Evaluating Creative Generative AI with Conceptual Blending

Suzie Xi
University of Notre Dame
wxi@nd.edu

Zhongheng Cheng
Columbia University
zc2737@columbia.edu

Deepak Kumar Soni
Case Western Reserve University
dks91@case.edu

Mark Turner
Case Western Reserve University
mark.turner@case.edu

Abstract

Human conceptualization is inherently creative and dynamic, yet as generative models become embedded in real-world creative workflows, the field still lacks principled frameworks for evaluating creativity. Current approaches reduce creativity to surface-level novelty, prioritizing what is easy to quantify over what is theoretically meaningful, and yielding evaluations that are unanchored, inconsistent, and difficult to reproduce. To address these limitations, this position paper introduces conceptual blending theory from cognitive linguistics, which models creativity as the integration of distinct conceptual spaces, producing emergent meaning. We argue that conceptual blending offers a principled framework for evaluating, understanding, and steering creativity in generative AI. We outline how creative generation can be operationalized through structured blending tasks built on FrameNet-based semantic structures, retrieval-augmented generation pipelines, and reproducible measures. Grounding evaluation in cognitive theory moves creativity assessment beyond ad hoc metrics. By integrating conceptual blending with computational infrastructure, our proposed framework supports an interpretable, principled, and reproducible way of evaluating creativity in generative AI.

1 Introduction

Generative AI is increasingly used in creative tasks such as writing and digital art, but current models remain constrained by their training data distributions, and we lack principled frameworks for evaluating their creative and conceptual capabilities (Bellemare-Pepin et al., 2026; Solaiman et al., 2025). This position paper introduces conceptual blending, a cognitive-linguistic account of creativity as the integration of distinct conceptual spaces producing emergent meaning, as a basis for evaluating generative AI (Turner, 2014; Fauconnier and Turner, 2008, 1998). Our notion of creativity is distinct from temperature tuning, stylistic prompting, or scale-induced emergence. These techniques may increase surface-level novelty, but they do not guarantee cross-domain integration or emergent meaning, and prompt-based approaches often improve output style without clarifying the underlying mechanism. Instead, we define creativity as the generation of outputs that integrate distinct conceptual domains to produce emergent structure or meaning not reducible to either source alone. Conceptual blending supports this view by enabling the systematic extraction and recombination of remote concepts in semantic space, offering a structured approach to evaluating creativity.

2 Our Position

We argue that conceptual blending provides a framework for evaluating creativity in generative AI that is interpretable, theory-grounded, and reproducible.

Beyond its theoretical appeal, our framework offers a principled way to address several persistent challenges in creative generative AI: output homogenization, where models converge on safe, familiar patterns; limited controllable steering, where there are few reliable mechanisms for guiding models toward more inventive output; and the weak theoretical grounding of existing creativity metrics, which too often reduce creativity to surface-level novelty. Where the above definition sets the criterion, conceptual blending makes it concrete, rendering conceptual integration explicit, interpretable, and measurable. To this end, we adopt FrameNet-based conceptual blending as a practical substrate: input frames provide structured semantic material, generated blends expose how models combine conceptual spaces, and rubric-based evaluation makes integration quality measurable and reproducible (Turner, 2014; Fauconnier and Turner, 2008, 1998).

3 Conceptual Blending Theory

The term *conceptual blending* is used to describe the general cognitive operation, while frame blending refers to a specific instance in which conceptual blending occurs between structured semantic frames. It opens new pathways for modeling divergent associative thinking, improving the explainability of the model, and designing experiments based on the FrameNet lexical resource (Baker et al., 1998; Ruppenhofer et al., 2016). We argue that creative generative AI systems that aim to produce cross-domain, meaning-constructive outputs (e.g., metaphor, narrative, theory borrowing) should be explicitly guided and evaluated using conceptual blending frameworks, rather than relying solely on scale, stochasticity, or surface-level prompting. Conceptual blending demonstrates a domain-general cognitive process, meaning that it relies on learning and inference mechanisms that operate broadly across content domains rather than being specialized for any single type of knowledge (Sloutsky, 2010). Its intellectual roots can be traced to Koestler's *The Act of Creation*, which identified a shared structural pattern underlying creative achievements in the arts, sciences, and humor (Koestler, 1964). Koestler termed this process the "bisociation of matrices," describing creativity as the integration of elements from previously unrelated matrices of thought into a new matrix of meaning through comparison, abstraction, categorization, analogy, and metaphor (Koestler, 1964). The idea of blending in relation to frame semantics was first formally introduced in the theorization of conceptual blending, which defines conceptual integration as a fundamental cognitive operation on par with analogy, recursion, mental modeling, conceptual categorization, and framing (Turner, 2014; Fauconnier and Turner, 1998). Blending is considered a basic mental process that generates new meanings, global insights, and conceptual compressions that make complex information easier to store and manipulate (Fauconnier and Turner, 2003).

The theory of conceptual blending, further developed in explorations of the "human spark" of creativity, views the creation of novel ideas as an active process of integrating conceptual structures from different mental spaces to produce an emergent "blended" space with its own unique properties (Turner, 2014; Fauconnier and Turner, 1998). Conceptual blending applies the network model of conceptual integration, including Input Space 1, Input Space 2, a Generic Space, and a Blended Space, to explain how shared organizing frames structure mappings across domains, as shown in Figure 1 (Fauconnier and Turner, 2003). Two input spaces are linked by a cross-space mapping and selectively projected onto a blended space. It assumes, as a core finding of cognitive science, that structure mapping and metaphorical projection are central to constructing reasoning and meaning. By aligning the topology of all spaces through a common frame, conceptual blending provides a systematic account of how complex conceptual relations are integrated into coherent mental models (Fauconnier and Turner, 2003).

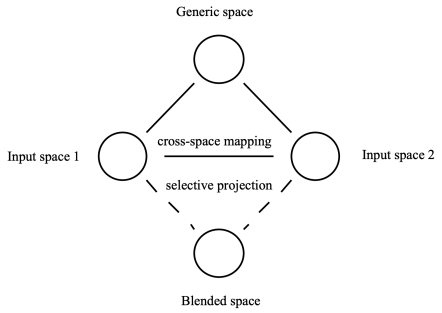

Figure 1: Four-space model of conceptual blending

The integration of computational methods with conceptual blending theory has evolved significantly since the theory was first introduced, exemplifying the convergence of explanatory and predictive research. Earlier efforts often relied on symbolic representations and structured knowledge models (Veale and O'Donoghue, 2000). Distributional semantics encodes the meaning of words as co-occurrence vectors, with word embeddings that capture the semantic relationships useful for identifying concepts to blend (Pereira et al., 2013). Specialized tools have emerged to support these tasks, such as Lutma for collaborative FrameNet expansion (Torrent et al., 2022). Although sequential reasoning paradigms such as Chain-of-Thought (CoT) improve performance on logical tasks (Wei et al., 2022b; Kojima et al., 2022), creative outputs often benefit from non-sequential, associative thinking, as explored in approaches like Leap-of-Thought (Zhong et al., 2024). Furthermore, techniques such as Retrieval-Augmented Generation (RAG) enable LLMs to incorporate external resources such as the FrameNet lexical resource (Lewis et al., 2020). Ongoing research reflects this evolution, progressing from early semantic metrics to integrated systems using LLMs, RAG, and

modern frameworks such as LangChain and LangGraph to parse FrameNet, generate frame blending examples, and automate blending processes (Xi, 2020; Cheng, 2024; Soni, 2025). Recent work also leverages the FrameNet lexical resource to evaluate models under prompt-based inference, finding that they can perform frame identification without explicit supervision (Chundru et al., 2025). These developments are leading toward models that not only produce creative output, but also align with broader trends in deploying large language models as intelligent agents, particularly in domains that require sophisticated analytical capacity and support the emergence of structured reasoning.

4 The Current State of Creativity in Generative AI

This section reviews the existing literature and highlights gaps that motivate our proposed framework. The early study of computational creativity addresses questions such as what creativity is and whether it can be explained scientifically (Boden, 2004). Early exploration of computational creativity includes self-transforming machines capable of generating novel output by making unfamiliar combinations of familiar ideas. Recently, the emergence of large language models has drawn more interest in evaluating and enhancing their creative capacity, which remains in a nascent state. In general, research in this area can be categorized into three main streams: first, evaluating the cognitive and creative capabilities of LLMs, often in comparison with humans; second, investigating human–AI collaboration in creative tasks; and third, exploring prompt engineering to elicit emergent creative behaviors in generative models.

One stream of the literature evaluates the extent to which generative models exhibit cognitive traits commonly associated with creativity, such as divergent thinking, semantic association, and conceptual flexibility. Prior work has assessed LLM-generated content using computational techniques including sentiment analysis, clustering, and classification (Dinu and Florescu, 2025). Other studies operationalize associative creativity by measuring semantic distance between otherwise unrelated concepts (Chen and Ding, 2023), and explicitly prompting models to connect disparate concepts has been shown to enhance their creative output (Mehrotra et al., 2024). Under specific task designs and prompting conditions, LLMs can outperform average human participants on measures of divergent creative association (Bellemare-Pepin et al., 2026), and broader benchmarks find their individual and collective creativity comparable to humans (Sun et al., 2025). At the same time, observed creative performance in LLMs remains highly sensitive to prompt structure and contextual framing (Zhao et al., 2025). However, much of the existing literature relies on narrowly defined benchmarks that inadequately capture the inherently context-dependent and subjective nature of creativity. Many evaluation approaches remain weakly grounded in established theories from cognitive science, linguistics, and psychology. In contrast, our position paper advances a cognitively and semantically grounded framework for understanding and evaluating creative generation, drawing on a body of work that has been extensively developed and examined across these disciplines (Turner, 2014; Baker et al., 1998; Fauconnier and Turner, 2008, 1998; Fillmore and Atkins, 2012; Veale and O'Donoghue, 2000; Torrent et al., 2024).

The second stream examines human–AI collaboration in creative tasks, focusing on how large language models augment, rather than replace, human creativity. Empirical studies of professional writers suggest that LLMs are particularly effective in supporting activities such as drafting, revision, and the generation of alternative expressions (Chakrabarty et al., 2024). Complementary work uses sentence-embedding-based analyses to evaluate model flexibility across content domains, finding that more adaptable models tend to achieve higher scores on creativity-related measures (Nath et al., 2024). Additionally, LLMs used as sounding boards narrowed the non-expert–expert gap, whereas use as ghostwriters gave no benefit and harmed experts (Chen and Chan, 2024).

The third stream investigates how emergent creative behaviors can be elicited through novel prompting strategies, and how such behaviors should be measured and bounded. One line of work reframes failure modes as opportunities: hallucinations, usually treated as defects, have been reinterpreted as a potential source of creativity (Jiang et al., 2024). Building on the idea that creativity can be drawn out through interaction, others design elaborate prompting protocols incorporating multi-voice narration, style transfer, and guided genre constraints, finding that the sophistication of the output tracks that of the prompt (Shanahan and Clarke, 2023). Alongside these optimistic accounts

Metric	Definition	Score 1	Score 5	Score 10
Completeness	Measures whether the generated output fully realizes the intended combination of the input frames.	One frame only, or key elements missing.	Both frames present, but unevenly developed.	Both frames clearly represented and integrated.
Clarity	Assesses the extent to which the generated text is well-formed, interpretable, and linguistically coherent.	Confusing, ungrammatical, or hard to interpret.	Understandable, but awkward or ambiguous.	Fluent, clear, and concise.
Relevance	Evaluates whether the output preserves meaningful semantic connections to the original input frames.	Loses meaning of one or both frames.	Preserves some relevant semantic elements.	Preserves core roles or relations from both frames.
Understanding Depth	Captures whether the response reflects abstract reasoning and non-trivial conceptual integration rather than superficial fusion.	Superficial juxtaposition only.	Some cross-frame mapping, but limited emergence.	Non-trivial integration with emergent meaning.
Coherence	Examines the logical consistency and integrative quality of the blended elements in the generated output.	Contradictory, inconsistent, or not unified.	Mostly understandable, with some weak links.	Logically consistent and well integrated.
Execution Time	Measures the time required to generate the output, serving as a proxy for system efficiency and scalability.	Slow or requires heavy manual intervention.	Moderate prompting or revision needed.	Efficient, minimal intervention, reproducible.

Table 1: Operational scoring rubric for evaluating conceptual blending in creative generative AI. The score columns operationalize each metric on a 1–10 scale.

run a set of cautionary findings. There is concern that LLMs may homogenize human creativity, narrowing rather than expanding ideation (Anderson et al., 2024). The assumption that creativity can simply be tuned through decoding parameters has likewise been challenged: temperature, often called the "creativity parameter," shows only a weak link to novelty (Peeperkorn et al., 2024). More fundamentally, some emergent behaviors attributed to model scale may be artifacts of evaluation, disappearing under closer inspection or different metrics (Schaeffer et al., 2023), a direct challenge to earlier claims that certain capabilities surface only at larger scales, marking qualitative shifts in model behavior (Wei et al., 2022a). These tensions have motivated efforts to evaluate creative behavior more rigorously. Automatic scoring systems have been developed to assess the originality of generated metaphors (DiStefano et al., 2025), and prompt-based protocols now probe phenomena ranging from divergent associative reasoning under the Leap-of-Thought paradigm (Zhong et al., 2024) to the systematic measurement of hallucination and factual reliability (Sato, 2025a).

To ground our theoretical foundation, we draw on the Four-C Model of Creativity, which distinguishes four levels: *Mini-C*, *Little-C*, *Pro-C*, and *Big-C* (Beghetto and Kaufman, 2007; Kaufman and Beghetto, 2009). We focus on the first two: *Mini-C*, the novel and personally meaningful interpretation of experiences in learning; and *Little-C*, everyday creative expression that, while not groundbreaking, still exhibits originality and value beyond randomness (Ilha Villanova and Pina e Cunha, 2021). Accordingly, our metrics target these two levels: Understanding Depth and Coherence assess whether a blend reflects the meaningful conceptual integration characteristic of Mini-C and Little-C, not the eminent originality of Pro-C or Big-C.

Grounding evaluation matters because current creativity evaluation faces four persistent limitations. First, there is no shared standard, since creativity is assessed inconsistently across models and benchmarks with no common protocol. Second, evaluation tends to emphasize surface novelty, favoring stylistic variation or lexical diversity over conceptual integration, and rarely capturing whether distinct conceptual domains are combined to construct meaning. Third, the theoretical grounding is weak, as existing approaches lack interpretable mechanisms anchored in cognitive or linguistic theory. Finally, there is no unified framework, so without a shared prompting or evaluation protocol, results are difficult to reproduce across models, datasets, and tasks.

Conceptual blending addresses these limitations by modeling creativity as the integration of multiple conceptual spaces, providing a theoretically grounded basis for evaluating generative outputs. To operationalize this view, we introduce a set of multidimensional metrics for assessing the quality of generated blends, which, as shown in Table 1, combine with structured prompts and model outputs to support a systematic and interpretable evaluation of creative generative AI. These metrics mark an initial step; developing shared annotation protocols and assessing inter-rater reliability remain open questions for future work.

5 Conceptual Blending as an Evaluation Framework for Creativity

5.1 Conceptual–Algorithmic Divide

Generative AI is increasingly used in creative domains. However, recent experimental evidence reveals a paradox in creative generative AI: while generative models can enhance individual creativity, they simultaneously reduce collective novelty by inducing convergence across outputs (Doshi and Hauser, 2024). This convergence suggests that creativity in the current state of GenAI is mainly driven by surface-level pattern reuse, and models generate variations rather than genuinely emergent conceptual structures. Although "language" is the foundational aspect of large language models, contributions from linguistics remain underutilized. We introduce conceptual blending as a theoretical and practical framework and argue that its core mechanisms can be induced in generative models. This position paper advocates the integration of conceptual blending into the research pipeline focused on creativity in generative AI.

5.2 Theoretical Discussion

We identify several persistent challenges and suggest how frame blending may help bridge the divide between conceptual and algorithmic approaches. First, there is no shared standard for evaluating creativity in generative systems. Without a consistent framework, it is difficult to compare benchmarks or generalize findings across studies. Conceptual blending theory offers a principled alternative by modeling creativity as a domain-general cognitive process through which distinct mental spaces are selectively integrated to produce emergent structure. Second, technical communities often lack deep engagement with foundational research in cognitive science, linguistics, and psychology. Without this grounding, evaluation metrics may lack theoretical validity. We provide a foundation for cross-disciplinary research into creativity and associative thinking. Third, there is currently no unified prompting framework for eliciting creativity in large language models. The FrameNet lexical resource enables structured metrics for grammatical and semantic evaluation, making the results more interpretable and reproducible.

Conceptual blending complements recent work on "Leap-of-Thought," which simulates non-sequential thinking by encouraging associative links between seemingly unrelated concepts (Zhong et al., 2024). Another study designed a quantitative framework to measure the cognitive responsiveness of LLMs, examining how models respond to fused versus isolated conceptual prompts (Sato, 2025b). As pathways to associative thinking and creative hallucination, prompt engineering approaches grounded in conceptual blending enable explicit control over the conceptual integration process, allowing models to actively shape how concepts are blended and to generate incommensurable concepts that reflect divergent associative reasoning. Such prompts can elicit imaginative outputs by encouraging the fusion of disparate concepts, turning what might otherwise be dismissed as hallucinations into sites of creative synthesis. This approach offers an opportunity to trace the origins of generative outputs within the model's reasoning process. This position paper also outlines future research directions in creative GenAI with conceptual blending, focusing on three key areas: using frame blending as a framework to better understand and facilitate the creativity and associative thinking of LLMs, designing quantitative experiments for model evaluation, and improving model explainability.

5.3 Technical Discussion

Beyond conceptual representations, several steering strategies guide generative models toward desired behaviors without fully retraining them. These approaches, including prompt engineering and chaining, retrieval-augmented generation (RAG), activation engineering via techniques such as sparse autoencoders (SAEs), and lightweight fine-tuning methods like LoRA, operate at different points in the generation pipeline to constrain, bias, or reweight model outputs while preserving the underlying model architecture (Lewis et al., 2020; Hu et al., 2022; Ng, 2011). Unlike ad hoc approaches, the proposed framework integrates diverse components, including model selection, data design, prompt engineering, human-AI interaction, and evaluation metrics, into a unified research pipeline. In this section, we propose a human-in-the-loop system based on frame blending to enhance the creativity of LLMs, discuss the potential technical challenges, and invite scholars to further develop this emerging field.

The **first component**, model integration, relies on a modular structure that allows for flexible loading and integration of fast-moving LLMs, making

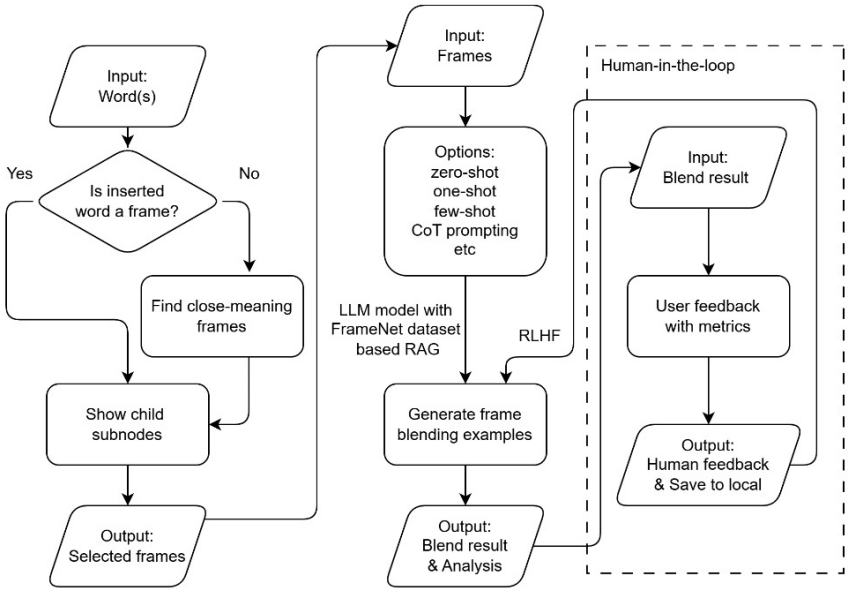

Figure 2: GenAI-integrated conceptual blending system flowchart

the system adaptable to evolving needs. The **second component** of the architecture is the reference lexical resources, including FrameNet and beyond (Baker et al., 1998; Ruppenhofer et al., 2016). The semantic richness of FrameNet provides an ideal scaffold for guiding creative generation. FrameNet entries are indexed and embedded using HuggingFaceEmbeddings, which enables the creation and maintenance of a persistent vector store for use in Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). To benefit cross-disciplinary collaborators, we also propose the architecture to support multimodal data, such as visual metaphors or image-based frame elicitation, to investigate how frame blending might generalize beyond textual modalities. These extensions draw on both corpus evidence and cognitive studies to enrich the representational framework. The **third component** is prompt engineering. Frame blending prompts are tailored to incorporate specific elements from lexical resources, including frame definitions, semantic roles, and example sentences, thereby offering contextually grounded guidance for language generation. The **fourth component** is the human-in-the-loop design. Our proposed trial system includes an interactive terminal interface optimized for use with GPU-based high-performance computing platforms. Researchers can input desired frames, select prompting strategies, and evaluate the generated blends. They can also provide qualitative feedback and rank outputs using defined evaluation metrics. This infrastructure supports techniques such as Reinforcement Learning from Human Feedback (RLHF) and helps refine the model's capacity for high-quality results (Ouyang et al., 2022).

Appendix Table A1 presents representative outputs from the proposed LLM-integrated frame blender system, demonstrating the capacity of the LLaMA series to perform conceptual integration through frame blending. Given two input frames, the model generates a blended sentence and a corresponding explanation that captures the emergent semantics arising from the mappings. For instance, in the Giving birth and Success or failure example, the model constructs a metaphorical blend that conceptualizes childbirth as a financial investment, illustrating how the model combines the two input frames into a single metaphorical mapping. These results illustrate that LLMs can produce outputs that exhibit interpretability, coherence, and cognitively plausible reasoning. To illustrate the potential of the proposed workflow, Figure 2 presents the system architecture, which consists of multiple interconnected components. The left sub-flow handles frame retrieval by checking whether an input word is a frame and, if not, identifying semantically close frames and their subframes. The right sub-flow integrates LLMs with FrameNet-enhanced retrieval-augmented generation (RAG) and incorporates prompt-based control (Lewis et al., 2020).

The generated blend and reasoning are then subject to human-in-the-loop evaluation, enabling iterative tuning. Appendix Figure A1 shows the system user interface, which includes interactive fields to input frame pairs, display generated blending outputs, and capture user evaluations. This user interface provides interactive demos that support transparent model auditing and facilitate human feedback collection, enabling systematic evaluation of the quality, novelty, and coherence of the generated blends. This system represents a step forward in operationalizing conceptual blending as a research direction and pipeline, with complementary developments improving its modularity, scalability, and automation. By making the blending process tractable and scalable, our position paper intends to lay the theoretical and computational foundation for steering creative GenAI, while motivating systematic investigation into how generative models can be interpreted through cognitive theory.

6 Alternative Views

6.1 Creativity Emerges Without Explicit Mechanisms or Evaluation

One alternative view holds that explicit steering mechanisms are unnecessary: models trained on increasingly diverse data will implicitly acquire the ability to generate creative, conceptually novel outputs. On this view, explicit evaluation and explainability matter little for creative generative AI, since creativity does not require transparent mechanisms and may even benefit from opaque ones. Black-box models are seen as capable of creative outputs without interpretable representations or principled evaluation. Recent evidence does show that generative AI can enhance perceived creativity through incremental elaboration and stylistic refinement (Doshi and Hauser, 2024). But such gains are often accompanied by increased similarity and anchoring across outputs, suggesting that without mechanisms for cross-domain integration, creativity remains surface-level rather than emergent or meaning-constructive (Doshi and Hauser, 2024). Incremental creativity, stylistic variation, and recombination around a single concept can typically be achieved through interpolation, retrieval, or stochastic variation over learned representations. Such models produce fluent outputs but rely on statistical associations from training data, and without identifiable points of intervention, emergence alone cannot deliver the control required for reproducible or targeted creative outcomes.

In contrast, conceptual blending is well suited to forms of creativity that depend on cross-domain integration, emergent structure, and meaning construction, such as metaphor creation, analogical reasoning, narrative innovation, and scientific or artistic insight that cannot be reduced to surface-level variation. By providing an explicit mechanism for combining conceptual structures, it enables systematic steering and reproducibility rather than occasional emergence. We therefore argue that the opposing view overlooks the role of explainability in conceptual control, transferability, and generalization. Black-box processes may appear creative in narrow settings, but without understanding the mechanisms behind generation, creativity risks remaining task-specific, non-transferable, and prone to collective homogenization. Conceptual blending addresses this by making semantic structures explicit and manipulable, allowing creators and researchers to steer generation toward specific conceptual relations rather than relying on stochastic variation.

6.2 Conceptual Blending Is Not Operational

Another alternative view is that conceptual blending is merely a theoretical construct with limited applicability to quantitative implementation. We argue instead that conceptual blending, combined with the FrameNet lexical resource, offers numerous opportunities for operationalization. Grounded in frame semantics and corpus evidence, FrameNet provides a rich, structured foundation for applying frame blending theory in quantitative experiments, with annotated example sentences, automatically tabulated valence reports, and active maintenance. The FrameNet database contains well over ten thousand lexical units spanning more than a thousand semantic frames, exemplified in hundreds of thousands of annotated sentences (Baker et al., 1998; Ruppenhofer et al., 2016). Whereas existing creativity benchmarks often lack theoretical grounding in cognitive science, this resource can serve as infrastructure for reproducible and controlled experimentation, letting researchers manipulate input variables and analyze the resulting creative outputs. Its structured semantics offer a cognitively informed basis for defining and measuring novelty, allowing researchers to construct prompts, specify controlled variables, and perform standardized quantitative evaluations (Baker et al., 1998; Ruppenhofer et al., 2016).

7 A Computational Trajectory: Operationalizing Conceptual Blending

7.1 Frame Embeddings

Prior computational work represents a sustained effort to integrate conceptual blending theory and FrameNet with emerging technologies, treating advances in NLP and large language models not merely as tools but as infrastructure for modeling key elements of human cognition. Beginning with the detection of blends in conversations about the future, the early phase laid the theoretical foundation by surveying the literature on conceptual blending, FrameNet, and semantic parsing, assessing existing parsing tools, and developing "Frame Embedding," a word-embedding-inspired approach for quantifying semantic relationships between frames (Xi, 2020). Building on these representations, subsequent work designed heuristic algorithms to identify candidate blends from co-occurrence and semantic proximity in text corpora.

7.2 Integrating FrameNet with LLMs

The following phase integrated FrameNet's structured data with LLMs to support frame-blending generation (Cheng, 2024). A parser extracted frame names, definitions, frame elements, lexical units, and inter-frame relationships from FrameNet into a structured format, with tools to visualize frame hierarchies. To ground outputs in factual FrameNet knowledge, this work implemented a Retrieval-Augmented Generation (RAG) pipeline using LlamaIndex, building a vector index from the parsed documents and querying it with LLaMA-family models for prompt construction (Cheng, 2024). A prompting module supported zero-shot, one-shot, few-shot, and chain-of-thought strategies, and an interactive terminal application provided an accessible interface for exploring frames and generating blends. This infrastructure makes FrameNet data computationally accessible and lays a foundation for interpretable creative AI systems.

7.3 Scalable and Automated Evaluation

The subsequent phase emphasized modularity, scalability, and automation. Evaluating the creativity and coherence of generated blends remains challenging, and evaluation methods continue to evolve alongside rapidly advancing models. To address this, structured evaluation tools were integrated into the interactive interface, enabling users to score blends on the metrics: completeness, clarity, relevance, understanding depth, coherence, and execution time. To scale this, an automated LangGraph stage (Soni, 2025) adds a validation agent that checks coherence and an evaluation agent that assigns LLM-generated scores and justifications from structured prompts.

8 Conclusion

This position paper argues that conceptual blending offers a principled framework for evaluating creative generative AI, one that is interpretable, theory-grounded, and reproducible. Rather than reducing creativity to surface-level novelty, fluency, or stylistic variation, we frame it as the integration of distinct conceptual spaces into coherent blends that yield emergent meaning, and we make that integration explicit and measurable through FrameNet-based frame blending, retrieval-augmented generation, and a multidimensional scoring rubric. This perspective directly targets three persistent weaknesses in current practice: output homogenization, limited controllable steering, and the weak theoretical grounding of existing creativity metrics. If the field continues to rely on ad hoc proxies for novelty, it will remain unable to distinguish stylistic variation from genuine conceptual innovation. We therefore advocate grounding creativity evaluation in cognitive linguistics and computational semantics, and we outline concrete pathways for operationalizing these insights to bridge the conceptual and algorithmic divide. We invite the community to build on this foundation by expanding the supporting infrastructure and developing principled methods for evaluating, explaining, and steering creativity in generative AI.

9 Limitations

As a position paper, this work proposes a framework rather than a validated pipeline, and several limitations follow from that scope. Our empirical illustration is intended to demonstrate feasibility, not generalizability; establishing the latter would require systematic studies across diverse models and prompting strategies. The proposed scoring rubric is theoretically motivated but not yet empirically validated—annotation protocols, inter-rater reliability, and benchmarking all remain to be established. Finally, scaling feedback collection and ensuring reproducibility in the human-in-the-loop design remain for future work.

References

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Ronald A Beghetto and James C Kaufman. 2007. Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of aesthetics, creativity, and the arts*, 1(2):73.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2026. Divergent creativity in humans and large language models. *Scientific Reports*, 16:1279.

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

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*, pages 132–155.

Honghua Chen and Nai Ding. 2023. Probing the "creativity" of large language models: can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888.

Zenan Chen and Jason Chan. 2024. Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12):9101–9117.

Zhongheng Cheng. 2024. Google summer of code 2024 frame blender.

Jayanth Krishna Chundru, Rudrashis Poddar, Jie Cao, and Tianyu Jiang. 2025. Do llms encode frame semantics? evidence from frame identification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29476–29488.

Anca Dinu and Andra-Maria Florescu. 2025. Testing language creativity of large language models and humans. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 426–436.

Paul V DiStefano, John D Patterson, and Roger E Beaty. 2025. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, 37(4):555–569.

Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290.

Gilles Fauconnier and Mark Turner. 1998. Conceptual integration networks. *Cognitive science*, 22(2):133–187.

Gilles Fauconnier and Mark Turner. 2003. Conceptual blending, form and meaning. *Recherches en communication*, 19:57–86.

Gilles Fauconnier and Mark Turner. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books.

Charles J Fillmore and Beryl T Atkins. 2012. Toward a frame-based lexicon: The semantics of risk and its neighbors. In *Frames, fields, and contrasts*, pages 75–102. Routledge.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Ana Luisa Ilha Villanova and Miguel Pina e Cunha. 2021. Everyday creativity: A systematic literature review. *The Journal of Creative Behavior*, 55(3):673–695.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

James C Kaufman and Ronald A Beghetto. 2009. Beyond big and little: The four c model of creativity. *Review of general psychology*, 13(1):1–12.

Arthur Koestler. 1964. *The Act of Creation*. Hutchinson, London.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*.

Surabhi S Nath, Peter Dayan, and Claire Stevenson. 2024. Characterising the creative process in humans and large language models. *arXiv preprint arXiv:2405.00899*.

Andrew Ng. 2011. Sparse autoencoder. *CS294A Lecture Notes*, 72:1–19.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Francisco Pereira, Matthew Botvinick, and Greg Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252.

Josef Ruppenhofer, Michael Ellsworth, Miriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute.

Makoto Sato. 2025a. Triggering hallucinations in llms: A quantitative study of prompt-induced hallucination in large language models. *arXiv preprint arXiv:2505.00557*.

Makoto Sato. 2025b. Waking up an ai: A quantitative framework for prompt-induced phase transition in large language models. *arXiv preprint arXiv:2504.21012*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581.

Murray Shanahan and Catherine Clarke. 2023. Evaluating large language model creativity from a literary perspective. *arXiv preprint arXiv:2312.03746*.

Vladimir M Sloutsky. 2010. Mechanisms of cognitive development: Domain-general learning or domain-specific constraints? *Cogn. Sci.*, 34(7):1125–1130.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, and 12 others. 2025. Evaluating the social impact of generative ai systems. In *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press.

Deepak Kumar Soni. 2025. Enhancing FrameBlender with LangChain and Ollama. Medium. Accessed: 2026-05-05.

Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. 2025. Large language models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 57:101870.

Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024. *Copilots for linguists: AI, constructions, and frames*. Cambridge University Press.

Tiago Timponi Torrent, Arthur Lorenzi, Ely Edison Matos, Frederico Belcavello, Marcelo Viridiano, and Maucha Andrade Gamonal. 2022. Lutma: A frame-making tool for collaborative FrameNet development. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 100–107.

Mark Turner. 2014. *The origin of ideas: Blending, creativity, and the human spark*. Oxford University Press.

Tony Veale and Diarmuid O'Donoghue. 2000. Computation and blending. *Cognitive linguistics*, 11(3/4):253–281.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Suzie Xi. 2020. Google summer of code 2020 frame blends nomination system.

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

A Appendix

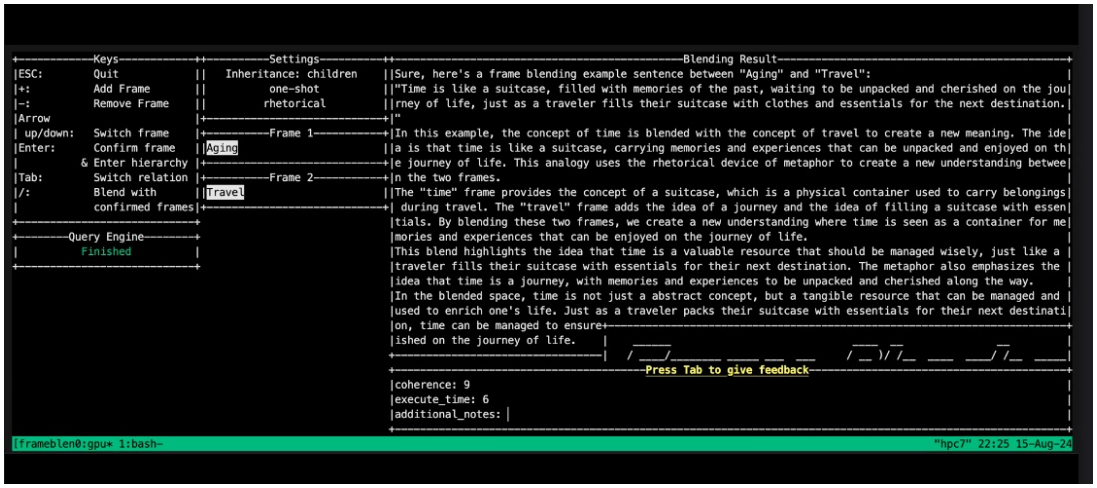

Figure A1: Interactive conceptual blending interface

Table A1: Sample conceptual blending results generated by the LLaMA series

Example 1	
Frame 1	Giving birth
Frame 2	Success or failure
Result	The successful delivery of her baby was like hitting a financial jackpot for Sarah, as every moment of the labor process felt like a valuable investment in her future.
Reasoning	The sentence creates a new understanding by mapping the concepts of time and money onto each other. The idea is that just as financial transactions involve investing time and money, giving birth is like a successful investment in one's future.
Example 2	
Frame 1	Emotion
Frame 2	Kinship
Result	Emotion is like kinship — worth investing in, but easily squandered if not managed properly.
Reasoning	In the blended space, emotions are seen as valuable resources that can be invested in social relationships, much like how money can be invested in relationships.
Example 3	
Frame 1	Aging
Frame 2	Delivery
Result	Time is like gold, aging quickly but valuable delivery matters most.
Reasoning	In the blended space, aging is conceptualized as a process that can be managed like a delivery service. Time is seen as a valuable resource that must be delivered efficiently, without being wasted or lost. The sentence emphasizes the value of time and the need to prioritize its delivery, much like how packages or gifts must be delivered promptly and efficiently.