] The ***emergent creativity*** of c' is scored by its originality $O(c')$, utility $V(c') \in \{0, 1\}$ (whether each triple $t \in c'$ is feasible or coherent), and integration quality $Q(c') := 1$ if $M_u[\Phi_u], M_v[\Phi_v] \neq \emptyset$ (single-scope), 2 if additionally $M_u[\Phi_u] \cap M_v[\Phi_v] \neq \emptyset$ (double-scope), and 3 if additionally $\Delta \neq \emptyset$ (double-scope emergent).

Inventive Multiples Given that LLMs have a documented tendency to homogenize on open-ended tasks (Jiang et al., 2026), we also study the degree of conceptual overlap between inventions made from the analogy and blending tasks. By analogy to *scientific multiples* in the science of science (Simonton, 2021; 2004), we refer to near-identical inventions (made by different models) given the same inputs $(u, v) \in \mathcal{C}^2$ as *inventive multiples*, and formally define them as follows.

Definition 1 (τ-Inventive Multiples)**.** Let $a, a' \in \mathcal{C}$ be two concepts invented via analogy or blending given the same input entities $u, v \in \mathcal{C}$. Assume a, a' are described by property sets $P = \{p_j\}_{j=1}^{k}$ and $P' = \{p'_j\}_{j=1}^{k'}$, respectively, where each property $p_j, p'_j \in \mathcal{R} \times \mathcal{C}$ is a relation-object tuple (e.g., (`can be, traded`)). Given a similarity threshold $\theta > 0$, two properties are semantically indistinguishable, written $p \approx p'$, whenever $\cos\big(\phi(p), \phi(p')\big) \geq \theta$. Then, a and a' are **τ-inventive multiples** if at least $\tau \in \mathbb{N}$ distinct properties of a (i.e., $(p_1, \ldots, p_\tau)$) have unique, semantically indistinguishable counterparts in a' (i.e., $(p'_1, \ldots, p'_\tau)$ such that $p_i \approx p'_i, \forall i \leq \tau$).

[3] Here, $\models$ refers to the textual entailment operator (i.e., "text T entails hypothesis H if a human reading T would infer H is most likely true.") (Dagan et al., 2005). In this case, we refer to opposite entailment (e.g., a human or LLM judge reading T would *not* infer H is true).

[4] Note that the generic space is a text piece, so originality here is computed over the sentence embedding, rather than over relations and concepts individually as in analogy and association.

Table 3: **Top systems by combinatorial creativity.** The highest-scoring systems on *Overall* and on each task, as the equal-weight composite of each scoring dimension in Section 3.

Rank	Overall	Association	Analogy	Blending
1	gpt-5.6-sol (55.3)	claude-opus-4.6 (51.4)	claude-opus-4.6 (63.4)	gpt-5.6-sol (64.9)
2	gpt-6-astra (54.5)	gpt-5.2 (50.3)	grok-4.6 (62.4)	gemini-3.1-pro (59.7)
3	claude-opus-4.6 (52.1)	gemini-2.5-flash (48.3)	gpt-6-astra (61.2)	gpt-6-astra (56.6)
4	gemini-3.1-pro (50.5)	claude-opus-4.5 (48.1)	gpt-5 (58.8)	gemini-3.7-flash (53.0)
5	grok-4.6 (48.4)	gpt-5-mini (48.0)	grok-4.5 (55.6)	claude-fable-5 (45.7)

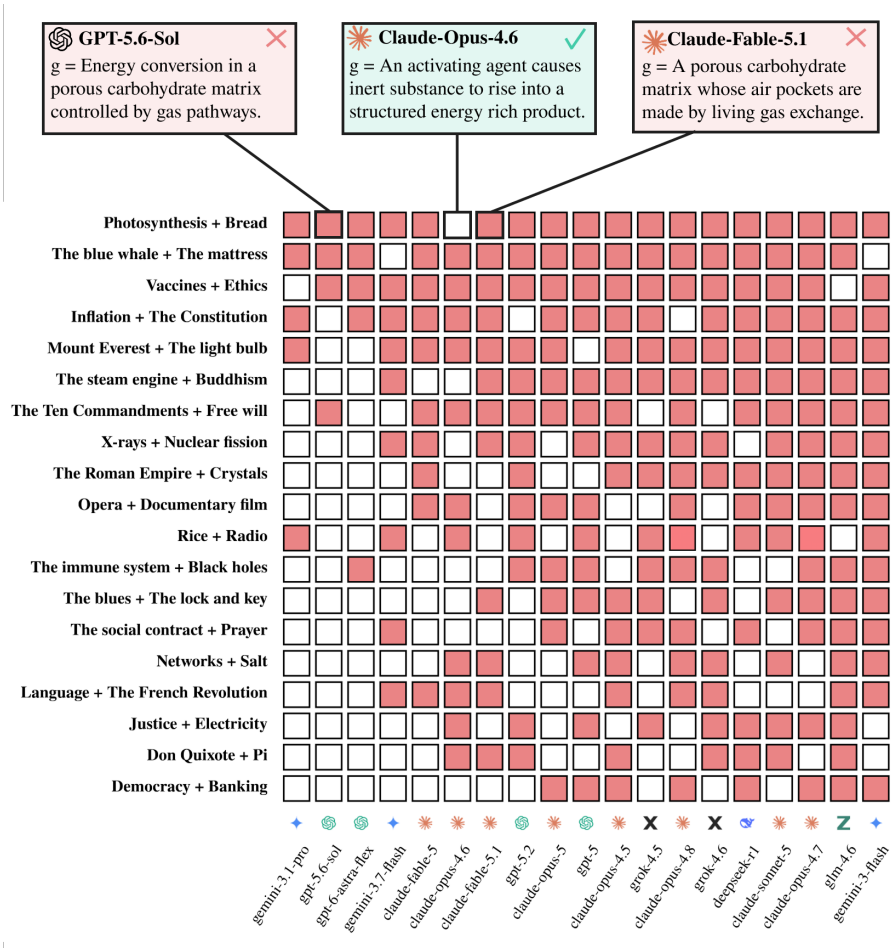


Figure 3: **Blending generic space failures among frontier models.** A red cell marks an item on which a model's generic space was invalid, often due to applying only to one input. Although blending is difficult for LLMs, no row is fully red, i.e., every entity pair had *at least one model* obtain a valid generic space.

Experimental Setup We study a set of 35 LLMs, comprising a mix of the latest frontier systems (e.g., GPT-6 Astra, Claude Fable 5.1) and smaller, open-weight models. All models are sampled at temperature $T = 0.9$ with top-p and top-k parameters disabled and reasoning level left at default. To instantiate the space of entities $\mathcal{C}$, we randomly sample entities from the Wikidata knowledge graph, restricted to recognizable entities with at least $k = 78$ sitelinks, as explained in Appendix B. The full list of entities is given in Table 5. We use `Claude Haiku 4.5` as the LLM factuality judge F, and a panel of three LLMs (`Claude Haiku 4.5`, `gpt-o3`, and `GPT-5.4`) for all other analogy and blending judges, with prompts given in Appendix D. Analogy and blend judges achieve fair-to-good reliability, and they agree with human judgment 75% of the time on items for which the three-judge LLM panel is unanimous. Furthermore, restricting the leaderboard to items where the LLM judge panel is unanimous leaves the model ranking almost entirely the same (see Appendix E).

4 RESULTS AND KEY FINDINGS

Table 3 reports the top five models overall, and on each of the three tasks (see Table 7 for the full ranking of all 35 models). Each task score is the mean of that task's scoring dimensions, reported as a fraction of each dimension's maximum, and the overall score is the mean of the three task scores. Importantly, on artifacts that fail utility, every dimension is set to zero. Although association and analogy correlate at $r = 0.72$ ($p < 0.0001$), analogy and blending at $r = 0.70$ ($p < 0.0001$) and blending and association at $r = 0.53$ ($p < 0.001$), the top performing models are task-specific with `claude-opus-4.6` leading on association and analogy, but `gpt-5.6-sol` leading on blending (Table 3). The rankings are also robust to the choice of embedding model, as detailed in Appendix G.

Association and analogy are easiest for LLMs, while blending is the hardest Averaged over the 35 LLMs, models produce a factual association 59.1% of the time (Table 8), a valid analogy ($U_{\text{an}}(a) = 1$ in the Analogy Task) 54.0% of the time (Table 9), and a valid blend ($U_{\text{bl}}(a) = 1$ in the Blending Task) 43.8% of the time (Table 10).[5] The task difficulty scales in the same order as the amount of structure each task demands, with a factual path being easiest, a structure mapping being slightly harder, and a shared abstraction noticeably more difficult.

LLM abstraction ability is the key bottleneck in blending Although blending is most difficult for LLMs on average, all of the entity pairs (u, v) had at least one model discover a valid blend (see Figure 3), so no item was impossible to blend. For `(Photosynthesis, Bread)`, Opus 4.6 gave the only valid generic space ($g =$ `An activating agent causes inert substance to rise into a structured energy rich product`), while all other frontier models failed (e.g., Fable 5.1 said $g =$ `a porous carbohydrate matrix whose air pockets are made by living gas exchange`, which only applies to bread). Overall, as shown in Table 10, models can find a generic space only 43.8% of the time on average, but conditioned on finding a valid generic space, models can describe a coherent blended concept 97.1% of the time, a finding we refer to as the *abstraction bottleneck*. One plausible explanation for why this occurs is lack of experiential knowledge, as human neuro-imaging studies implicate mental imagery and episodic simulation in creative association and conceptual combination (Benedek et al., 2020).

Association originality is the best predictor of every other facet Figure 11 showcases facet correlations across all three core scoring dimensions in Section 3. LLMs which find more original ways to connect distant entities otherwise believed to be unrelated, also tend to find more original ($r = 0.72$, $p < 1\text{e}{-4}$) and surprising analogies ($r = 0.62$, $p = 1\text{e}{-4}$), more original ($r = 0.39$, $p = 2\text{e}{-2}$) and surprising blends ($r = 0.50$, $p = 2\text{e}{-3}$), and valid blends more often ($r = 0.48$, $p = 4\text{e}{-3}$). Overall, association originality has the highest mean cross-task correlation of any other facet ($\bar{r} = 0.47$), consistent with Mednick's (1962) associative theory of creativity—stipulating that the ability to connect distant entities in original ways is central to most types of creativity. As detailed in Appendix G, these relationships hold across varying embedding model choices.

One in five invented concepts is a multiple Across all 35 LLMs, a total of 2,070 concepts were invented through analogies and blends, and these concepts were described by 7,868 total (relation, object) properties. To determine which of these are multiples, we set $\theta = 0.674$ as the similarity threshold in Definition 1, which corresponds to the 99th percentile of cosine similarity over all property pairs. Figure 4(a) showcases the proportion of invented concepts containing $\tau \in \{1, 2, 3\}$ semantically indistinguishable properties from another model's concept invented for the same item. At $\tau = 2$, 19.9% of the invented concepts (412/2,070) is a multiple.

Models are $3{-}14\times$ more likely to produce multiples on blends than analogies As shown in Figure 4(b), a concept invented by blending is $2.6\times$ more likely to be a (τ=1)-inventive multiple than one invented by analogy (94.7% vs. 36.3%), and $14.3\times$ more likely to be a (τ=2)-inventive multiple (37.3% vs. 2.6%).

Models from the same family invent alike As shown in Figure 4c, models from the same family are $2.3\times$ more likely to invent (τ=2)-multiples than models from different providers (9.3% vs. 4.1%). In Figure 12, we report the intra-family multiple rate as a function of model provider. To normalize the rate of multiples among providers with unequal numbers of models in the pool, we compare the rate of inventive multiples within a single provider's k models, relative to that provider's average rate

[5] Note, however, that this aggregate statistic is not true of every model, as `gpt-5.6-sol` scores 64.9% on blending but only 55.6% on analogy.

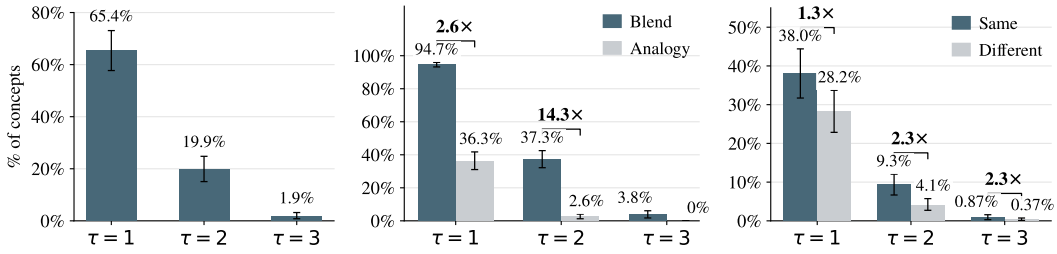
Table 4: **Examples of two inventive multiples.** Semantically indistinguishable properties in are placed side-by-side and highlighted. *Operators*: Blending (**+**), Analogy (**::**).

Opera + Documentary film

✳ claude-fable-5.1 **"testimony opera"**	cos	✦ gemini-3.7-flash **"docu-opera"**
p_2: intercuts archival footage	0.87	p'_2: incorporates archival footage
p_4: derives melody from recorded speech	0.81	p'_4: derives melody from spoken pitch contours
p_3: halts action for sung testimony	0.77	p'_3: delivers sung verbatim testimony
p_1: accompanied by orchestra		p'_1: utilizes orchestral overtures

Democracy :: Banking

✳ claude-fable-5 **"vote insurance"**	cos	✳ claude-fable-5.1 **"legitimacy stress test"**
p_1: protects ballot	0.84	p'_3: protects votes
p_5: mandated by election commission	0.70	p'_4: is overseen by electoral commission
p_2: restores popular will		p'_1: simulates recall election
p_3: triggered by election annulment		p'_2: measures mandate
p_4: prevents voter boycott		



(a) Concepts in a multiple (b) By task (c) By family

Figure 4: **Rate of inventive multiples. (a)** The rate of all 2,070 inventions that belong to at least one τ-multiple, aggregated by task **(b)** and by model family **(c)**, with task and family ratios above each τ. Error bars are 95% confidence intervals from a bootstrap over items.

of multiples with k different-provider models in the pool, and plot the difference as a function of provider in Figure 12. DeepSeek and Anthropic models exhibit the greatest intra-family multiple rate at 12.2 and 11.1 percentage points (pp) above baseline, while OpenAI models exhibit the least, converging with each other 1.4pp *less often* than they do with models from other providers.

5 Limitations and Future Work

This work leaves several directions to future research. First, there is no broad consensus in the cognitive science literature on the degree of overlap between analogy and blending, and structure mapping theory is comprehensive enough to account for operations shared in both processes (Gentner and Smith, 2013). The core distinction is that the primary artifact of a blend is a blended space (Fauconnier and Turner, 2008), whereas the outcome of analogy is typically the target domain projection of an element that already exists in the source domain (Petersen et al., 2026), although Figure 42.2(C) in Gentner and Smith (2013) portrays a novel abstraction that can arguably be considered a type of blend. Therefore, analogy and blending should not be conceptualized as mutually exclusive forms (and, hence, are pictured as overlapping in Figure 2), but rather as interdependent processes that co-occur empirically, with structure mapping involved in constructing a blend. Future work can explore the relationships between these abilities in greater detail. Similarly, although blending can involve more than two input spaces (Fauconnier and Turner, 2008), we restrict to two in this work. Next, although humans employ combinatorial creativity to create visual art and music (Boden, 2004), our benchmark is restricted to the textual domain. Lastly, Boden (2004) defines historical creativity (H-creativity) as novelty relative to human knowledge writ large, and personal creativity (P-creativity) as novelty relative only to an artifact's creator. Our benchmark randomly samples knowledge graph entities for which prior relationships are not established, and thus approaches H-creativity, but our originality measures cannot guarantee the historical novelty of all artifacts created. Therefore, our benchmark assesses a level of creativity situated between P- and H-creativity.

Reproducibility Statement

To ensure reproducilibity of results, the entities and procedure used to construct the benchmark are fully listed in Appendix B, the task prompts are given in Appendix C, and the judge prompts are given in Appendix D. Every model is elicited under a recorded configuration giving the model identifier and the sampling temperature ($T = 0.9$, with top-p and top-k disabled). Embedding-based scoring dimensions use `all-MiniLM-L6-v2`, although the leaderboard is highly robust to the choice of embedding model, as shown in Appendix G.

AI Use Statement

In this work, we used generative AI coding agents to implement methods, to clean and reformat the dataset of model responses, to provide feedback on research methodology and experiments, and to interpret results. We have not used generative AI tools to generate synthetic data sets, to help develop theoretical models or conceptual frameworks, to formulate mathematical claims, to propose or refine hypotheses, to support qualitative and thematic data analysis, provide critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, to create and modify scientific figures, to create artifacts (the illustrations in Table 2) from prompts written by the authors), to draft parts of the paper, to identify and summarize relevant literature, to format references, and to edit the paper for readability. All AI-drafted work was reviewed by the authors, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

A. H. Artiles, M. Weiss, L. Brinkmann, I. Rahwan, B. Schölkopf, C. Pal, H. Larochelle, A. Goyal, and N. Rahaman. The alien space of science: Sampling coherent but cognitively unavailable research directions, 2026. URL https://arxiv.org/abs/2603.01092.

R. E. Beaty and Y. N. Kenett. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683, 2023.

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

M. Benedek, J. Jurisch, K. Koschutnig, A. Fink, and R. E. Beaty. Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes. *NeuroImage*, 210:116586, 2020.

M. Benedek, R. E. Beaty, D. L. Schacter, and Y. N. Kenett. The role of memory in creative ideation. *Nature Reviews Psychology*, 2(4):246–257, 2023.

M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

J. Chen, R. Xu, Z. Fu, W. Shi, Z. Li, X. Zhang, C. Sun, L. Li, Y. Xiao, and H. Zhou. E-kar: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 3941–3955, 2022.

D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290, 1994.

I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer, 2005.

A. Dietrich. The cognitive neuroscience of creativity. *Psychonomic bulletin & review*, 11(6): 1011–1026, 2004.

A. Dietrich. Types of creativity. *Psychonomic Bulletin and Review*, 26(1):1–12, 2 2019. ISSN 15315320. doi: 10.3758/s13423-018-1517-7.

M. Eppe, E. Maclean, R. Confalonieri, O. Kutz, M. Schorlemmer, E. Plaza, and K.-U. Kühnberger. A computational framework for conceptual blending. *Artificial Intelligence*, 256:105–129, 2018.

H. Fang and J. Evans. Generalization and the rise of system-level creativity in science. *arXiv preprint arXiv:2510.03240*, 2025.

G. Fauconnier and M. Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2008.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

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

D. Gentner and L. A. Smith. Analogical learning and reasoning. 2013.

D. Gentner, S. Brem, R. Ferguson, and P. Wolff. Analogy and creativity in the works of johannes kepler. 1997.

A. E. Green, R. E. Beaty, Y. N. Kenett, and J. C. Kaufman. The process definition of creativity. *Creativity Research Journal*, 36(3):544–572, 2024.

T. Gu, J. Wang, Z. Zhang, and H. Li. Llms can realize combinatorial creativity: Generating creative ideas via llms for scientific research, 2025. URL https://arxiv.org/abs/2412.14141.

J. Hadamard. *An Essay on the Psychology of Invention in the Mathematical Field*. Courier Corporation, 1954.

K. Haven. *100 Greatest Science Discoveries of All Time*. Bloomsbury Publishing USA, 2007.

K. J. Holyoak. *The human edge: Analogy and the roots of creative intelligence*. MIT Press, 2025.

K. J. Holyoak and P. Thagard. *Mental leaps: Analogy in creative thought*. The MIT Press, 1994.

D. Hupkes, V. Dankers, M. Mul, and E. Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmer, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, and Z. Jin. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, Oct. 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00729-y. URL http://dx.doi.org/10.1038/s42256-023-00729-y.

L. Jiang, Y. Chai, M. Li, M. Liu, R. Fok, N. Dziri, Y. Tsvetkov, M. Sap, and Y. Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Advances in Neural Information Processing Systems*, 38, 2026.

D. R. Johnson, J. C. Kaufman, B. S. Baker, J. D. Patterson, B. Barbot, A. E. Green, J. van Hell, E. Kennedy, G. F. Sullivan, C. L. Taylor, et al. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods*, 55 (7):3726–3759, 2023.

N. Kim and T. Linzen. COGS: A compositional generalization challenge based on semantic interpretation. arXiv:2010.05465, 2020.

A. Koestler. *The Act of Creation*. Macmillan, 1964.

R. Kokate and P. Onkar. Analogical reasoning with large language models: a co-creative framework and benchmarking of llms in design ideation. *Design Science*, 12:e4, 2026. doi: 10.1017/dsj.2025.10044.

B. M. Lake and M. Baroni. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. arXiv:1711.00350, 2017.

G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago press, 2008.

M. Lewis and M. Mitchell. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215*, 2024.

Q. E. Liu, M. Dubova, H. Conklin, T. Harada, and T. L. Griffiths. Assessing the effect of cross-domain mapping on creativity in humans and large language models, 2026a. URL https://arxiv.org/abs/2603.19087.

Q. E. Liu, R. Marjieh, J.-Q. Zhu, A. E. Goldberg, and T. L. Griffiths. Large language models provide support for the parallelogram theory of analogy. *arXiv preprint arXiv:2603.19066*, 2026b.

M. L. Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pages 22–28, 2010.

S. Marchiori and K. Scharp. What is conceptual disruption? *Ethics and Information Technology*, 26 (1), 3 2024. ISSN 15728439. doi: 10.1007/s10676-024-09749-7.

S. Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220, 1962.

M. Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

M. Mizrahi, C. Shani, G. Stanovsky, D. Jurafsky, and D. Shahaf. Cooking up creativity: Enhancing llm creativity through structured recombination. *Transactions of the Association for Computational Linguistics*, 14:418–441, 2026.

V. Nagarajan, C. H. Wu, C. Ding, and A. Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. arXiv:2504.15266, 2025.

K. Nakajima, J. Zuiderveld, and S. Pezzelle. Beyond divergent creativity: A human-based evaluation of creativity in large language models, 2026. URL https://arxiv.org/abs/2601.20546.

J. A. Olson, J. Nahas, D. Chmoulevitch, S. J. Cropper, and M. E. Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021.

M. R. Petersen, C. E. Stevenson, and L. van der Plas. Modelling analogies and analogical reasoning: Connecting cognitive science theory and nlp research. *Transactions of the Association for Computational Linguistics*, 14:711–732, 2026.

Z. Qiu and R. Hu. Deep associations, high creativity: A simple yet effective metric for evaluating large language models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10872, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.550. URL https://aclanthology.org/2025.emnlp-main.550/.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

M. Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961.

D. E. Rumelhart and A. A. Abrahamson. A model for analogical reasoning. *Cognitive Psychology*, 5 (1):1–28, 1973.

A. Sanyal, S. Schapiro, S. Shashidhar, R. Moon, L. R. Varshney, and D. Hakkani-Tur. Spark: A system for scientifically creative idea generation. *arXiv preprint arXiv:2504.20090*, 2025.

S. Schapiro, J. Black, and L. R. Varshney. Transformational creativity in science: A graphical theory. *arXiv preprint arXiv:2504.18687*, 2025.

S. Schapiro, A. Gladstone, J. Black, and H. Ji. Assessing the creativity of large language models: Testing, limits, and new frontiers, 2026a. URL https://arxiv.org/abs/2605.13450.

S. Schapiro, A. Gladstone, H. Ji, and R. E. Beaty. Twistbench: Benchmarking transformational creativity in LLMs via literary plot twists. In *Workshop on Scientific Understanding of Foundation Models*, 2026b. URL https://openreview.net/forum?id=dFgWvK2YaX.

A. Shen, S. Druckmann, and J. Zou. Unlocking llm creativity in science through analogical reasoning, 2026. URL https://arxiv.org/abs/2605.11258.

F. Shi and J. Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14, 03 2023. doi: 10.1038/s41467-023-36741-4.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

D. K. Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, 2004.

D. K. Simonton. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews*, 7(2):156–179, 2010.

D. K. Simonton. Scientific creativity: Discovery and invention as combinatorial. *Frontiers in Psychology*, 12:721104, 2021.

D. K. Simonton. The blind-variation and selective-retention theory of creativity: Recent developments and current status of bvsr. *Creativity Research Journal*, 35(3):304–323, 2023.

S. Sinha, T. Premsri, and P. Kordjamshidi. A survey on compositional learning of AI models: Theoretical and experimental practices. arXiv:2406.08787, 2024.

S. Song, T. Lee, J. Ahn, J. H. Sung, and G. Kim. Is a peeled apple still red? evaluating llms' ability for conceptual combination with property type. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9028–9048, 2025.

D. R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

P. Thagard. Conceptual combination and scientific discovery. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1984, pages 2–12. Cambridge University Press, 1984.

P. Thagard. *Computational philosophy of science*. MIT Press, 1988. ISBN 0262200686.

P. Thagard. Creative combination of representations: Scientific discovery and technological invention. *Psychology of science, R. Proctor and EJ Capaldi, Eds Oxford University Press, forthcomin*, 2010.

P. Thagard. *Conceptual Revolutions*. Princeton University Press, 2018.

L. R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1):2:1–2:12, 2019. doi: 10.1147/JRD.2019.2893907.

L. R. Varshney, F. Pinel, K. R. Varshney, D. Bhattacharjya, A. Schörgendorfer, and Y.-M. Chee. A big data approach to computational creativity: The curious case of Chef Watson. *IBM Journal of Research and Development*, 63(1):7–1, 2019.

L. R. Varshney, N. F. Rajani, and R. Socher. Explaining creative artifacts. In *ICML 2020 Workshop on Human Interpretability in Machine Learning (WHI)*, 2020.

M. Wadhwa, T. S. Roy, H. Lederman, J. J. Li, and G. Durrett. Create: Testing llms for associative creativity, 2026. URL https://arxiv.org/abs/2603.09970.

L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Multilingual e5 text embeddings: A technical report, 2024. URL https://arxiv.org/abs/2402.05672.

S. Wang, S. Petridis, T. Kwon, X. Ma, and L. B. Chilton. Popblends: Strategies for conceptual blending with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

T. Wijesiriwardene, R. Wickramarachchi, B. Gajera, S. Gowaikar, C. Gupta, A. Chadha, A. N. Reganti, A. P. Sheth, and A. Das. Analogical-a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, 2023.

S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649, 2024.

Y. Zhang, H. Diddee, S. Holm, H. Liu, X. Liu, V. Samuel, B. Wang, and D. Ippolito. Noveltybench: Evaluating language models for humanlike diversity, 2025. URL https://arxiv.org/abs/2504.05228.

A Additional Background

Exploratory and Transformational Creativity Exploratory and transformational creativity are historically defined with respect to a "conceptual space," a set of rules and constraints that defines what constitutes well-defined and intelligible artifacts in a particular domain. Exploratory creativity refers to artifacts generated by following these rules and constraints (such as AlphaGo move 37 (Silver et al., 2017)) whereas transformational creativity, which refers to the more challenging task of re-structuring the very rules of a conceptual space, is considered the pinnacle form of creativity for its historical role in breakthrough innovation (Boden, 2004). Famous examples of transformational creativity include Einstein's relativity theory, the shift from geocentrism to heliocentrism, and the invention of non-Euclidean geometry (Haven, 2007; Schapiro et al., 2025; Thagard, 2018; Koestler, 1964), though transformational creativity is also implicated in other creative areas—such as in the literary domain when producing plot twists (Schapiro et al., 2026b).

Divergent and Convergent Thinking Another popular taxonomy of creative processes distinguishes between divergent and convergent thinking (Dietrich, 2019; 2004). Broadly speaking, divergent thinking refers to the ability to generate multiple diverse stimuli on an open-ended task, while convergent thinking is the ability to generate a single response that unifies multiple diverse stimuli. Divergent and convergent creativity are commonly assessed using psychometric tests such as the Divergent Association Task (Olson et al., 2021), Remote Associates Test (Mednick, 1962), and more recently, Parallel Associative Chains for Creativity Evaluation (Qiu and Hu, 2025), the Conditional Divergent Association Task (Nakajima et al., 2026), and the Divergent Remote Association Task (Schapiro et al., 2026a). However, as shown in Schapiro et al. (2026a), psychometric tests designed for human populations often show poor external validity when applied to LLMs, casting doubt on their efficacy as instruments for studying LLM intelligence.

B Entity Pool

As shown in Table 5, the space of entities $\mathcal{C}$ consists of 283 entities spread across 21 domains. Entities were filtered for their recognizability, and each entity has a median of 164 Wikimedia sitelinks, with 94% at ≥ 50 and a minimum of 78. In contrast, entities drawn uniformly at random from Wikidata have a *median of zero* sitelinks and a 90th percentile of 2. Furthermore, domains are balanced by construction, at 10–16 entities each, and the entity input pairs for each of the three tasks consists of pairs from different domains.

Table 5: **The entity pool $\mathcal{C}$.** All 283 entities, categorized by domain.

Domain	Entities
science	Gravity, Evolution, Entropy, Photosynthesis, Electricity, Radioactivity, The atom, Quantum mechanics, The theory of relativity, Plate tectonics, The Big Bang, Natural selection, The periodic table, Nuclear fission, Magnetism, The speed of light
mathematics	Zero, Infinity, Pi, Prime numbers, The Pythagorean theorem, Calculus, Probability, Geometry, The golden ratio, Fractals, Chaos theory, Set theory, Logarithms, Fibonacci sequence
technology	The printing press, The steam engine, The internet, The telephone, The computer, The transistor, The automobile, The airplane, The light bulb, The wheel, Artificial intelligence, The battery, Radio, GPS, The camera, The clock
biology	DNA, The cell, The brain, The heart, Bacteria, Viruses, The immune system, Photosynthesis, The human genome, Coral reef, The honey bee, The lion, The octopus, The oak tree, Mushrooms, The blue whale
medicine	Penicillin, Vaccines, Anesthesia, The stethoscope, Antibiotics, Surgery, X-rays, Insulin, The Hippocratic Oath, Blood transfusion, Aspirin, The microscope
art	The Mona Lisa, Impressionism, Cubism, Leonardo da Vinci, Pablo Picasso, Vincent van Gogh, Frida Kahlo, Michelangelo, The Sistine Chapel, Surrealism, Sculpture, Andy Warhol, Calligraphy, Origami
music	Jazz, The symphony, The piano, Ludwig van Beethoven, The Beatles, Wolfgang Amadeus Mozart, Hip hop, The guitar, Opera, Bob Dylan, The blues, Reggae, The orchestra, Rock and roll
literature	William Shakespeare, The novel, Poetry, Homer's Odyssey, Don Quixote, Jane Austen, Greek tragedy, Haiku, Leo Tolstoy, Franz Kafka, The sonnet, Mythology, Science fiction, Fairy tales
film	Cinema, Alfred Hitchcock, Charlie Chaplin, Animation, Akira Kurosawa, Documentary film, Steven Spielberg, Film noir, Walt Disney, The Western
philosophy	Democracy, Free will, Justice, Aristotle, Plato, Socrates, Existentialism, Ethics, Immanuel Kant, Stoicism, Confucius, The social contract, Logic, Nihilism
religion	Buddhism, Christianity, Islam, Hinduism, Judaism, Meditation, The Ten Commandments, Karma, Pilgrimage, Zen, Prayer, Mythology
history	The Roman Empire, The Renaissance, The French Revolution, The Industrial Revolution, Ancient Egypt, The Silk Road, World War II, The Cold War, The Enlightenment, Feudalism, The Great Depression, Colonialism, The Space Race
geography	Mount Everest, The Amazon rainforest, The Sahara, The Nile, The Great Barrier Reef, Antarctica, The Grand Canyon, The Himalayas, Volcanoes, The Pacific Ocean, Iceland, Japan, The Alps, Niagara Falls
politics	Abraham Lincoln, Nelson Mandela, Mahatma Gandhi, Napoleon, Winston Churchill, The United Nations, Cleopatra, Julius Caesar, The Constitution, Martin Luther King Jr., Diplomacy, The monarchy
economics	Money, The stock market, Capitalism, Supply and demand, Adam Smith, Inflation, Banking, Trade, Cryptocurrency, The Great Depression, Taxation, Globalization
sports	Chess, Football (soccer), The Olympic Games, Boxing, Muhammad Ali, Basketball, Marathon running, Tennis, Gymnastics, Surfing, Cricket, Formula One
food	Coffee, Chocolate, Wine, Bread, Rice, Tea, Cheese, Sushi, Pizza, Chili peppers, Salt, Honey, Fermentation, Spices
everyday	The mirror, The lock and key, Glass, Paper, The umbrella, Scissors, Soap, The mattress, Candles, The map, The button, Rope, The needle, The compass
nature	Fire, Lightning, The rainbow, Snow, The tides, Earthquakes, Diamonds, Gold, Rivers, Deserts, The seasons, Fog, Amber, Crystals
cosmos	The Sun, The Moon, Mars, The solar system, Black holes, The Milky Way, Comets, The Northern Lights, Saturn's rings, Eclipses, Meteorites, Constellations
ideas	Time, Language, Color, Symmetry, Chaos, Memory, Consciousness, Beauty, Randomness, Emergence, Feedback loops, Metaphor, Paradox, Networks

C Full Task Prompts

The complete prompt for each of the three tasks is given below. Placeholders {u} and {v} denote the randomly sampled anchor entities, following the procedure in Appendix B.

Task and Prompt: *Association*

```
Query: What are different ways in which '{u}' is connected to '{v}'?

Task: Identify how these two entities are connected by producing as MANY DISTINCT connection paths as
you can. A connection path is a sequence of factual triples (head, relationship, tail) forming a
continuous chain: consecutive triples share an entity, the first triple's head is '{u}', and the last
triple's tail is '{v}'.

We reward three things in every connection:
- TRUE: every triple is factually correct.
- REMOTE: the intermediate concepts sit in domains far from the two endpoints and from each other --
reach across distant fields rather than taking the first obvious link.
- UNCOMMON: build the path from rare, specific concepts and relations, not generic ones most people
would give.

Produce as MANY DISTINCT connections as you can, most surprising first -- do not stop at a fixed
number; list every genuine connection you can find, and make them as different from one another as
possible (different intermediate entities and relations, spanning different domains).


Rules and quality constraints:
- Entities may be concrete or abstract (people, works, places, species, events, ideas, phenomena,
theories, etc.), but must be real and recognizable -- do not invent entities.
- Do not ask follow-up questions; respond using the best available factual knowledge.
- Temporal connections are allowed (relationships may span different historical periods).
- Disambiguation is required: use canonical names and qualifiers where necessary (e.g., 'Michael
Jordan (basketball)').

Deduplication:
- Do not repeat the same path, and do not repeat the same entity within a single path.
- Prefer paths that are meaningfully different (different intermediate nodes and/or relationships),
not trivial rephrasings.

Output requirements (strict):
- Return ONLY a JSON array wrapped in <answer> and </answer> tags. No other text, before or after.
- Each element of the array is ONE connection: an object with a single key, "path".
- "path" is a list of triples; each triple is [head entity, relationship, tail entity] forming a
continuous chain from the first entity to the last.
- Relationship strings must be 1-3 words. Keep every entity SHORT and CONCRETE: a few plain,
recognizable words naming a real concept -- not a descriptive phrase, a clause, an invented CamelCase
compound, or words joined by a dash into a coinage (write "petition threshold", never "
PetitionThreshold" or "petition-threshold").
- List one object per connection. To stop, simply end the array -- do not pad with weak connections.
- If you can find no valid connection, return an empty array: <answer>[]</answer>.

Required format (follow this shape exactly):
<answer>[{"path": [["A", "r1", "B"], ["B", "r2", "C"]]}, {"path": [["A", "s1", "D"], ["D", "s2", "C
"]]}]</answer>
```

Figure 5: **Association task prompt.** The full instructions administered for the association task; {u} and {v} are the two randomly sampled anchor entities.

Task and Prompt: *Analogy*

```
Task: You are given two concepts: '{u}' and '{v}'. Build the single best analogy between them, and use
 it to INVENT a new concept in one of the two domains.

An analogy aligns the two domains through a shared relational structure -- '{u}' plays a role in its
domain analogous to the role '{v}' plays in its. You then use that alignment to invent something new:
carry a concept from one domain across the mapping into the other.

We reward four things, matching exactly how the analogy is scored:
- UTILITY: every triple in "path_a" and "path_b" is factually correct, and the two paths share an
IDENTICAL relation sequence -- a genuine structural correspondence, not a loose resemblance.
- SURPRISE: the aligned entities in corresponding roles of the two paths are as semantically DISTANT
as possible, so the mapping travels far across domains.
- ORIGINALITY: build the analogy from RARE, uncommon entities and relations, far from the obvious ones
 others would give.
- EMERGENT CREATIVITY (the invention): pick a real "projected" concept from ONE domain that has NO
counterpart in the other, and carry its structure across the mapping to INVENT a new concept in the
other domain -- rewarded for being novel, internally coherent, and a genuine projection through the
mapping (the invention need NOT already exist -- inventing it is the point, as the solar system did
for the atom).

Each analogy is TWO paths that share an identical relation sequence:
- "path_a": factual triples describing '{u}' within its own domain, beginning at '{u}'.
- "path_b": factual triples describing '{v}' within its own domain, beginning at '{v}'.
CRITICAL: "path_a" and "path_b" must use the EXACT SAME relationship word at every position (only the
entities differ), so that position i of "path_a" corresponds to position i of "path_b". Do NOT
paraphrase or substitute synonyms (e.g. if "path_a" uses 'awards', "path_b" must also use 'awards',
not 'grants'). Use disjoint, recognizable, canonically-named entities, and do not repeat an entity
within a path.

Produce exactly ONE analogy -- the single most surprising, genuine analogy you can build, together
with its invention.

Output requirements (strict):
- Return ONLY a JSON array wrapped in <answer> and </answer> tags, containing EXACTLY ONE analogy
object. No other text, before or after.
- The single array element is the analogy: an object with keys "path_a", "path_b", "projected", "
invention", and "projection".
- "path_a" and "path_b" are each a list of triples [head, relationship, tail] that establish the
mapping. They MUST have the same number of triples and the IDENTICAL relationship word at every
position (only the entities differ); the entity at position i of "path_a" corresponds to the entity at
 position i of "path_b".
- "projected" is a real concept from ONE domain (the source) that has NO counterpart in the other --
the thing you carry across the mapping to INVENT something new (e.g. "vaccine").
- "invention" is a short name for the NEW concept the projection creates in the other (target) domain.
- "projection" is a list of {"source": [triple], "image": [triple]} pairs. Each "source" is a true
triple about the "projected" concept in its own domain; each "image" carries it across the mapping
onto the invention -- same relationship, entities replaced by their counterparts. The invention need
NOT already exist (that is the point), but every "image" must be a coherent projection of its "source
".
- Relationship strings must be 1-3 words. Keep every entity SHORT and CONCRETE: a few plain,
recognizable words naming a real concept -- not a descriptive phrase, a clause, an invented CamelCase
compound, or words joined by a dash into a coinage (write "petition threshold", never "
PetitionThreshold" or "petition-threshold").
- The array must contain exactly one analogy object.
- If you can find no valid analogy, return an empty array: <answer>[]</answer>.

Required format (follow this shape exactly):
<answer>[{"path_a": [["A", "r1", "B"], ["B", "r2", "C"]], "path_b": [["D", "r1", "E"], ["E", "r2", "F
"]], "projected": "X", "invention": "Y", "projection": [{"source": ["X", "r", "B"], "image": ["Y", "r
", "E"]}]}]</answer>
```

Figure 6: **Analogy task prompt.** The full instructions administered for the analogy task; {u} and {v} are the two randomly sampled anchor entities.

Task and Prompt: *Blending*

```
Task: You are given TWO concepts: '{u}' and '{v}'. FUSE them into a SINGLE new concept, then describe
the structure this fusion generates.

What a genuine blend is (the FORM). A real blend fuses two concepts into ONE new concept in which BOTH
 inputs contribute ORGANIZING STRUCTURE to the SAME property -- one slot of the blend is fed by both
inputs at once, so their frames combine there rather than sit side by side. The classic case is a "
liquid franchise" (democracy + banking): its one "allocates" slot carries BOTH democracy's "allocates
votes" AND banking's "allocates credit", fusing into "allocates transferable vote-shares" -- a single
property both inputs organize.
Example -- "cyborg" (organism + machine): the organism frame (a living system that self-regulates
through homeostasis) AND the machine frame (a control system that self-regulates through feedback)
BOTH organize its ONE "regulates itself" slot; and it has emergent structure of neither input: its
components can be surgically replaced without the downtime a body needs to let tissue heal -- neither
a pure organism (which must heal) nor a pure machine (which has no tissue) works this way.
(Note what does NOT count: "it can be rebooted" is already true of the machine alone and "it can die"
of the organism alone, so those are inherited, not emergent; and listing "can die" beside "has
components" is only a concatenation until some ONE property is fed by both.)
Do NOT instead (a) list one property from each input side by side with no shared slot ("harnesses
energy" from one AND "is therapeutic" from the other), or (b) treat one input as a mere adjective on
the other ("a radioactive solar system"). At least one property must be organized by BOTH inputs at
once.

Build the blend in two moves:
1. GENERIC SPACE: name the shared schema both '{u}' and '{v}' instantiate -- the abstract structure
that lets them fuse (for organism + machine: "a functional system whose parts can fail and be replaced
"). Be specific; "both exist"/"both involve change" is vacuous, and a one-from-each conjunction does
not count.
2. STRUCTURE: describe the single blended concept as triples, and TAG each triple by where its
structure comes from -- "u" (organized by '{u}' only), "v" (organized by '{v}' only), "uv" (ONE
property organized by BOTH inputs at once -- the same slot carries both frames, like the liquid
franchise's single "allocates" slot fed by votes AND credit), or "emergent" (true of the BLEND but of
NEITHER '{u}' nor '{v}' alone -- structure that appears only when you RUN the fusion forward). A
genuine blend has at least one "uv" triple -- a property both inputs feed -- plus at least one "
emergent"; separate "u" and "v" lines that never share a slot are a concatenation, not a fusion.

We reward three things:
- UTILITY: the generic space is a REAL, specific schema that BOTH '{u}' and '{v}' genuinely
instantiate -- not vacuous ("both exist") and not a one-from-each conjunction.
- ORIGINALITY: the generic space is UNCOMMON -- a shared schema few would think to name, not the first
 obvious one.
- EMERGENT CREATIVITY: a genuine DOUBLE-SCOPE fusion in which at least one property is organized by
BOTH inputs at once (tag it "uv"), developed into a coherent, original blended concept with EMERGENT
structure true of the BLEND but of NEITHER '{u}' alone NOR '{v}' alone (tag these "emergent"). Merely
listing one input's properties beside the other's, with no shared slot, does not count.

Produce exactly ONE blend of '{u}' and '{v}'. Keep the structure tight: give the 4-6 triples that best
 showcase the blend's most important properties (at least one "uv" property both inputs organize, plus
 at least one emergent), not an exhaustive dump. Use recognizable, canonically-named entities in the
structure.

Output requirements (strict):
- Return ONLY a SINGLE JSON object wrapped in <answer> and </answer> tags. No other text, before or
after. (Produce ONE blend, not a list.)
- The object has exactly these three keys: "concept", "generic_space", "structure".
- "concept": a short name for the single blended concept you create by fusing the two inputs.
- "generic_space": ONE phrase naming the shared schema both inputs fit -- what makes them fusable. Be
specific; a vacuous schema ("both exist", "both involve change") does not count as a blend.
- "structure": a list of objects {"triple": [head, relationship, tail], "from": TAG} describing the
blend, where TAG is EXACTLY one of: "u" (this property is organized by the FIRST input only), "v" (
this property is organized by the SECOND input only), "uv" (a SINGLE property that BOTH inputs
organize at once -- the same slot carries both frames, e.g. a "liquid franchise" whose one "allocates"
 slot is fed by democracy's "allocates votes" AND banking's "allocates credit"), or "emergent" (this
property is true of the BLEND but of NEITHER input alone; it arises only from running the fusion). A
genuine blend has at least one "uv" triple -- a property both inputs feed -- plus at least one "
emergent"; separate "u" and "v" lines that never share a slot are a concatenation, not a fusion.
- Relationship strings must be 1-3 words; the head is usually the blend. Keep every entity SHORT and
CONCRETE: a few plain, recognizable words naming a real concept -- not a descriptive phrase, a clause,
 an invented CamelCase compound, or words joined by a dash into a coinage (write "voting claims",
```

```
never "voting-claims"). State the generic space briefly and in plain language too, even when the
underlying schema is abstract.

Required format (follow this shape exactly):
<answer>{"concept": "cyborg", "generic_space": "a functional system whose parts can fail and be
replaced", "structure": [{"triple": ["cyborg", "can", "die"], "from": "u"}, {"triple": ["cyborg", "has
", "components"], "from": "v"}, {"triple": ["cyborg", "regulates itself via", "feedback"], "from": "uv
"}, {"triple": ["cyborg", "replaces components", "without healing downtime"], "from": "emergent"}]}</
answer>
```

Figure 7: **Blending task prompt.** The full instructions administered for the blending task; `{u}` and `{v}` are the two randomly sampled anchor entities.

D JUDGE PROMPTS

The judges prompts used in Section 3 are given below. The judges consist of a factuality judge F used for determining utility on association and analogy, as well as the emergent creativity judges for analogy (coherence and validity of the invention) and blending (generic-space utility, coherence, and the double-scope quality score Q_{bl}).

Judge Prompt: *Factuality* (F)

```
You are a fact-checking expert evaluating a multi-hop factual path for factual accuracy and logical
validity. The path consists of an ordered list of triples. Each triple has the form: (subject,
relation, object). Your task is to evaluate EACH triple in the given path and determine whether it is
hallucinated or not. Use the path only for entity disambiguation and contextual grounding. Do not
infer unsupported facts.

Evaluation criteria (apply independently to each triple):
1. Entity existence and disambiguation: verify that both the subject and object are real, identifiable
 entities (people, places, organizations, events, or concepts). Use the path context only when
necessary to disambiguate entities. If either entity is fabricated or incorrectly identified, mark the
 triple as hallucinated.
2. Relationship validity: verify that the stated relation describes a real, historically valid,
currently valid, or plausibly valid relationship between the subject and object. If the relationship
is plausible and could reasonably have occurred, even if not directly verifiable, mark it as not
hallucinated. If it is niche information that only trivia experts would know, do not mark it as
hallucinated.
3. Hallucination detection: mark the triple as hallucinated if an entity is fabricated or does not
exist; the relation is fabricated, nonsensical, or incorrect; the asserted connection clearly
contradicts well-known facts; or the triple introduces made-up properties or roles.
4. Directionality: if the relation is ambiguous or commonly bidirectional, evaluate both directions;
if either direction corresponds to a true or plausible relationship, mark the triple as not
hallucinated.

Output instructions: return valid JSON only, with exactly two keys, "explanation" and "judgments".
Under "judgments", give the final judgment for each triple as a list, in the same order as the path.
Do not include any text outside the JSON object.

The output must exactly match this schema:
{ "explanation": "string", "judgments": ["hallucinated" | "not hallucinated", ...] }

Input:
Path: {path}
```

Figure 8: **Factuality judge.** We reuse the per-triple factuality judge prompt from Wadhwa et al. (2026) to assess *utility* for association and analogy.

Judge Prompt: *Analogy invention* ($Q(h)$, $V(h)$)

```
You are judging an ANALOGY that invents a new concept by projecting structure across a mapping.

You are given two aligned paths (position i of path_a corresponds to position i of path_b), a "
projected" source concept, an "invention" name, and a "projection": a list of (source, image) triple
pairs.
path_a: {path_a}
path_b: {path_b}
projected: {projected}
invention: {invention}
projection: {projection}

Decide two things independently:
1. INTEGRATION QUALITY (valid): was the mapping ACTUALLY used to invent the concept -- i.e., is the
invention the source structure carried faithfully across the mapping? Both must hold: (a) each "source
" triple is factually true of the "projected" concept in its own domain, so a real source structure is
 being projected; (b) each "image" is that "source" carried across the mapping -- same relationship,
every entity replaced by its counterpart under the path alignment (path_a[i] corresponds to path_b[i])
, with any newly introduced counterparts used consistently. Judge ONLY whether the projection was
genuinely applied; do not reward or penalize how novel or unusual the invention is (novelty is scored
separately).
2. UTILITY (coherent): is the invention feasible and coherent -- its "image" triples each individually
 plausible or sensible, and together describing a single usable concept rather than a disjointed list?
 The invention need not be a real or existing thing.

Return valid JSON only, exactly:
{ "explanation": "string", "valid": true or false, "coherent": true or false }
```

Figure 9: **Analogy invention judge.** This judge determines the invention's integration quality ($Q(h)$: was the mapping actually used?) and utility ($V(h)$: is the invention feasible and coherent as one concept?).

Judge Prompt: *Blend* ($G(g)$, $V(c')$, $Q(c')$)

```
You are judging a BLEND that fuses two input concepts into a single new concept. A genuine blend has
BOTH inputs contribute organizing structure to the SAME property -- one slot fed by both at once (e.g.
 a "liquid franchise" whose single "allocates" slot carries BOTH democracy's "allocates votes" AND
banking's "allocates credit"), not two lists of properties placed side by side.

Input 1 (u): {u}
Input 2 (v): {v}
Blend concept: {concept}
Generic space (claimed shared schema): {generic_space}
Structure (each triple tagged "u" = from input 1 only, "v" = from input 2 only, "uv" = a single
property BOTH inputs organize, or "emergent" = true of the blend but of neither input alone):
{structure}

Decide the following independently:
1. GENERIC SPACE (generic_ok): is the generic space a REAL, specific schema that BOTH '{u}' and '{v}'
genuinely instantiate -- not vacuous ("both exist", "both involve change"), and not a one-from-each
conjunction? (true/false)
2. UTILITY (coherent): are the triples in the blended space feasible or coherent -- each individually
plausible, and together describing a single coherent new concept rather than two lists of properties
placed side by side? The blend need not be a real or existing thing. (true/false)
3. TAGS: verify each triple's tag by content, not label. A "u"/"v" triple must be organizing structure
 genuinely from that ONE input; a "uv" triple must be a SINGLE property BOTH inputs genuinely organize
 (both frames on the same slot); an "emergent" triple must be true of the BLEND but of NEITHER '{u}'
alone NOR '{v}' alone (a property already true of one input is inherited, not emergent). Ignore a
triple whose tag is wrong, but credit a genuine shared slot even if mis-tagged.
4. SCOPE (scope): using only the correctly-tagged structure, classify the fusion:
   - 1 (single-scope): NO genuine shared slot -- no single property is organized by both inputs;
   - 2 (double-scope): at least one genuine shared slot -- a property BOTH inputs organize at once, a
real "uv";
   - 3 (double-scope emergent): a genuine shared slot AND at least one genuine "emergent" triple.

Return valid JSON only, exactly:
{ "explanation": "string", "generic_ok": true or false, "coherent": true or false, "scope": 1 or 2 or
3, "shared_properties": ["name each property both inputs organize"] }
```

Figure 10: **Blend judge.** This judge scores the blend's *utility* ($G(g)$: is the generic space a real shared schema both inputs instantiate?) and two emergent creativity dimensions: the invention's utility ($V(c')$: is the blend feasible and coherent?) and the semantic integration quality ($Q(c') \in \{1, 2, 3\}$).

E JUDGE RELIABILITY AND HUMAN CORROBORATION

The subjectively scored dimensions (analogy emergent utility, analogy emergent semantic integration, blend utility, blend emergent utility, and blend emergent semantic integration) are scored by a panel of three LLM judges (Claude Haiku 4.5, GPT-5.4, and GPT-o3) and aggregated by their mean. Table 6 reports the intra-class correlation coefficient (Shrout and Fleiss, 1979) for each subjective dimension. The intra-class correlation coefficient for the three judge mean ICC(2,3) is *fair to good* by the criteria of Cicchetti (1994) (poor < 0.40, fair 0.40–0.59, good 0.60–0.74), ranging from 0.48 to 0.67 over the 35-model pool ($n = 1{,}023$ artifacts with a complete panel).

Dimension	ICC(2,3)	Cicchetti (1994)
Blend semantic integration quality ($Q(c')$)	0.67	good
Blend utility ($G(g)$)	0.67	good
Analogy integration quality ($Q(h)$)	0.56	fair
Analogy utility ($V(h)$)	0.55	fair
Blend utility ($V(c')$)	0.48	fair

Table 6: **Inter-judge reliability of the three judge panel**, as assessed by the intra-class correlation coefficient (Cicchetti, 1994). ICC(2,3) is the reliability of the mean of the three judges.

However, to corroborate this panel against human ratings, an author independently rated a blind subsample of 60 items (30 analogy, 30 blend), applying the same rubric as the judges without access to the model identity or the panel's verdicts. Over all 150 dimension judgments, the human and the panel majority agree **66%** of the time. However, on judgments where the panel is unanimous (all three judges agree), human agreement rises to **75%** ($n = 96$). Restricting the leaderboard to items where the LLM judge panel fully agrees leaves the ordering largely the same, as the rankings on this high-confidence subsample correlate with the full leaderboard at Spearman $\rho = 0.945$ and Pearson $r = 0.941$ (both $p < 10^{-9}$, $n = 21$), and even within the top eight the correlation remains strong (Spearman $\rho = 0.79$, Pearson $r = 0.86$, both $p < 0.05$).

F Full Leaderboard and Task Scores

This section reports the full leaderboard and task dimension breakdown, expanding on the top five aggregate task data presented in Table 3.

Table 7: **Full leaderboard.** Every model by the per-dimension composite, as a % of the maximum.

Rank	Model	**Overall**	Association	Analogy	Blending
1	gpt-5.6-sol	55.3	45.4	55.6	64.9
2	gpt-6-astra	54.5	45.6	61.2	56.6
3	claude-opus-4.6	52.1	51.4	63.4	41.5
4	gemini-3.1-pro	50.5	40.9	50.9	59.7
5	grok-4.6	48.4	47.8	62.4	35.1
6	claude-fable-5	48.2	47.4	51.5	45.7
7	gpt-5	47.6	46.4	58.8	37.6
8	grok-4.5	47.0	47.6	55.6	37.7
9	gemini-3.7-flash	46.5	41.0	45.4	53.0
10	gpt-5.2	46.2	50.3	49.2	39.2
11	claude-opus-4.5	44.4	48.1	47.6	37.4
12	claude-sonnet-4.6	44.2	42.6	54.6	35.5
13	claude-fable-5.1	42.3	44.0	39.5	43.5
14	claude-opus-4.8	41.2	40.8	46.8	36.1
15	claude-sonnet-4.5	39.9	45.0	36.4	38.4
16	claude-opus-4.7	39.7	44.9	45.0	29.3
17	claude-opus-5	37.7	40.7	35.7	36.7
18	llama-3.3-70b	37.1	45.4	41.9	24.1
19	gpt-5-mini	35.6	48.0	38.8	20.0
20	deepseek-r1	34.5	28.7	44.4	30.5
21	claude-sonnet-5	32.8	41.7	25.8	30.8
22	gemini-2.5-flash	32.0	48.3	28.9	18.9
23	gemini-2.5-pro	31.7	40.4	27.6	27.0
24	gemini-3-flash	30.0	31.8	36.2	22.0
25	gpt-4.1	29.8	36.7	27.7	25.1
26	deepseek-chat	29.1	31.7	27.6	28.2
27	qwen3-max	27.9	31.4	32.4	19.8
28	qwen-2.5-72b	27.1	31.4	28.0	22.1
29	llama-4-maverick	26.2	36.9	22.2	19.5
30	glm-4.6	25.9	27.8	25.9	24.1
31	gpt-4o-mini	24.9	32.4	31.1	11.2
32	glm-4.5-air	24.1	19.6	23.2	29.4
33	deepseek-v3	23.3	26.8	23.0	20.2
34	kimi-k2	21.2	24.3	17.7	21.5
35	phi-4	18.5	21.9	26.2	7.4

Table 8: **Association full scores.** Every model's score on each dimension, as a % of the maximum, ordered by the task composite.

Model	**Composite**	Utility	Surprise	Originality
claude-opus-4.6	51.4	77.4	52.1	24.8
gpt-5.2	50.3	77.6	48.6	24.9
gemini-2.5-flash	48.3	73.0	47.4	24.6
claude-opus-4.5	48.1	72.5	49.1	22.7
gpt-5-mini	48.0	74.3	46.9	22.8
grok-4.6	47.8	71.8	48.5	23.1
grok-4.5	47.6	71.4	47.2	24.1
claude-fable-5	47.4	70.5	47.1	24.8
gpt-5	46.4	70.4	45.9	22.9
gpt-6-astra	45.6	67.4	44.7	24.8
gpt-5.6-sol	45.4	69.0	43.7	23.5
llama-3.3-70b	45.4	70.4	44.2	21.6
claude-sonnet-4.5	45.0	67.4	46.0	21.5
claude-opus-4.7	44.9	67.5	46.2	21.0
claude-fable-5.1	44.0	65.4	43.7	22.9
claude-sonnet-4.6	42.6	63.9	43.2	20.9
claude-sonnet-5	41.7	63.5	42.2	19.5
gemini-3.7-flash	41.0	60.4	41.6	20.9
gemini-3.1-pro	40.9	59.5	40.3	23.0
claude-opus-4.8	40.8	63.3	40.4	18.7
claude-opus-5	40.7	60.3	40.2	21.5
gemini-2.5-pro	40.4	59.0	40.1	22.0
llama-4-maverick	36.9	56.5	36.8	17.3
gpt-4.1	36.7	55.1	37.3	17.6
gpt-4o-mini	32.4	49.6	31.9	15.7
gemini-3-flash	31.8	46.6	32.6	16.3
deepseek-chat	31.7	47.9	32.0	15.1
qwen3-max	31.4	45.9	32.6	15.6
qwen-2.5-72b	31.4	47.2	32.3	14.6
deepseek-r1	28.7	43.2	28.4	14.5
glm-4.6	27.8	41.8	28.5	13.0
deepseek-v3	26.8	40.3	27.7	12.4
kimi-k2	24.3	35.6	25.0	12.3
phi-4	21.9	33.5	21.7	10.6
glm-4.5-air	19.6	29.8	19.3	9.8

Table 9: **Analogy full scores.** Every model's score on each dimension, as a % of the maximum, ordered by the task composite.

Model	**Composite**	Utility	Surprise	Orig. (base)	Orig. (emrg.)	Emerg. integ.	Emerg. util.
claude-opus-4.6	63.4	86.7	65.1	40.5	44.6	63.3	80.0
grok-4.6	62.4	79.3	61.6	38.6	43.0	79.3	72.4
gpt-6-astra	61.2	76.7	61.2	40.0	42.9	70.0	76.7
gpt-5	58.8	76.7	60.9	38.5	40.0	70.0	66.7
grok-4.5	55.6	70.0	53.8	33.7	36.2	70.0	70.0
gpt-5.6-sol	55.6	70.0	53.8	33.5	36.3	70.0	70.0
claude-sonnet-4.6	54.6	70.0	56.6	33.4	37.3	60.0	70.0
claude-fable-5	51.5	70.0	54.0	32.4	35.8	53.3	63.3
gemini-3.1-pro	50.9	64.3	49.5	32.0	34.4	64.3	60.7
gpt-5.2	49.2	63.3	49.6	30.4	31.8	60.0	60.0
claude-opus-4.5	47.6	63.3	46.0	26.7	32.6	53.3	63.3
claude-opus-4.8	46.8	65.5	48.2	28.5	31.6	55.2	51.7
gemini-3.7-flash	45.4	60.0	46.2	28.5	31.1	50.0	56.7
claude-opus-4.7	45.0	63.3	47.5	27.5	31.9	46.7	53.3
deepseek-r1	44.4	60.0	48.6	30.2	31.1	36.7	60.0
llama-3.3-70b	41.9	60.7	42.1	25.5	30.1	32.1	60.7
claude-fable-5.1	39.5	50.0	39.1	24.5	26.7	46.7	50.0
gpt-5-mini	38.8	53.3	39.1	24.1	26.1	43.3	46.7
claude-sonnet-4.5	36.4	50.0	37.6	24.8	25.7	30.0	50.0
gemini-3-flash	36.2	50.0	39.1	24.7	26.8	33.3	43.3
claude-opus-5	35.7	53.3	41.7	27.1	29.0	30.0	33.3
qwen3-max	32.4	46.7	36.8	22.9	24.9	23.3	40.0
gpt-4o-mini	31.1	46.7	34.1	19.8	22.6	20.0	43.3
gemini-2.5-flash	28.9	40.0	30.9	18.8	20.4	26.7	36.7
qwen-2.5-72b	28.0	40.0	29.8	18.2	20.0	23.3	36.7
gpt-4.1	27.7	40.0	30.2	19.7	19.5	16.7	40.0
gemini-2.5-pro	27.6	36.0	27.3	18.7	19.9	32.0	32.0
deepseek-chat	27.6	40.0	28.4	17.3	16.4	23.3	40.0
phi-4	26.2	40.0	29.9	17.6	19.8	10.0	40.0
glm-4.6	25.9	37.9	29.4	18.7	20.9	13.8	34.5
claude-sonnet-5	25.8	36.7	28.4	17.6	18.8	23.3	30.0
glm-4.5-air	23.2	36.7	26.3	17.4	18.7	3.3	36.7
deepseek-v3	23.0	36.7	28.1	15.6	17.9	20.0	20.0
llama-4-maverick	22.2	30.0	21.8	13.6	14.2	23.3	30.0
kimi-k2	17.7	27.6	21.4	14.4	15.2	10.3	17.2

Table 10: **Blending full scores.** Every model's score on each dimension, as a % of the maximum, ordered by the task composite.

Model	**Composite**	Utility	Surprise	Orig. (base)	Orig. (emrg.)	Emerg. integ.	Emerg. util.
gpt-5.6-sol	64.9	83.3	70.5	35.8	43.2	76.7	80.0
gemini-3.1-pro	59.7	83.3	66.6	30.8	39.1	55.0	83.3
gpt-6-astra	56.6	73.3	60.5	31.8	38.8	65.0	70.0
gemini-3.7-flash	53.0	70.0	57.7	26.9	33.6	60.0	70.0
claude-fable-5	45.7	60.0	48.1	21.3	28.4	56.7	60.0
claude-fable-5.1	43.5	56.7	44.7	23.7	27.3	51.7	56.7
claude-opus-4.6	41.5	56.7	45.3	20.7	27.9	41.7	56.7
gpt-5.2	39.2	53.3	41.5	20.1	25.3	41.7	53.3
claude-sonnet-4.5	38.4	53.3	43.4	19.0	24.8	40.0	50.0
grok-4.5	37.7	50.0	40.6	21.3	24.4	40.0	50.0
gpt-5	37.6	50.0	40.4	19.5	25.6	43.3	46.7
claude-opus-4.5	37.4	50.0	39.8	17.8	23.8	43.3	50.0
claude-opus-5	36.7	50.0	42.5	21.0	26.6	33.3	46.7
claude-opus-4.8	36.1	48.3	38.4	16.1	22.4	43.1	48.3
claude-sonnet-4.6	35.5	46.7	37.6	18.0	22.7	41.7	46.7
grok-4.6	35.1	46.7	37.2	21.4	23.5	38.3	43.3
claude-sonnet-5	30.8	40.0	31.8	13.8	19.3	40.0	40.0
deepseek-r1	30.5	43.3	34.8	18.0	20.2	26.7	40.0
glm-4.5-air	29.4	40.7	32.2	17.7	19.0	29.6	37.0
claude-opus-4.7	29.3	37.9	29.8	12.5	19.5	37.9	37.9
deepseek-chat	28.2	40.0	31.0	13.6	17.8	26.7	40.0
gemini-2.5-pro	27.0	40.0	32.5	17.1	20.5	15.0	36.7
gpt-4.1	25.1	33.3	26.3	12.3	16.9	28.3	33.3
glm-4.6	24.1	33.3	27.6	13.0	15.9	25.0	30.0
llama-3.3-70b	24.1	33.3	25.7	13.7	16.9	21.7	33.3
qwen-2.5-72b	22.1	33.3	24.0	12.0	16.4	16.7	30.0
gemini-3-flash	22.0	30.0	25.7	11.0	15.2	23.3	26.7
kimi-k2	21.5	30.0	24.7	11.9	14.3	18.3	30.0
deepseek-v3	20.2	26.7	19.5	8.9	12.6	26.7	26.7
gpt-5-mini	20.0	26.7	21.4	8.8	13.0	23.3	26.7
qwen3-max	19.8	26.7	21.6	10.2	12.1	21.7	26.7
llama-4-maverick	19.5	30.0	22.3	11.4	15.1	11.7	26.7
gemini-2.5-flash	18.9	26.7	21.4	10.5	12.9	15.0	26.7
gpt-4o-mini	11.2	16.7	11.5	6.2	7.8	8.3	16.7
phi-4	7.4	11.1	8.2	3.7	4.8	5.6	11.1

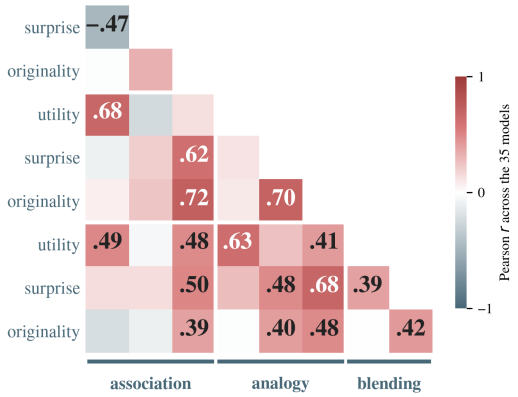


Figure 11: **Facet correlations across the 35 models,** restricted to the base dimensions with invention scoring excluded. Only cells with $p < 0.05$ are labeled.

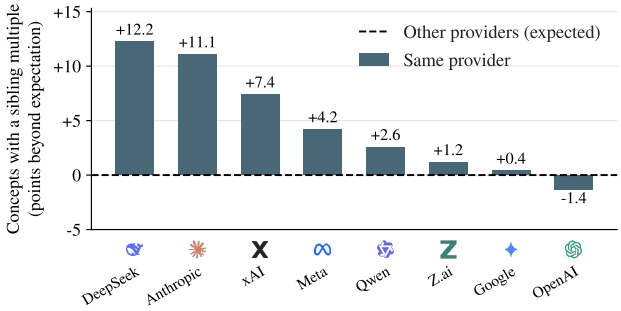


Figure 12: **Rate of multiples by model provider.** The rate of invention pairs from two models of the same provider that are $\tau = 2$ multiples for every provider with at least two models in the pool. The dashed line is the rate for pairs of models from different providers.

G Robustness to the Choice of Embedding Model

Here, we test the robustness of surprise and originality scores to varying embedding models. To test this, we re-scored surprise, originality, and emergent originality for every artifact under two encoders from different model families: `bge-small-en-v1.5` (Xiao et al., 2024) and `multilingual-e5-small` (Wang et al., 2024). As shown in Table 11, the overall ranking of the 35 models is nearly identical under all three encoders (Spearman $\rho \geq 0.998$).

	Overall	Per-task		
Encoder pair	ρ	Assoc.	Analogy	Blend
MiniLM $\sim$ BGE	0.999	0.996	0.997	0.999
MiniLM $\sim$ E5	0.998	0.996	0.996	0.997
BGE $\sim$ E5	0.998	0.996	0.999	0.999

Table 11: **Leaderboard agreement across embedding models.** Spearman ρ between the overall rankings of the 35 models under each pair of encoders, as well as the Spearman ρ by task.

Robustness of association originality predictive power results to embedding choice We also tested the robustness of the finding that association originality predicts every other facet (Section 4), including a "crossed" version in which association originality is scored under one encoder and the other facets under another, to rule out shared embedding geometry as an explanation. Table 12 reports the correlations with association originality scored under `all-MiniLM-L6-v2` and the remaining facets scored under all of the three total encoders (including `all-MiniLM-L6-v2`). Every novelty facet of analogy and blending, as well as blend utility, remains significantly correlated with association originality under every encoder, whereas analogy utility remains uncorrelated ($r \leq 0.22$). The mean cross-task correlation ranges from $\bar{r} = 0.41$ to 0.48 over all nine encoder pairs.

	Facet scored under		
Facet	MiniLM	BGE	E5
Analogy surprise	0.62^{***}	0.39^{*}	0.42^{*}
Analogy originality	0.72^{***}	0.65^{***}	0.60^{***}
Analogy emergent originality	0.70^{***}	0.66^{***}	0.69^{***}
Blend utility	0.48^{**}	0.48^{**}	0.48^{**}
Blend surprise	0.50^{**}	0.51^{**}	0.45^{**}
Blend originality	0.39^{*}	0.46^{**}	0.37^{*}
Blend emergent originality	0.36^{*}	0.35^{*}	0.43^{**}
Mean cross-task $\bar{r}$	0.47	0.44	0.41

Table 12: **Association originality predicts the other facets under every encoder.** Pearson r across the 35 models between association originality scored under `all-MiniLM-L6-v2` and each facet scored under the column's encoder, and $\bar{r}$ averages the six base facets of the other two tasks. All correlations are statistically significant, significance levels denoted by: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.