Beyond Noise: Characterizing Creative Potential in Unverifiable LLM Hallucinations

Yu Yan[1] **Chunhong Zhang**[1] **Haiyu Zhao**[1] **Ziyang Zeng**[1] **Zihao Liu**[1]
Yongkang Wu[1] **Jianzhou Diao**[1] **YiJie Chen**[1] **Shujie Wang**[1] **Zheng Hu**[1]*
[1]Beijing University of Posts and Telecommunications
{yanyu2023, zhangch, huzheng}@bupt.edu.cn

Abstract

In knowledge-intensive creative tasks, Large Language Models (LLMs) often generate outputs that extend beyond established knowledge, making direct verification against current evidence impractical. Unlike factual hallucinations checked against ground truth, such outputs arise naturally in creative generation, where extending beyond current knowledge is often the goal. Yet prior work debates whether hallucination should be suppressed or embraced without empirically analyzing this unverifiable subclass. On the ideation evaluation side, existing work focuses on individual outputs without characterizing the unverifiable space as a whole. To address this gap, we propose a novelty-verifiability characterization that distinguishes *Creative Synthesis* (Region A) from *Groundless Fabrication* (Region B), and study it through a *conceptual creation* task where LLMs synthesize novel scientific concepts. Through 32,400 generations across three technical domains and 1,080 human judgments, we find that Region A is non-negligible (4.7%) and robust, persisting across generation strategies, models, domains, and embedding choices. A retrospective recovery experiment further shows that LLMs can approximate post-cutoff scientific concepts in controlled combinatorial settings. Our findings suggest that the unverifiable space is not uniformly noise but exhibits empirically distinguishable internal structure, providing an empirical basis for more selective hallucination governance.[1]

1 Introduction

Large Language Models (LLMs) are increasingly used for knowledge creation, including novel protein design (Shin et al., 2021), molecule generation (Bagal et al., 2022), code synthesis (DeLorenzo et al., 2024), story writing (Gómez-Rodríguez and Williams, 2023; Buz et al., 2024). In such creative generation settings, LLMs are expected not merely to reproduce established knowledge, but to extend beyond it. Consequently, many outputs are inherently *unverifiable*: they fall beyond what existing evidence can directly confirm or refute. Following the hallucination taxonomy of Huang et al. (2025), we refer to such beyond-boundary generations as *unverifiable hallucinations*. Unlike factual hallucinations in question-answering, where outputs can be checked against ground truth and should be suppressed (Bang et al., 2023; Ji et al., 2023b), scientific discovery workflows often require models to propose candidates that extend beyond currently established evidence.

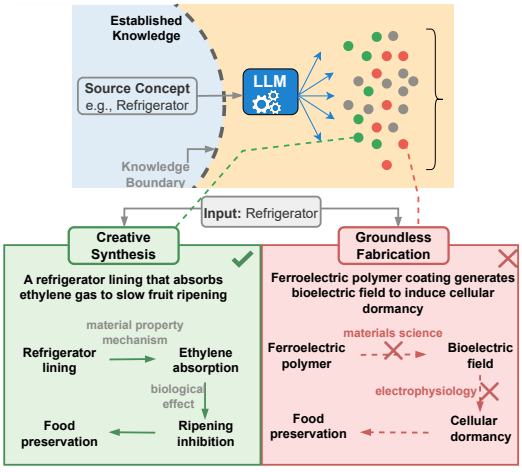


Figure 1: Characterization of unverifiable hallucinations and their internal heterogeneity.

However, prior work has left a basic question unanswered: *what structure exists within the space of unverifiable hallucinations?* On the intervention side, hallucination mitigation studies typically aim to suppress unsupported generations (Dhuliawala et al., 2024; Ji et al., 2023b), while work linking hallucination to creativity suggests such outputs may contribute to creative generation (Jiang et al., 2024). Despite their opposing stances, both re-

*Corresponding author.
[1]https://github.com/YuLab1/llm-concept-creation

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12102–12124
July 2-7, 2026

main at the level of broad conceptual claims about hallucination, without fine-grained empirical analysis of the unverifiable subclass. On the evaluation side, scientific ideation systems assess individual ideas (Si et al., 2024; Guo et al., 2025), but they do not characterize the internal structure of unverifiable outputs at the population level. This leaves a key interpretability gap: we lack a fine-grained characterization of the unverifiable space and of how its different regions vary in grounding, coherence, and creative potential.

As shown in Figure 1, unverifiable outputs can differ fundamentally. For instance, given *refrigerator* as an input concept, an LLM may produce *a refrigerator lining that absorbs ethylene gas to slow fruit ripening*—a coherent extrapolation grounded in established food-chemistry principles. It may also produce *a refrigerator whose ferroelectric polymer coating generates a bioelectric field to induce cellular dormancy in stored produce*—a concept that chains real terminology from materials science, electrophysiology, and botany into a plausible-sounding but mechanistically groundless narrative. Both outputs are unverifiable, yet without fine-grained characterization, they would be indistinguishably grouped, obscuring outputs with genuine creative potential.

To address this gap, we propose a Novelty–Verifiability characterization that maps unverifiable outputs into two structurally distinct regions: *Creative Synthesis* (Region A) and *Groundless Fabrication* (Region B). We ground this investigation in a *conceptual creation* task, where LLMs are guided to synthesize new scientific or technical concepts by modifying, recombining, or extending existing domain knowledge. Inspired by psychological theories of human creativity (Guilford, 1950; BESEMER and TREFFINGER, 1981), we design complementary generation strategies that probe both the *process* and *product* dimensions of creative output (see Figure 2). Through 32,400 generations across three technical domains and 1,080 human judgments, we identify three key findings:

- **Region A exists and is non-negligible:** Among highly novel outputs, a non-negligible subset (4.7%) is judged both novel and plausible by domain experts, indicating that unverifiable hallucinations are not uniformly noise. This subset persists even under stricter thresholds.
- **Region A is reproducible across conditions:** Different generation strategies access Region A through distinct behaviors—unconstrained generation explores broader novelty, while structured generation preserves stronger semantic grounding. This pattern is consistent across domains, models, and embedding choices, suggesting that Region A is not a strategy-specific artifact.
- **Proximity to real-world innovation:** In a retrospective recovery experiment under a controlled combinatorial setting, LLM generation recovers 85% of real scientific concepts published after the model's training cutoff, suggesting that LLMs can approach real-world innovation in relatively simple recombination scenarios.

2 Related Work

Hallucination and Creativity in LLMs LLM hallucinations have traditionally been viewed as a sign of reduced reliability, motivating a wide range of detection and mitigation strategies (Manakul et al., 2023; Dhuliawala et al., 2024; Ji et al., 2023b; Li et al., 2023). However, recent studies have begun to reframe hallucinations as a potential source of creative association. From a cognitive science perspective, Jiang et al. (2024) argued that hallucination and creativity share underlying mechanisms, while Lee (2023) analyzed their relationship using probabilistic and information-theoretic formulations. In scientific settings, Yuan et al. (2025) further showed that guided hallucination can facilitate drug discovery. Unlike prior work, we focus on characterizing unverifiable hallucinations as a distinct subclass whose internal structure may reflect latent creativity rather than mere noise.

Scientific Idea Generation with LLMs Recent advances have shifted LLMs from passive assistants toward more active roles in scientific idea generation and hypothesis formulation. In the general domain, IdeaBench (Guo et al., 2025) and LiveIdeaBench (Ruan et al., 2026) evaluate LLMs' ability to generate research ideas from literature summaries. Several exploratory studies on LLM-generated ideas in AI research highlighted the potential for automated idea generation with human guidance (Si et al., 2024; Feng et al., 2025; Qiu et al., 2025). Furthermore, studies in natural sciences have explored LLM-based idea generation (Ciucă et al., 2023; Buehler, 2024). In contrast, prior works mainly focus on benchmarking generative capability or demonstrating domain-specific utility. Our work instead focuses on how unver-

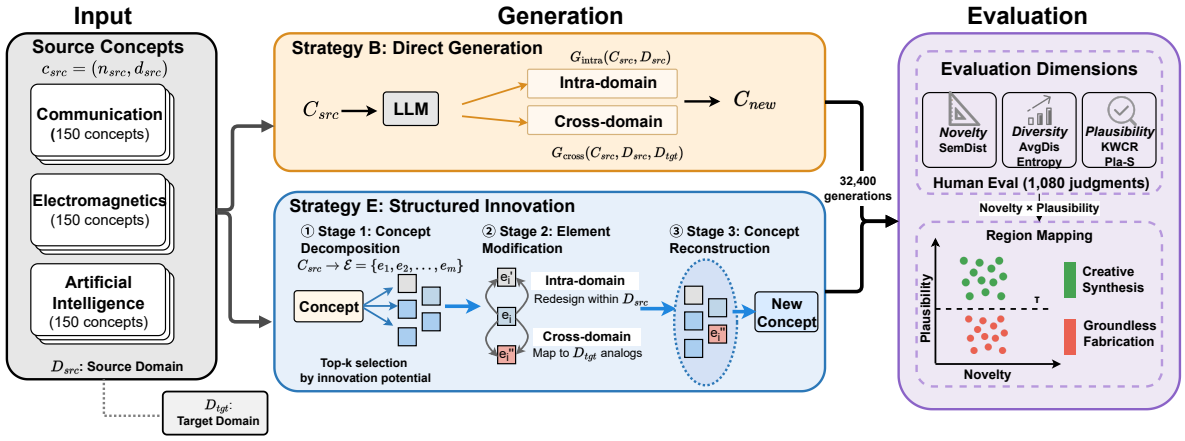


Figure 2: Overview of our proposed framework for exploring conceptual creativity in LLMs.

ifiable conceptual outputs should be interpreted, especially when they may reflect creative value.

3 Problem Formulation

3.1 Novelty-Verifiability Conceptualization

To analyze the relationship between unverifiable hallucinations and creativity, we propose a two-dimensional conceptual framework based on *novelty* and *verifiability*, as shown in Figure 3. We divide verifiable outputs into two categories: *Knowledge Reproduction* (Quadrant IV) and *Empirical Innovation* (Quadrant I). Knowledge Reproduction reflects the memorization and compression of training data (Carlini et al., 2021; Kiyomaru et al., 2024). Empirical Innovation refers to novel hypotheses that remain grounded in existing knowledge or expert consensus (Qi et al., 2023; Yang et al., 2024). For example, the Scideator system (Radensky et al., 2026) explores human–LLM collaboration for grounded scientific idea generation. In contrast, outputs lacking both novelty and verifiability fall into *Degenerate Generation* (Quadrant III), which captures structural failures or meaningless sequences (Huang et al., 2025; Ji et al., 2023a). Distinct from these categories, our investigation focuses on *Boundary Exploration* (Quadrant II). Although these outputs lack immediate empirical support, structurally coherent and semantically well-formed cases may still indicate unexplored regions of the knowledge space. We further divide this quadrant into two representative sub-regions.

Region A: Creative Synthesis These outputs are internally coherent and plausibly connected to existing knowledge frameworks, despite lacking external factual verification. This behavior may

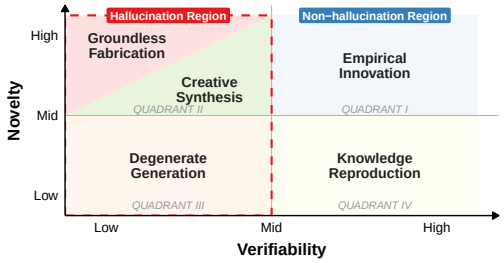


Figure 3: A quadrant-based taxonomy of generative outputs based on **novelty** and **verifiability**.

stem from the high-dimensional semantic representations of LLMs, which enable interpolation between related concepts and extrapolation into sparsely explored semantic regions (Mikolov et al., 2013; Park et al., 2024). While not yet formalized as knowledge, such outputs may exhibit developmental or exploratory potential.

Region B: Groundless Fabrication In contrast, these outputs exhibit apparent novelty but lack meaningful grounding in context, domain knowledge, or common sense. Although they may contain loosely related semantic elements, they fail to establish coherent logical connections with existing concepts. As a result, they produce fragmented or irreconcilable information that cannot be reliably integrated into existing knowledge frameworks.

3.2 Task Formalization

We formalize the conceptual creation task as follows. Let the source concept be $C_{src} = (n_{src}, d_{src})$ where n_{src} denotes the concept name, d_{src} its definition. Let $\mathcal{D}_{src}$ denote the source domain and $\mathcal{K}$ the relevant knowledge base. The goal is to generate a novel concept $C_{new} = (n_{new}, d_{new})$. Within

Quadrant II, we operationalize Region A and Region B using a plausibility function $\mathcal{P}(\cdot)$:

$$\text{Region A:} \quad \{C_{new} \mid \mathcal{P}(C_{new} \mid C_{src}, \mathcal{K}) \geq \tau\} \quad (1)$$

$$\text{Region B:} \quad \{C_{new} \mid \mathcal{P}(C_{new} \mid C_{src}, \mathcal{K}) < \tau\} \quad (2)$$

where τ denotes the plausibility sufficiency threshold. In practice, we estimate $\mathcal{P}(\cdot)$ using automatic metrics and human evaluation. We consider two generation pathways: (1) *Intra-domain extension* denoted by $G_{intra}(C_{src}, \mathcal{D}_{src})$, which transforms the source concept within its original domain; (2) *Cross-domain mapping* denoted by $G_{cross}(C_{src}, \mathcal{D}_{src}, \mathcal{D}_{tgt})$, which synthesizes interdisciplinary concepts by mapping the source concept into a target domain $\mathcal{D}_{tgt}$.

4 Methods

We employ two complementary generation strategies: an unconstrained approach that captures LLMs' natural generative behavior, and a structured approach that enforces compositional constraints. The outputs are evaluated using both automatic metrics and human judgment.

4.1 Baseline: Direct Generation

This baseline strategy captures the natural creative behavior of LLMs with minimal intervention. Given a source concept $C_{src} = (n_{src}, d_{src})$, the model receives only the concept name n_{src} and definition d_{src}, and is prompted to generate a related novel concept $C_{new} = (n_{new}, d_{new})$. For intra-domain generation, the prompt constrains outputs to remain within $\mathcal{D}_{src}$; for cross-domain generation, it specifies the target domain $\mathcal{D}_{tgt}$.

4.2 Structured Element-Based Innovation

Inspired by psychological research on divergent-convergent creative processes (Guilford, 1950), we propose a modular three-stage generation framework implemented using LangChain (Chase, 2022), as shown in Figure 2.

4.2.1 Stage 1: Concept Decomposition

In this phase, the LLM analyzes the input concept definition d_{src} to extract a set of core semantic elements $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$, where each element represents a modular component of the concept. The model then scores each element for its *innovation potential*, defined as its potential to yield creative variants under modification. We select the top-k elements with the highest scores from the extracted set $\mathcal{E}$, denoted as $\mathcal{E}_{top}$.

4.2.2 Stage 2: Element Modification

Different transformation methods are applied to the selected elements depending on the innovation pathway. For *intra-domain generation* ($G_{intra}(C_{src}, \mathcal{D}_{src})$), the model is prompted to redesign each element within the constraints of the source domain $\mathcal{D}_{src}$, producing domain-consistent variants. For *cross-domain generation* ($G_{cross}(C_{src}, \mathcal{D}_{src}, \mathcal{D}_{tgt})$), the selected elements are mapped to analogous or complementary elements from a target domain $\mathcal{D}_{tgt}$ based on functional or theoretical consistency. This yields a set of modified element variants $\mathcal{V} = \{e'_1, e'_2, \ldots, e'_k\}$.

4.2.3 Stage 3: Concept Reconstruction

The modified element variants $\mathcal{V}$ are reintegrated into the original conceptual context to form a coherent new concept $C_{new} = (n_{new}, d_{new})$. In this stage, the model reintegrates the modified elements into the original conceptual context to produce a coherent new concept while maintaining plausibility with respect to the source concept C_{src} and the relevant knowledge base $\mathcal{K}$.

4.3 Automatic Metrics

We evaluate generated concepts on three dimensions: *novelty*, *diversity*, and *plausibility*.

4.3.1 Novelty

Semantic Distance (SemDist) We measure novelty as the cosine distance between the embeddings of the source and generated concepts: $\text{SemDist}(C_{src}, C_{new}) = 1 - \frac{\mathbf{e}_{C_{src}} \cdot \mathbf{e}_{C_{new}}}{\|\mathbf{e}_{C_{src}}\| \|\mathbf{e}_{C_{new}}\|}$, where $\mathbf{e}_{C_{src}}$ and $\mathbf{e}_{C_{new}}$ denote the embeddings of the source and generated concepts. Larger values indicate greater conceptual departure.

4.3.2 Diversity

We assess diversity from both local geometric and global distributional perspectives.

Pairwise Distance (AvgDis) AvgDis captures diversity at the pairwise embedding level by computing the average pairwise cosine distance. Given a set of generated concepts $\{C_1, C_2, \ldots, C_T\}$, AvgDis is defined as: $\text{AvgDis} = \frac{2}{T(T-1)} \sum_{1 \leq i < j \leq T} \left(1 - \frac{\mathbf{e}_{C_i} \cdot \mathbf{e}_{C_j}}{\|\mathbf{e}_{C_i}\| \|\mathbf{e}_{C_j}\|}\right)$. Higher values indicate the broader conceptual exploration.

Semantic Entropy We measure diversity at the distributional level using semantic entropy. For

each generation set, we compute concept embeddings, project them into a lower-dimensional semantic space, and cluster them into semantic groups. Let $p(c)$ denote the proportion of generated concepts assigned to cluster c. Semantic entropy is computed as $H = -\sum_c p(c) \log p(c)$. Higher entropy implies broader distribution across functionally distinct semantic regions.

4.3.3 Plausibility Metrics

Keyword Coverage (KWCR) We measure plausibility-related semantic preservation using Keyword Coverage (KWCR), which evaluates how well the generated concept retains the core semantic lineage of the source concept. We use KeyBERT (Grootendorst, 2020) to extract top-N_1 weighted keywords $\mathcal{W}_{src} = \{(k_i, w_i)\}_{i=1}^{N_1}$ from the source concept, where w_i is the normalized importance weight of keyword k_i. For the generated concept, we extract the top-N_2 keywords and denote the resulting keyword set as $\mathcal{W}_{new} = \{k'_j\}_{j=1}^{N_2}$. KWCR is computed as: $\text{KWCR}(C_{src}, C_{new}) = \sum_{i=1}^{N_1} (w_i \cdot \max_{k' \in \mathcal{W}_{new}} \text{sim}(k_i, k'))$, where $\text{sim}(k_i, k')$ denotes the cosine similarity between the embeddings of source keyword k_i and generated keyword k'. Higher values indicate stronger preservation of core meaning.

LLM-as-Judge Plausibility (Pla-S) We leverage the LLM's broad contextual knowledge to assess whether C_{new} aligns with C_{src} in a scientifically coherent manner. LLM-based judges are prone to overconfidence, often assigning high plausibility scores to outputs that appear coherent but lack sufficient theoretical support (Geng et al., 2024). To mitigate this issue, we adopt a bidirectional verification scheme inspired by the dialectical interplay of supportive inference and critical scrutiny in scientific discovery. *Forward Progression* (S_{fwd}) assesses whether C_{new} can be coherently inferred from C_{src}, while *Backward Scrutiny* (S_{bwd}) identifies theoretical violations. Accordingly, we operationalize the plausibility function as $\mathcal{P}(C_{new} \mid C_{src}, \mathcal{K}) = \frac{1}{2}(S_{fwd} + S_{bwd})$. Detailed prompts are provided in Appendix B.

4.4 Human Evaluation

We conduct a human evaluation to validate our automatic metrics and to provide human judgments for Region A and Region B classification. The study follows established best practices for text generation assessment (van der Lee et al., 2019) and expert annotation for high-cognitive-load tasks (Clark et al., 2021). Eight researchers from the author team, all with relevant domain backgrounds, serve as annotators. Before the main evaluation, all eight annotators rate a fully overlapped calibration set of 25 samples to align their interpretation of the rubric and scoring criteria. We use a stratified set of 360 samples covering 6 domains (3 intra-domain and 3 cross-domain) and two generation strategies, with 30 samples per cell. Each sample is independently rated by three annotators under a fully blind setting, resulting in 1,080 human judgments. Annotators rate each generated concept on three 5-point Likert scales: *novelty*, *plausibility*, and *heuristic value*. Final scores are computed by averaging ratings.

5 Experiments

In this section, we aim to address the following research questions:

RQ1: Are unverifiable hallucinations merely random noise, or do they in fact contain plausible creative extensions?

RQ2: If Region A exists, what properties characterize these concepts, and can different generation strategies explore this space effectively?

RQ3: Can LLM-generated concepts approximate real scientific innovations, exhibiting human-like creativity in relatively straightforward scenarios?

5.1 Experimental Setup

We conduct experiments on three representative domains: *Communication*, *Electromagnetics*, and *Artificial Intelligence*, constructing a glossary of 150 concepts per domain. We evaluate three widely used LLMs: GPT-4o (OpenAI, 2024b), GPT-4o-mini (OpenAI, 2024b), and Qwen-max (Qwen Team, 2025). For each source concept, we generate baseline and structured outputs under both intra-domain and cross-domain pathways. For the structured strategy, we set $k = 3$. Each setting is repeated three times, yielding 24 generations per source concept and 32,400 generations in total across all settings. We use `all-MiniLM-L6-v2`[2] as the embedding model. Additional data details are provided in Appendix A.1.

5.2 RQ1: Plausible Extensions vs. Noise

5.2.1 Quantitative Analysis

Automatic Evaluation As a preliminary step, we examine automatic metrics across all generation

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Metric	Method	Intra-domain									Cross-domain								
		Comm			Ele			AI			Comm+Ele			Comm+AI			Ele+AI		
		4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen
SemDist	B	0.466	0.426	0.376	0.413	0.360	0.336	0.425	0.364	0.342	0.386	0.342	0.347	0.374	0.365	0.362	0.371	0.373	0.378
	E	0.359	0.313	0.312	0.321	0.282	0.284	0.335	0.305	0.279	0.319	0.286	0.269	0.351	0.344	0.327	0.358	0.360	0.339
KWCR	B	0.660	0.687	0.736	0.719	0.755	0.781	0.695	0.740	0.737	0.711	0.739	0.754	0.767	0.776	0.780	0.793	0.796	0.798
	E	0.765	0.802	0.804	0.814	0.847	0.844	0.768	0.790	0.796	0.760	0.784	0.801	0.769	0.770	0.791	0.782	0.785	0.806
AvgDis	B	0.337	0.267	0.240	0.311	0.308	0.243	0.299	0.262	0.226	0.266	0.179	0.163	0.164	0.169	0.124	0.149	0.147	0.115
	E	0.310	0.231	0.221	0.296	0.223	0.215	0.303	0.242	0.219	0.254	0.208	0.207	0.222	0.225	0.190	0.198	0.204	0.193
Pla-S	B	3.439	3.852	3.931	3.463	3.730	3.831	3.579	3.900	3.922	3.473	3.752	3.714	3.955	3.945	3.962	3.895	3.898	3.882
	E	3.505	3.914	3.936	3.541	3.862	3.911	3.711	3.899	3.937	3.674	3.752	3.785	3.919	3.923	3.924	3.810	3.776	3.808

Table 1: Transposed quantitative evaluation across domains and models. B: Prompt-Only Baseline, E: Structured Element-Based Innovation (highlighted with gray background). Models: 4o (GPT-4o), 4o-mini (GPT-4o-mini), Qwen (Qwen-max).

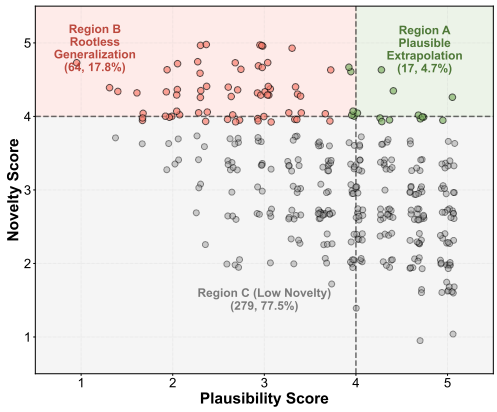


Figure 4: Scatter plot of generated concepts in the novelty-plausibility space.

Threshold (τ)	Region A	Region B	Region C
3.5 (Inclusive)	20 (5.6%)	61 (16.9%)	279 (77.5%)
4.0 (Current)	17 (4.7%)	64 (17.8%)	279 (77.5%)
4.5 (Strict)	6 (1.7%)	75 (20.8%)	279 (77.5%)

Table 2: Sensitivity analysis of region proportions under different plausibility thresholds, with Novelty ≥ 4 fixed.

conditions (Table 1) to assess whether generated concepts exhibit non-negligible plausibility rather than being dominated by random noise. Across all domains and generation strategies, Pla-S scores are consistently above 3, suggesting that novel concepts are not predominantly random noise.

Human Evaluation We then use human evaluation data to operationalize Region A and Region B based on novelty and plausibility ratings:

- **Region A (Creative Synthesis):** Novelty ≥ 4 AND Plausibility ≥ 4
- **Region B (Groundless Fabrication):** Novelty ≥ 4 AND Plausibility < 4

We adopt a plausibility threshold of 4, since a score of 3 in our rubric still allows minor flaws, such as logical leaps or terminology misuse. Figure 4 reveals an empirical separation among highly novel concepts (Novelty ≥ 4), with 4.7% falling into Region A and 17.8% into Region B. Table 2 shows that the presence of Region A is robust to different threshold choices: Region A remains non-zero even under a stricter threshold of $\tau = 4.5$. These results indicate that unverifiable hallucinations are not uniformly implausible, but include a non-negligible subset that maintains both novelty and plausibility. This subset aligns with our notion of *Creative Synthesis*, providing support for the existence of plausible creative extensions within unverifiable hallucinations.

5.2.2 Qualitative Analysis

To further characterize the nature of Region A concepts, we examine representative cases in light of recent literature. Although generated concepts are novel, we identify related studies that provide indirect or complementary validation. For example, concepts *Time-Domain Adaptive Neural BEM*, *Adversarial Predistortion in Microwave RF Systems*, and *Split-Step Neural Electromagnetic Method* extend classical methods along recognized computational and physical dimensions. These examples suggest that some LLM-generated concepts, although exploratory, are plausible extensions grounded in established problem formulations and supported by related research. Detailed analysis is provided in Appendix A.7.3.

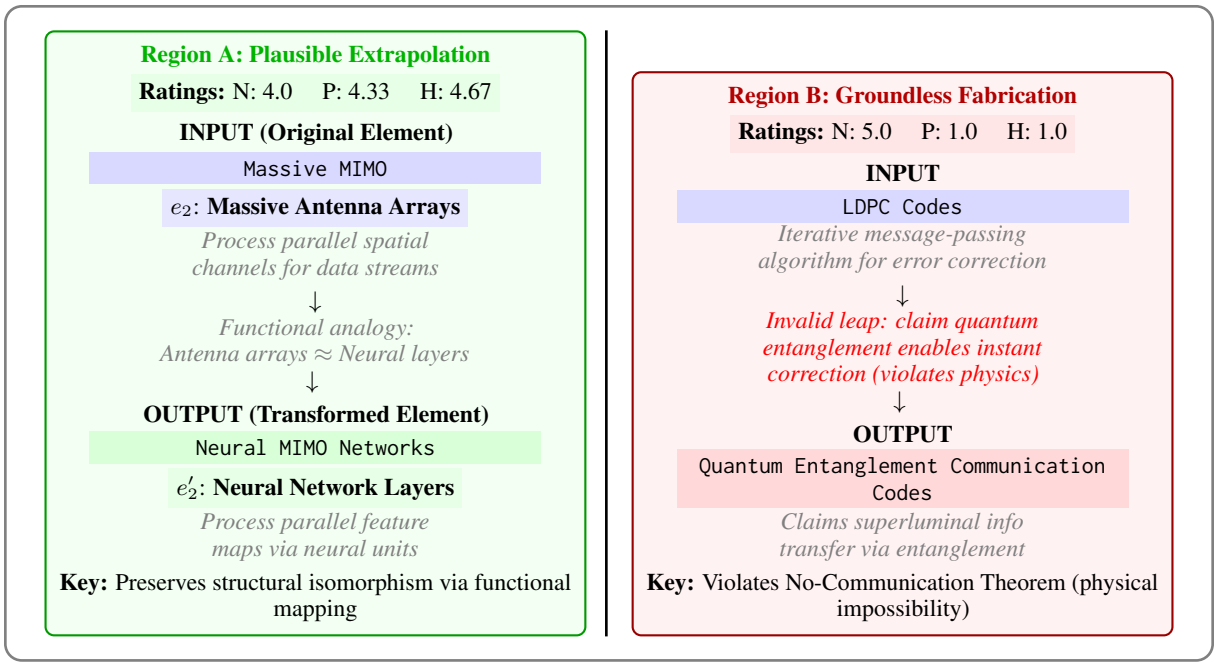


Figure 5: Contrastive case study of Region A vs B. **Massive MIMO** uses antenna arrays for parallel spatial processing; **Neural MIMO Networks** applies functional analogy to neural layers. **LDPC Codes** perform error correction via iterative decoding; **Quantum Entanglement Communication Codes** incorrectly claims quantum entanglement enables instantaneous correction, violating fundamental physics.

5.3 RQ2: Properties and Exploration Strategies

Having established the existence of Region A, we investigate its distinguishing properties and how different generation strategies explore this space.

5.3.1 Exploration through Complementary Strategies

Table 1 reports evaluation results for two complementary generation strategies. The structured strategy (Strategy E) generally yields lower semantic distance (SemDist) with higher keyword coverage (KWCR), reflecting conservative yet structurally grounded exploration. The baseline strategy (Strategy B) exhibits higher novelty (SemDist) with more variable coherence, reflecting more exploratory and divergent generation. Diversity metrics (AvgDis) further reveal contrasting patterns: Strategy E reduces dispersion in single-domain tasks but increases it under cross-domain generation, reflecting structured divergence when bridging disciplinary boundaries. Semantic entropy analysis further confirms that cross-domain transfer broadens distributional coverage (see Figure 11 in Appendix). For plausibility, Strategy E achieves higher scores in single-domain and Comm+Ele settings, while the gap becomes smaller in cross-domain configurations (Comm+AI, Ele+AI). Both strategies maintain moderate-to-high Pla-S scores across most settings, indicating that Region A is accessible through multiple generation pathways. We further verify these findings using SciBERT (Beltagy et al., 2019) embeddings, as shown in Table 4.

Domain	Novelty	Plausibility	Heuristic Value
Comm	3.49 ± 0.78	3.25 ± 1.07	3.23 ± 0.76
AI	2.98 ± 0.67	3.89 ± 0.74	3.20 ± 0.74
Ele	3.82 ± 0.75	3.31 ± 0.99	3.46 ± 0.81
Comm+AI	2.53 ± 0.49	4.38 ± 0.59	2.92 ± 0.63
Comm+Ele	2.52 ± 0.71	3.81 ± 0.90	2.61 ± 0.73
Ele+AI	3.35 ± 0.64	3.92 ± 0.79	3.56 ± 0.86

Table 3: Human Evaluation Results by Domain.

5.3.2 Domain-Specific Analysis

Domain characteristics substantially shape generation outcomes (Figure 6). Within single-domain settings, Electromagnetics exhibits the highest absolute KWCR, reflecting its comparatively stronger structural organization. Communication achieves the largest KWCR gains under Strategy E, suggesting a heightened sensitivity to explicit scaffolding, while Electromagnetics shows the most consistent improvements in plausibility across models. In cross-domain settings, Comm+AI achieves

the highest plausibility, suggesting a stronger degree of conceptual compatibility between the two domains, whereas Ele+AI shows high baseline KWCR with negligible gains, suggesting limited room for additional gains from Strategy E. These patterns suggest that Region A's accessibility varies with domain-specific conceptual structure and the compatibility between domains.

5.3.3 Model Comparison Analysis

Models exhibit differentiated responsiveness to structural intervention, as shown in Figure 6. **Qwen-max** demonstrates the most stable performance with consistent E-induced improvements and the smallest variance across settings. **GPT-4o-mini** shows the highest sensitivity to structural constraints, exhibiting the largest KWCR gains in intra-domain settings but greater variability under cross-domain transfer. **GPT-4o** achieves competitive plausibility across both intra- and cross-domain settings, balancing stability and adaptability.

5.3.4 Validation of Automatic Metrics

To assess whether our automatic metrics provide meaningful signals aligned with human judgment, we compute Pearson correlations between automated scores and expert annotations on a subset of 360 generated concepts. **Pla-S** exhibits a strong positive correlation with human-rated plausibility ($r = 0.60$, $p < 0.001$). While this correlation does not imply equivalence to human judgment, it supports the use of Pla-S as a scalable proxy for plausibility in large-scale analysis. In contrast, **SemDist** shows a weaker but statistically significant correlation with human novelty ratings ($r = 0.21$, $p < 0.001$). This result is expected, as human perceptions of novelty integrate domain relevance, conceptual coherence, and utility beyond distributional distance alone. Therefore, SemDist is used to characterize relative semantic dispersion and exploratory trends, rather than as an independent measure of human-perceived creativity, and serves to identify potential candidates. In our empirical analysis, automatic metrics serve as scalable proxies for preliminary assessment, while human expert evaluation is used for validation and interpretation.

Summary Region A is not a strategy-specific or model-specific artifact but a reproducible phenomenon that persists across generation strategies, domains, models, and embedding choices.

5.4 RQ3: Retrospective Concept Recovery

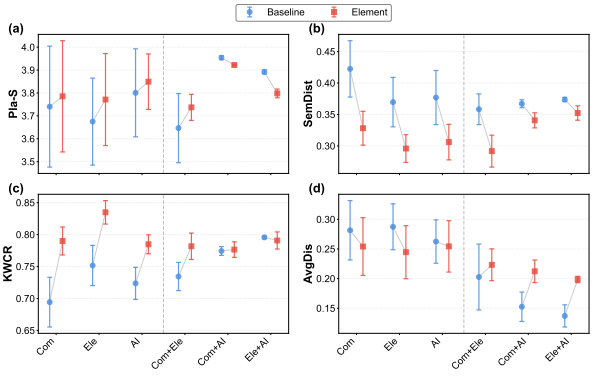


Figure 6: Comparative performance of Baseline (B) and Element (E) methods.

Experimental Setup Inspired by the combinatorial ideation paradigm proposed in recent work (Zhao et al., 2025), we conduct a retrospective concept recovery experiment to assess the proximity of LLM-generated concepts to real-world innovations. Cross-domain synthesis provides a controllable setting for this evaluation, because each target concept can be viewed as a verifiable "child" concept retrospectively derived from two cross-domain "parent" concepts. Focusing on the Communication+AI intersection, which exhibits the highest plausibility in human evaluation (Table 3), we collect 20 real composite concepts from recent literature published in 2024 or later, as listed in Table 7. Each target concept is decomposed into constituent elements, yielding two concept pools: $C_{comm} = \{c_1, c_2, \ldots, c_M\}$ (Communication concepts) and $C_{AI} = \{a_1, a_2, \ldots, a_N\}$ (AI concepts). We use GPT-3.5-Turbo-0125 (OpenAI, 2024a), whose knowledge cutoff is September 1, 2021, creating a clear temporal gap from the target concepts. For each cross-domain pair (c_i, a_j), the model generates a composite concept, yielding $M \times N = 400$ candidates. Among these, 20 candidates correspond to the original parent pairings underlying the 20 real target concepts. We use GPT-4o to compare these aligned candidates against their corresponding targets using full definition-level descriptions and to output a binary consistency decision (consistent or inconsistent). For the remaining generated concepts, we use GPT-4o to score their plausibility. Prompts are provided in Appendix B.

Main Results Figure 7 presents the recovery analysis. Among the 20 target concepts, 17 (85%) were successfully recovered, indicating that the model can independently derive realistic innovations through the systematic combination of do-

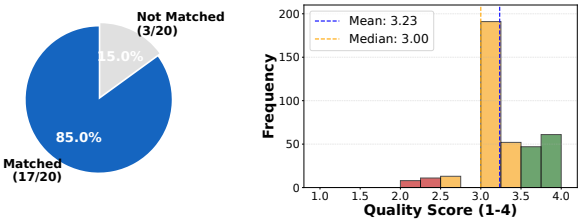


Figure 7: Evaluation results of composite concept generation. **Left:** Number of target concepts successfully recovered. **Right:** Plausibility score distribution of generated concepts that do not match any target concept.

main knowledge. The three unrecovered concepts matched at the name level but differed in their definitions, reflecting variations in technical details or application context. For the remaining 383 generated combinations, the average plausibility score reaches 3.23/4, indicating that many exhaustively generated combinations have a high probability of falling within Region A. This suggests that LLM-based concept creation extends beyond reproducing innovations found in the training corpus. Overall, these results suggest that in relatively simple combinatorial scenarios, LLM-based concept creation can approximate real scientific innovations and exhibit human-like creativity. The plausibility matrix is visualized as a heatmap in Figure 16.

5.5 Case Study

Figure 5 shows the representative cases. The Region A concept *Neural MIMO Networks* maps antenna arrays to neural network layers via functional analogy, preserving core structural logic while achieving genuine cross-domain novelty. Both components perform parallel processing on independent channels, making the analogy grounded and the resulting concept internally coherent. In contrast, the Region B concept *Quantum Entanglement Communication Codes (QECC)* attempts to bridge macroscopic coding theory and microscopic quantum mechanics, yet provides no coherent mechanism connecting the two scales. It chains loosely related terminology into a superficially novel narrative, but the causal links between iterative decoding and quantum entanglement remain unexplained and semantically fragmented. This contrast illustrates the core distinction: Region A outputs extend source knowledge through structurally grounded transformations, while Region B outputs assemble domain-spanning jargon without establishing meaningful conceptual connections. More cases are provided in Appendix A.7.3.

6 Conclusion

In this paper, we introduce a novelty-verifiability characterization of unverifiable hallucinations, distinguishing Region A (Creative Synthesis) from Region B (Groundless Fabrication). Across three domains, our results show that unverifiable hallucinations are not uniformly random noise: a non-negligible subset remains both novel and plausible, and this phenomenon is reproducible across different generation strategies and experimental conditions. Further, in a controlled retrospective recovery setting, LLM-generated concepts show meaningful proximity to real-world scientific concepts. These results suggest that hallucination in LLMs should be interpreted more selectively: beyond reliability failure, it may also contain a meaningful space of plausible conceptual extrapolation.

Limitations

This work has several limitations that point to important directions for future research.

Our evaluation of plausibility relies in part on LLM-based scoring mechanisms. However, these scores often vary only modestly across candidates, making it difficult to reliably distinguish plausible innovations from implausible ones. This suggests that current models may still struggle to capture fine-grained semantic feasibility, especially when evaluating subtle or compositional variations.

Our analysis of hallucination remains at a descriptive level, focusing on observable patterns in output divergence. We do not yet investigate the deeper internal mechanisms, such as activation pathways, attention dynamics, or representational drift, that may underlie these phenomena.

Our empirical investigation is conducted on a sampled dataset across three technical domains (Communication, Electromagnetics, and AI), which may limit the generalizability of the findings to other domains or knowledge structures. While our results show consistent patterns across these domains, the extent to which Region A generalizes to the humanities, social sciences, or other knowledge-intensive fields remains an open question.

Our retrospective concept recovery experiment adopts a controlled and verifiable setting, but this design may also simplify the discovery process and therefore does not fully test a model's ability to identify suitable source concepts independently. In addition, the recovery benchmark remains rela-

tively limited in scale, which constrains the strength of conclusions that can be drawn about broader scientific innovation patterns. Future work could extend this setting to open-ended source concept discovery before synthesis.

References

A. Aimi, G. Di Credico, H. Gimperlein, and C. Guardasoni. 2025. Adaptive time-domain boundary element methods for the wave equation with neumann boundary conditions. *Computers & Mathematics with Applications*, 198:196–213.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. 2022. Molgpt: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9):2064–2076.

Eranga Bandara, Safdar H. Bouk, Sachin Shetty, Sandip Roy, Ravi Mukkamala, Abdul Rahman, Peter Foytik, Xueping Liang, Ng Wee Keong, and Kasun De Zoysa. 2025. Llama-recipe — fine-tuned meta's llama llm, pbom and nft enabled 5g network-slice orchestration and end-to-end supply-chain verification platform. In *2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC)*, pages 1–6.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

SUSAN P. BESEMER and DONALD J. TREFFINGER. 1981. Analysis of creative products: Review and synthesis. *The Journal of Creative Behavior*, 15(3):158–178.

Thomas Bonnafont, Benjamin Chauvel, and Abdelmalek Toumi. 2024. A deep split-step wavelet model for the long-range propagation. In *2024 18th European Conference on Antennas and Propagation (EuCAP)*, pages 1–5.

Markus J. Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Preprint*, arXiv:2403.11996.

Tolga Buz, Benjamin Frost, Nikola Genchev, Moritz Schneider, Lucie-Aimée Kaffee, and Gerard de Melo. 2024. Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of Reddit's showerthoughts. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 291–307, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Haoye Chai, Yuan Yuan, and Yong Li. 2025. Mobiworld: World models for mobile wireless network. *Preprint*, arXiv:2507.09462.

Harrison Chase. 2022. LangChain.

Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. 2023. Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. *Preprint*, arXiv:2306.11648.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. 2024. Creativeval: Evaluating creativity of llm-based hardware code generation. In *2024 IEEE LLM Aided Design Workshop (LAD)*, pages 1–5. IEEE.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Dayu Fan, Rui Meng, Song Gao, and Xiaodong Xu. 2025. Kgrag-sc: Knowledge graph rag-assisted semantic communication. *Preprint*, arXiv:2509.04801.

Tao Feng, Yihang Sun, and Jiaxuan You. 2025. Grapheval: A lightweight graph-based llm framework for idea evaluation. *Preprint*, arXiv:2503.12600.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. 2026. Cache-to-cache: Direct semantic communication between large language models. *Preprint*, arXiv:2510.03215.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Yunlong Gu, Meng Xu, Jiguang Li, Qilei Li, Zhao Huang, Mengshan Li, Lixin Guan, and Mikko Valkama. 2025. Wikan: Lightweight kolmogorov–arnold networks for accurate indoor wifi localization. *Pervasive and Mobile Computing*, 114:102121.

Joy Paul Guilford. 1950. Creativity. *American psychologist*, 5(9):444.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5888–5899, New York, NY, USA. Association for Computing Machinery.

Jiahui Hu, Dan Wang, Zhibo Wang, Xiaoyi Pang, Huiyu Xu, Ju Ren, and Kui Ren. 2025. Federated large language model: Solutions, challenges and future directions. *IEEE Wireless Communications*, 32(4):82–89.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.06647.

Faranak Khosravi and Mehdi Shadaram. 2025. Retentive time series: A scalable machine learning model for traffic prediction in elastic optical networks. *IEEE Access*, 13:116569–116585.

Yongjun Kim, Jihong Park, Mehdi Bennis, and Junil Choi. 2025. Resilient llm-empowered semantic mac protocols via zero-shot adaptation and knowledge distillation. *Preprint*, arXiv:2505.21518.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

Katherine Van Koevering and Jon Kleinberg. 2024. How random is random? evaluating the randomness and humaness of llms' coin flips. *Preprint*, arXiv:2406.00092.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10).

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Guangming Liang, Mingjie Yang, Dongzhu Liu, Paul Henderson, and Lajos Hanzo. 2025. Environment-aware channel inference via cross-modal flow: From multimodal sensing to wireless channels. *Preprint*, arXiv:2512.04966.

Peng Liao, Xuyu Wang, Yingxin Shan, Lingling An, and Shiwen Mao. 2025. Wireless sensing in artificial intelligence of things: A general quantum machine learning framework. *IEEE Network*, 39(3):207–214.

Xuanyu Liu, Shijian Gao, Boxun Liu, Xiang Cheng, and Liuqing Yang. 2025. Wifo-cf: Wireless foundation model for csi feedback. *Preprint*, arXiv:2508.04068.

Yanzhen Liu, Zhijin Qin, Yongxu Zhu, and Geoffrey Ye Li. 2026. Enabling green wireless communications with neuromorphic continual learning. *Preprint*, arXiv:2502.17168.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

OpenAI. 2024a. GPT-3.5 Turbo [large language model].

OpenAI. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *Preprint*, arXiv:2311.05965.

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

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2026. Human-llm compound system for scientific ideation through facet recombination and novelty evaluation. *Preprint*, arXiv:2409.14634.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2026. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Preprint*, arXiv:2412.17596.

Huaguang Shi, Kaibo Jin, Xiaoquan Ren, Wei Li, and Yi Zhou. 2026. Channelmamba: A mamba-driven selective state-space model for channel prediction of high-mobility mimo in 6g iot. *IEEE Transactions on Wireless Communications*, 25:5291–5305.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. 2021. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):2403.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Zhihao Tao and Athina P. Petropulu. 2025. Meta-learning-driven gflownets for 3d directional modulation in mobile wireless systems. *Preprint*, arXiv:2511.06188.

Jingwen Tong, Zijian Li, Fang Liu, Wei Guo, and Jun Zhang. 2026. Wirelessagent++: Automated agentic workflow design and benchmarking for wireless networks. *Preprint*, arXiv:2603.00501.

Jingwen Tong, Jiawei Shao, Qiong Wu, Wei Guo, Zijian Li, Zehong Lin, and Jun Zhang. 2024. Wirelessagent: Large language model agents for intelligent wireless networks. *Preprint*, arXiv:2409.07964.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Haoyang Wang, Jiaxing He, and Lizhen Cui. 2025a. An advanced indoor localization method based on xlstm and residual multimodal fusion of uwb/imu data. *Electronics*, 14(13).

Xudong Wang, Jiacheng Wang, Lei Feng, Dusit Niyato, Ruichen Zhang, Jiawen Kang, Zehui Xiong, Hongyang Du, and Shiwen Mao. 2025b. Wireless hallucination in generative ai-enabled communications: Concepts, issues, and solutions. *Preprint*, arXiv:2503.06149.

Xuesong Wang, Mo Li, Xingyan Shi, Zhaoqian Liu, and Shenghao Yang. 2025c. Diffusion-aided task-oriented semantic communications with model inversion attack. *Preprint*, arXiv:2506.19886.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565, Bangkok, Thailand. Association for Computational Linguistics.

Yucheng Yu, Peng Chen, Xiao-Wei Zhu, Jianfeng Zhai, and Chao Yu. 2022. Continual learning digital predistortion of rf power amplifier for 6g ai-empowered wireless communication. *IEEE Transactions on Microwave Theory and Techniques*, 70(11):4916–4927.

Shuzhou Yuan, Zhan Qu, Ashish Yashwanth Kangen, and Michael Färber. 2025. Can hallucinations help? boosting llms for drug discovery. *Preprint*, arXiv:2501.13824.

Xinran Zhao, Boyuan Zheng, Chenglei Si, Haofei Yu, Ken Liu, Runlong Zhou, Ruochen Li, Tong Chen, Xiang Li, Yiming Zhang, and Tongshuang Wu. 2025. The ramon llull's thinking machine for automated ideation. *Preprint*, arXiv:2508.19200.

Fenghao Zhu, Xinquan Wang, Chongwen Huang, Richeng Jin, Qianqian Yang, Ahmed Al Hammadi, Zhaoyang Zhang, Chau Yuen, and Mérouane Debbah. 2024. Robust continuous-time beam tracking with liquid neural network. In *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, page 4878–4883. IEEE.

A Appendix

A.1 Dataset Construction and Concept Selection Criteria

To evaluate the proposed framework across diverse conceptual spaces, we constructed three distinct datasets, each comprising 150 concept entries from different technical domains: **Communication**, **Electromagnetics**, and **Artificial Intelligence**. These three domains were selected as illustrative examples due to their rich terminological ecosystems, clear conceptual hierarchies, and relevance to knowledge reasoning and generative innovation.

Concepts are selected according to the following three criteria:

(1) Technical Specificity Each term must refer to a concrete technical method, phenomenon, or computational mechanism, rather than to high-level system architectures (e.g., "cloud platform") or broad application scenarios (e.g., "smart city"). This constraint ensures that the terms are sufficiently grounded to allow meaningful semantic manipulation and generation.

(2) Intermediate Complexity We exclude elementary concepts (e.g., "neuron" or "signal") in favor of terms with moderate complexity—those that typically appear in mid-level academic discussions and have room for theoretical recombination or expansion. This ensures that the task challenges both the creativity and generalization capabilities of the model.

(3) Scope Coverage Within each domain, our datasets encompass multiple subfields and specialized areas, as illustrated in Figure 8. This granular categorization enables us to examine innovation across different levels of conceptual abstraction and disciplinary scopes. In addition, the selected terms are drawn from relatively mature subdomains to ensure conceptual coherence and enable more stable evaluation of generated concepts.

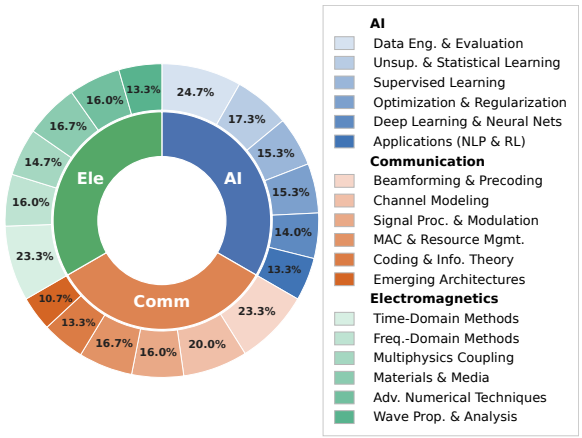


Figure 8: A hierarchical sunburst visualization of concept distributions across three domains. The outer ring shows the relative proportions of sub-domains within each domain.

A.2 Supplementary Analysis: Incremental Creativity Pattern

To understand LLMs' natural generative tendencies in conceptual creation, we conduct supplementary experiments using a minimal-intervention setting: models receive only the source concept name and definition, without structural constraints or intermediate reasoning steps (Section 4.1). We analyze 1,800 sampled outputs and classify them into three categories using an LLM-based classifier:

- **Rephrasing:** representing a simple restatement of the source text;
- **Incremental:** involving local modifications, subtle variations, or reconfigurations of existing elements;
- **Radical:** exhibiting significant semantic divergence or logical disconnection.

We observe a clear empirical pattern: LLMs predominantly generate new concepts via incremental semantic modifications rather than radical conceptual shifts. As illustrated in Figure 9, most generations fall into the incremental category, suggesting a preference for conservative recombination and local variation.

A.3 Temperature Effects

Temperature is a key parameter that influences the generated text, controlling the randomness of the model outputs (Koevering and Kleinberg, 2024). As an illustrative example within the communication domain, we employ GPT-4o to conduct tests in various temperature settings: {0.1, 0.3, 0.5, 0.7, 0.9}. As shown in Figure 10, we make the follow-

Metric	Method	Intra-domain									Cross-domain								
		Comm			Ele			AI			Comm+Ele			Comm+AI			Ele+AI		
		4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen
SemDist	B	0.082	0.075	0.061	0.075	0.063	0.059	0.086	0.068	0.063	0.072	0.064	0.064	0.073	0.074	0.072	0.074	0.073	0.075
	E	0.065	0.054	0.058	0.061	0.052	0.056	0.071	0.060	0.058	0.054	0.052	0.046	0.061	0.062	0.058	0.065	0.065	0.062
KWCR	B	0.903	0.907	0.921	0.901	0.899	0.913	0.893	0.906	0.909	0.918	0.919	0.925	0.923	0.920	0.924	0.908	0.907	0.911
	E	0.927	0.935	0.935	0.924	0.930	0.925	0.912	0.922	0.928	0.936	0.938	0.945	0.936	0.933	0.940	0.922	0.918	0.928
AvgDis	B	0.046	0.037	0.029	0.041	0.037	0.031	0.044	0.035	0.031	0.036	0.025	0.022	0.027	0.028	0.021	0.029	0.028	0.023
	E	0.044	0.032	0.029	0.042	0.032	0.031	0.046	0.036	0.033	0.036	0.032	0.030	0.035	0.034	0.030	0.036	0.035	0.033
Pla-S	B	3.439	3.852	3.931	3.463	3.730	3.831	3.579	3.900	3.922	3.473	3.752	3.714	3.955	3.945	3.962	3.895	3.898	3.882
	E	3.505	3.914	3.936	3.541	3.862	3.911	3.711	3.899	3.937	3.674	3.752	3.785	3.919	3.923	3.924	3.810	3.776	3.808

Table 4: (SciBERT) Transposed quantitative evaluation across domains and models. B: Prompt-Only Baseline, E: Structured Element-Based Innovation (highlighted with gray background). Models: 4o (GPT-4o), 4o-mini (GPT-4o-mini), Qwen (Qwen-max).

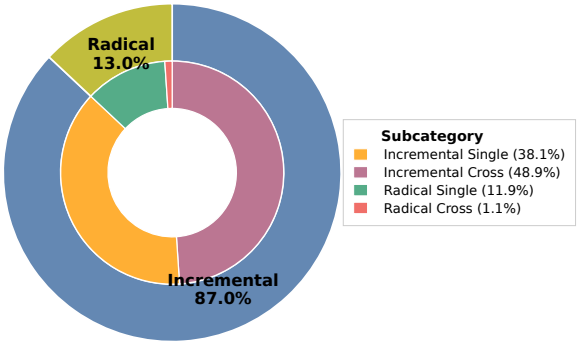


Figure 9: Distribution of concept classifications stratified by innovation category and domain scope.

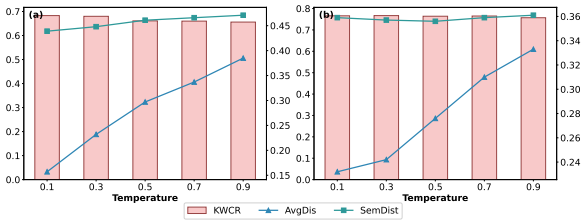


Figure 10: Average performance of generated concepts across temperature settings for the Prompt-Only Baseline (left) and the Structured Element-Based Innovation approach (right).

ing observations. First, regardless of the generation strategy, AvgDis increases noticeably as temperature rises, indicating higher diversity. In contrast, SemDist remains relatively stable across temperature values, with only a slight upward trend in the prompt-only baseline. KWCR also shows only minor fluctuations, suggesting that semantic lineage preservation is largely unaffected by temperature changes.

A.4 Robustness Check with SciBERT

To verify that our findings are not sensitive to the choice of embedding model, we replicate the main quantitative evaluation (Table 4) using SciBERT (Beltagy et al., 2019) embeddings. As shown in Table 4, overall trends remain similar when using SciBERT, suggesting that the main observations are relatively stable across embedding models.

A.5 Detailed Plausibility Results

Table 5 provides a detailed breakdown of forward reasoning (S_{fwd}), backward criticism (S_{bwd}), and their mean (Pla-S) across all experimental configurations. A notable asymmetry emerges: S_{fwd} scores are consistently high (>4.3), while S_{bwd} scores remain substantially lower (2.5–3.0), with a gap often exceeding 1.5 points. This suggests that LLMs exhibit overconfidence in forward concept extension but struggle with backward critical scrutiny. It therefore motivates the use of bidirectional evaluation for plausibility assessment.

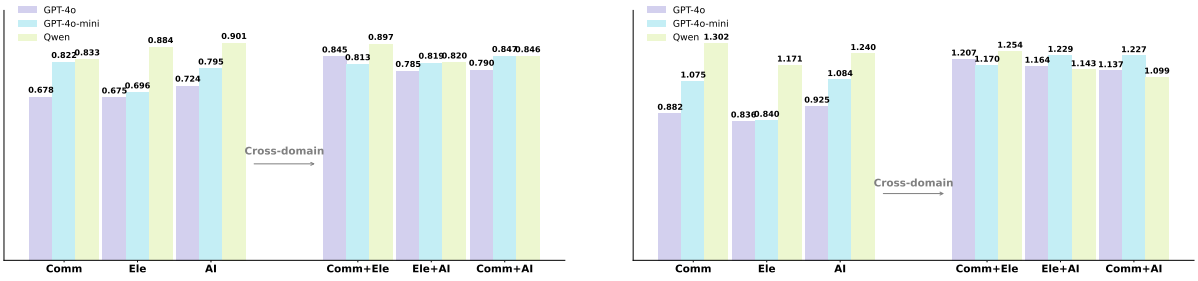
A.6 Additional Representation-Level Analyses

A.6.1 Semantic Diversity Analysis

Semantic Entropy As illustrated in Figure 11, in the single-domain setting, the AI domain exhibits the highest semantic entropy among all domains, suggesting a broader conceptual dispersion. Furthermore, when cross-domain signals are introduced, semantic entropy increases notably in the Communication and Electromagnetics domains. These findings suggest that cross-domain transfer expands the semantic space explored by the model, enabling more diverse conceptual variations.

Metric	Method	Intra-domain									Cross-domain								
		Comm			Ele			AI			Comm+Ele			Comm+AI			Ele+AI		
		4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen	4o	4o-mini	Qwen
S_{fwd}	B	4.351	4.756	4.896	4.427	4.678	4.781	4.545	4.861	4.883	4.340	4.616	4.636	4.929	4.914	4.934	4.865	4.850	4.825
	E	4.435	4.878	4.898	4.464	4.805	4.863	4.628	4.823	4.891	4.486	4.592	4.651	4.860	4.862	4.868	4.706	4.638	4.688
S_{bwd}	B	2.527	2.949	2.967	2.500	2.783	2.885	2.613	2.942	2.965	2.607	2.889	2.794	2.982	2.978	2.991	2.927	2.949	2.940
	E	2.576	2.955	2.979	2.619	2.924	2.963	2.796	2.980	2.988	2.864	2.914	2.923	2.980	2.988	2.983	2.917	2.916	2.932
Pla-S	B	3.439	3.852	3.931	3.463	3.730	3.831	3.579	3.900	3.922	3.473	3.752	3.714	3.955	3.945	3.962	3.895	3.898	3.882
	E	3.505	3.914	3.936	3.541	3.862	3.911	3.711	3.899	3.937	3.674	3.752	3.785	3.919	3.923	3.924	3.810	3.776	3.808

Table 5: Detailed plausibility evaluation across domains and models. B: Prompt-Only Baseline, E: Structured Element-Based Innovation (highlighted with gray background). Models: 4o (GPT-4o), 4o-mini (GPT-4o-mini), Qwen (Qwen-max).



(**a**) (**b**)

Figure 11: Semantic entropy comparison across Structured Element-Based Innovation strategy (**a**) and domain scopes (**b**).

A.6.2 Semantic Deviation Analysis

Domain	Euc.Mean	Euc.Var	Ang.Mean	Ang.Var
Comm	0.841	0.006	37.59	69.58
Comm+Ele	0.828	0.006	31.51	30.76
Comm+AI	0.789	0.007	36.21	35.97

Table 6: Distribution of semantic deviation across datasets. Euclidean distance captures the magnitude of deviation from the source concept embedding, while the angle metric reflects directional semantic divergence.

To further characterize the geometric properties of semantic shifts under different domain combinations, we compute both the Euclidean distance and the angular deviation. This analysis is restricted to **Structured Element-Based Innovation** outputs due to their higher generation density. For each source concept C_{src} with n generated concepts $\{C_{new}^{(1)}, \ldots, C_{new}^{(n)}\}$, we compute the Euclidean distance between each generated embedding and the source embedding:

$$d_i = \|\mathbf{e}_{C_{new}^{(i)}} - \mathbf{e}_{C_{src}}\|_2, \quad i = 1, \ldots, n \tag{3}$$

We then calculate the mean distance $\mu_d = \frac{1}{n}\sum_{i=1}^{n} d_i$ and variance $\sigma_d^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \mu_d)^2$ to quantify the magnitude and consistency of semantic displacement. To assess the directional consistency of semantic shift, we define a reference direction as the vector from the source embedding to the centroid of generated embeddings:

$$\mathbf{v}_{ref} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_{C_{new}^{(i)}} - \mathbf{e}_{C_{src}} \tag{4}$$

For each generated concept, we compute the angular deviation from this reference direction:

$$\theta_i = \arccos\left(\frac{(\mathbf{e}_{C_{new}^{(i)}} - \mathbf{e}_{C_{src}}) \cdot \mathbf{v}_{ref}}{\|\mathbf{e}_{C_{new}^{(i)}} - \mathbf{e}_{C_{src}}\|_2 \|\mathbf{v}_{ref}\|_2}\right) \tag{5}$$

We compute the mean angle μ_θ and variance σ_θ^2 for each source concept, and report dataset-level averages of these statistics. As shown in Table 6, single-domain generation (Comm) exhibits the largest mean Euclidean distance alongside substantially higher angular variance, reflecting a more exploratory yet scattered semantic trajectory. In contrast, cross-domain configurations (Comm+Ele, Comm+AI) exhibit markedly reduced angular variance while maintaining comparable displacement magnitudes. This pattern indicates that the current

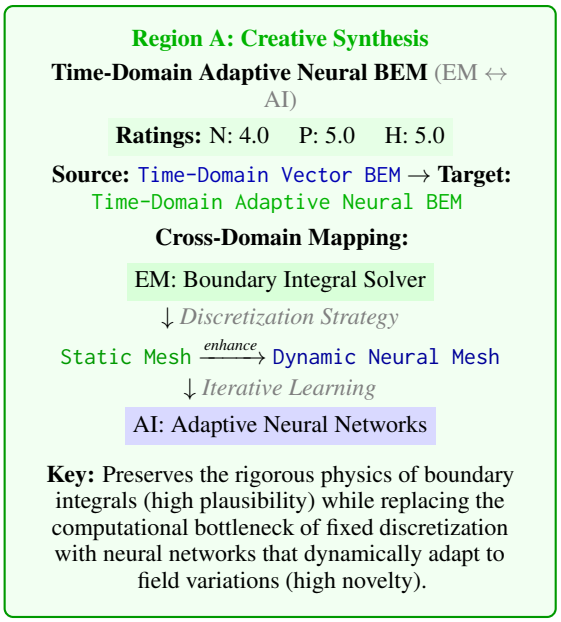


Figure 12: Region A example: Time-Domain Adaptive Neural BEM demonstrates constraint-respecting innovation by optimizing physical solvers with AI.

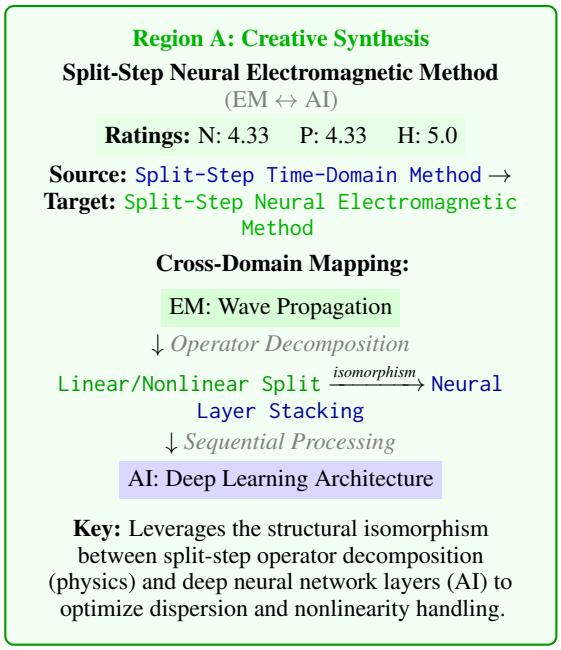


Figure 13: Region A example: Split-Step Neural Electromagnetic Method demonstrates structural alignment between physical operators and neural layers.

structured element-based generation strategy imposes stronger structural constraints during cross-domain transformations, leading to a more coherent directional shift in semantic space.

A.7 Case study

A.7.1 Region A

Region A represents outputs characterized by creative synthesis, where novel concepts maintain strong coherence with established knowledge while extending beyond current verification boundaries. As illustrated in Figures 12, 13, and 14, the examples showcase how Region A outputs bridge domains while maintaining internal consistency and technical plausibility.

A.7.2 Region B

Region B encompasses outputs that exhibit groundless fabrication, and lack coherent grounding in established knowledge frameworks. As shown in Figure 15, these outputs typically involve invalid logical leaps or scale mismatches. They bridge concepts from incompatible domains without establishing a coherent mechanism, resulting in semantically disconnected outputs despite superficial structural novelty.

A.7.3 Technical Plausibility and Literature Alignment

To explain why Region A concepts are technically plausible, we conduct a focused case-based analysis grounded in recent literature. Because these generated concepts are novel by construction, our goal is not to identify identical prior proposals, but to verify whether these concepts align with established problem formulations and recognized innovation trajectories.

Case 1: Neural Operator Substitution. The generated concept *Time-Domain Adaptive Neural BEM* extends classical boundary element methods by introducing neural surrogates for computationally expensive components. Recent work by Aimi et al. (2025) establishes adaptive time-domain BEM as a frontier approach for wave propagation, while explicitly highlighting the high cost of residual-based error estimation. Our concept follows a well-established AI-acceleration paradigm by replacing this bottleneck with a learned operator, representing a plausible efficiency-oriented extrapolation rather than a departure from the underlying physics.

Case 2: Proactive Robustness Strategy. The concept of *Adversarial Predistortion in Microwave RF Systems* addresses the same core challenge iden-

Region A: Creative Synthesis

Adversarial Predistortion in RF Systems (EM ↔ AI)

Ratings: N: 4.67 P: 4.33 H: 5.0

Source: `Microwave RF Linearization` →
Target: `Adversarial Predistortion`

Cross-Domain Mapping:

EM: Signal Linearization

↓ *Robustness Strategy*

`Standard Predistortion` $\xrightarrow{transfer}$ `Adversarial Scenarios`

↓ *Distortion Simulation*

AI: Adversarial Training (GANs)

Key: Applies the AI concept of "adversarial training" to RF systems, creating proactive distortion scenarios to train the predistorter, significantly enhancing robustness against non-linearities.

Figure 14: Region A example: Adversarial Predistortion demonstrates mechanism transfer from AI robustness to RF linearity.

Region B: Groundless Fabrication

Quantum Entanglement Enhanced Scattering (EM ↔ Quantum)

Ratings: N: 4.33 P: 1.33 H: 2.0

Source: `Multi-Body Coupled Scattering` →
Target: `Entanglement Enhanced Scattering`

Cross-Domain Mapping:

EM: Macroscopic Scattering

↓ *Scale Mismatch / Logic Gap*

`Mutual Coupling` $\xrightarrow{invalid\ jump}$ `Quantum Entanglement`

↓ *Violates No-Communication Thm.*

Quantum: Instantaneous Info Exchange

Key: A "semantic leap" that conflates macroscopic wave coupling with microscopic quantum entanglement. Violates the No-Communication Theorem by claiming instantaneous information transfer for adaptive control.

Figure 15: Region B example: Demonstrates a "Geometric Orphan" where high-level jargon is combined without respecting physical scale or causality.

tified in 6G RF systems—unpredictable and time-varying hardware non-linearities. While prior studies emphasize reactive adaptation through continual learning (Yu et al., 2022), our concept introduces a complementary proactive strategy via adversarial training. This shift reframes robustness as an intrinsic system property, aligning with emerging work on adversarially robust RF learning rather than contradicting existing approaches.

Case 3: Architectural Isomorphism. The *Split-Step Neural Electromagnetic Method* exploits the structural correspondence between split-step solvers for parabolic wave equations and layered neural architectures. This mapping has been independently validated by Bonnafont et al. (2024), who demonstrate that physical operator splitting can directly inform neural network design. Such model-driven alignment ensures that the generated concept preserves physical constraints while enabling learning-based acceleration.

Overall, these cases indicate that Region A concepts arise from constraint-preserving scientific extrapolation. Rather than arbitrary hallucinations, they extend established methods along recognized computational and physical dimensions.

B Prompts

Concept Decomposition

```
You are an expert in {domain}. Analyze the given concept definition and extract its core elements. Then, evaluate the innovation potential of each element on a scale of 1 to 5 (higher means more potential for modification and innovation) without any explanation.
Concept Name:
{concept_name}
Concept definition:
{concept_definition}
```

Intra-Domain Element Modification

```
You are an expert in innovative {domain}. Your task is to creatively transform a specific technical element related to a {domain} concept.
Your task is to:
1.Analyze the element's role within its technical and functional
```

context.
2.Creatively redesign or recombine it to produce a novel variant with distinct utility.
3.Ensure the innovation reflects a substantive transformation, not just superficial modification.
Input Context:
Concept Name: {concept_name}
Concept definition: {concept_definition}
Element Name: {element_name}

Intra-Domain Concept Reconstruction

You are an expert in the field of {domain}. Given an original concept and a distinctly innovated core element, your task is to revise the original concept by substituting the old element with this new, innovative one.
Original Concept definition: {concept_definition}
Original Element to be replaced: {raw_element} New Element Information:
-New Element: {json.dumps(new_element, indent=2)}
Make sure that:
-The newly revised concept is logically consistent and technically robust.
-The innovative element is seamlessly and meaningfully integrated, bringing novel and substantial value to the concept.

Intra-Domain Prompt-Only Baseline

You are an expert in {domain}. Your task is to creatively evolve the following concept within the {domain} domain. Rather than merely modifying or extending existing knowledge, your goal is to generate a novel concept by rethinking, restructuring, or recombining core elements in an original way.
The newly derived concept should:
-Represent a meaningful reconfiguration of existing principles within {domain}.
-Offer a clear advancement, reinterpretation, or novel direction compared to the original.
-Be technically rigorous, logically coherent, and precisely articulated.
Original Concept Name: {concept_name}
Original Concept definition: {concept_definition}

Cross-Domain Element Modification

You are an expert in both {domain1} and {domain2}. Your task is to analyze the core elements of a given concept from {domain1}, select the most appropriate analogous or complementary concept from {domain2}, and establish a formal mapping relationship between these two concepts.
For each core element, please:
1.Identify a corresponding element or principle from {domain2} that exhibits functional, structural, or theoretical alignment.
2.Justify this mapping by referencing foundational principles, technical mechanisms, or emergent synergies between the domains. **Original Concept**: {concept_name}
Concept definition: {concept_definition}
Elements Name: {element_name}

Cross-Domain Concept Reconstruction

You are an expert in both the fields of {domain1} and {domain2}. Your task is to incorporate this transformation into the original {domain1} concept based on the mapping results, so as to reconstruct a brand-new concept.

This new concept should possess the following characteristics:
-Fuse the principles from both the {domain1} field and the {domain2} field in a meaningful way.
-Present a concept that is technically sound and innovative.
Concept definition: {concept_definition}
Raw element: {raw_element}
Mapped Electromagnetic Element: {json.dumps(mapped_data, indent=2)}

Cross-Domain Prompt-Only Baseline

You are an expert in both {domain1} and {domain2}. Your task is to creatively generate a novel cross-domain concept by meaningfully integrating the following concept from {domain1} with relevant principles, methodologies, or paradigms from {domain2}. The fusion should go beyond superficial combination —it must reflect a deep, innovative integration that leverages the strengths and insights of both fields.
The new concept should:
-Combine core ideas, models, or approaches from both {domain1} and {domain2} in a technically coherent and creative way.
-Offer improvements or advancements in the context of both fields.
-Be technically rigorous and innovative.
Original Concept Name: {concept_name}
Original Concept definition: {concept_definition}

Plausibility Forward Reasoning

You are a Principal Engineer in the field of {domain}. The original term is {original_term}, and a newly proposed concept is {generated_term}.
Please reason in detail how this new concept could have evolved from the original one, considering technical background, principles, and potential innovations in the field of domain} technologies.
Finally, rate the plausibility of this reasoning on a scale from 1 to 5, and briefly explain your reasoning. Respond in JSON with keys "score" (int 1–5) and "rationale" (string).
Example(JSON):
{{
"score": 4,
"rationale": "Because ... "
}}

Plausibility Backward Criticism

You are a Principal Engineer in the field of {domain}. Please critically evaluate the feasibility of the new concept {generated_term}. From a technical, resource, or engineering standpoint, what challenges or limitations might arise in implementing this concept? Identify potential issues and rate its overall feasibility on a scale from 1 to 5. The more severe or numerous the issues, the lower the feasibility score. Briefly explain your reasoning.
Respond in JSON with keys "score" (1–5) and "rationale" (string).
Example(JSON):
{{
"score": 2,
"rationale": "Potential issues are ... "
}}

Concept Classification Prompt

You are an expert taxonomy analyst for scientific concepts. Your task is to classify the relationship between a Source Concept and a

Generated Concept into one of the following three categories.
Input Data:
- Source Concept: {source_name} - {source_def}
- Generated Concept: {gen_name} - {gen_def}
Categories:
1.[Rephrasing]: The generated concept is merely a paraphrase or a trivial retrieval of the source. No new semantic information is added.
2.[Incremental]: The generated concept extends the source with logical but predictable modifications (e.g., adding a common attribute, combining with a closely related concept). It represents a "safe" step.
3.[Radical]: The generated concept deviates significantly from the source. It introduces entirely new paradigms, OR it is logically disconnected/nonsensical.
Output JSON format only, no additional content:
{
"category": "Rephrasing" or "Incremental" or "Radical",
"explanation": "1-2 sentence explanation"
}

Composite Concept Generation Prompt (RQ3)

You are an expert in wireless communication and artificial intelligence. Your task is to naturally combine two concepts into an innovative composite concept.
Communication Concept:
{com_concept}
AI Concept:
{ai_concept}
Please create a novel composite concept that naturally integrates these two concepts. The combination should be technically meaningful and innovative.
Provide your answer in the following JSON format:
{{
"concept_name":
"A concise and catchy name for the composite concept",
"definition":
"A clear 1-2 sentence definition explaining what this composite concept is and how it works"
}}
Requirements:
1.The concept name should be creative and reflect both domains
2.Keep the definition concise (1-2 sentences)
Output only the JSON, no additional explanation.

Definition Semantic Match Prompt (RQ3)

You are an expert in wireless communication and AI. Your task is to determine if two concept definitions describe essentially the same composite concept.
Baseline Concept Definition:
{baseline_def}
Generated Concept Definition:
{generated_def}
Please determine if these two definitions describe the same or highly similar composite concept.
Consider:
1.Core functionality and purpose
2.Technical approach and methods
3.Application domain and use cases
Answer with ONLY a JSON object containing a single boolean field:
{{
"is_match": true or false
}}
Output only the JSON, no additional text.

Quality Evaluation Prompt (RQ3)

```
You are an expert in wireless communication and AI. Your task is to evaluate the quality and plausibility of a composite concept.
Communication Concept:
{com_concept}
AI Concept:
{ai_concept}
Composite Concept Name:
{concept_name}
Definition:
{definition}
Please evaluate the plausibility and logical coherence of this composite concept on a scale of 1-4:
-4:Good plausibility, reasonable integration, some technical value
-3:Moderately plausible, acceptable integration, limited novelty
-2:Weak plausibility, forced integration, unclear benefit
-1:Not plausible, illogical integration, no clear value
Important: The maximum score is 4. Do not use 5.
Answer with a JSON object: {{
"score": 1 to 4 (integer),
"reasoning": "Brief explanation"
}}
Output only the JSON, no additional text.
```

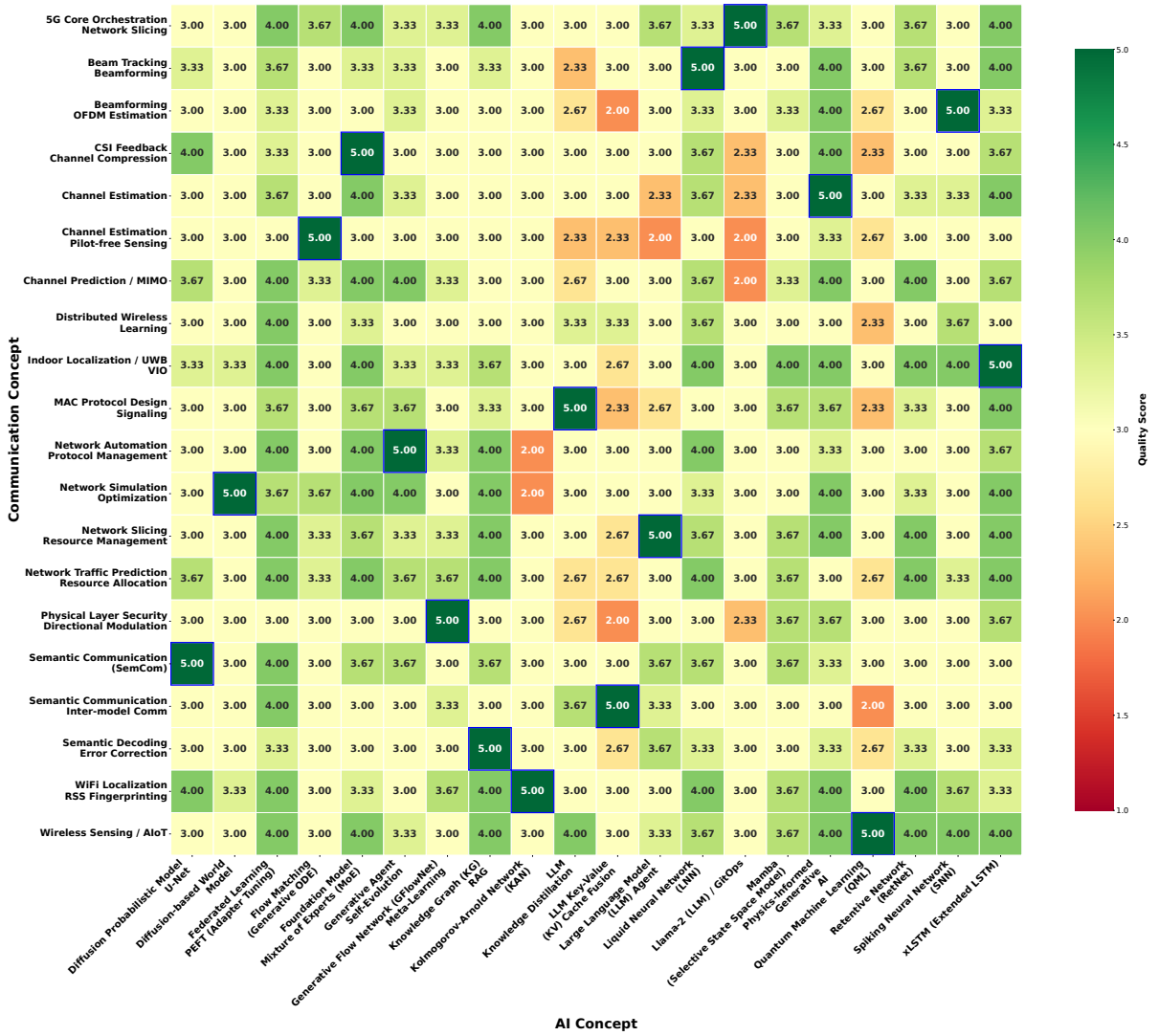


Figure 16: Heatmap visualization of quality scores for all 400 composite concepts. The 20×20 matrix represents the plausibility evaluation results, where each cell corresponds to a composite concept formed by combining one communication concept with one AI concept.

No.	Composite Concept Definition	Source Paper
1	**MobiWorld**: A generative world model for simulating and optimizing mobile wireless networks via counterfactual reasoning.	*MobiWorld: World Models for Mobile Wireless Network(Chai et al., 2025)*
2	**WirelessAgent**: An autonomous agent framework using LLMs for closed-loop wireless network management tasks like slicing.	*WirelessAgent: Large Language Model Agents for Intelligent Wireless Networks(Tong et al., 2024)*
3	**WiFo-CF**: A wireless foundation model designed to compress and reconstruct heterogeneous CSI feedback across varying configurations.	*WiFo-CF: Wireless Foundation Model for CSI Feedback(Liu et al., 2025)*
4	**DiffSem**: A semantic communication framework using diffusion models for noise-resilient semantic transmission and reconstruction.	*Diffusion-aided Task-oriented Semantic Communications with Model Inversion Attack(Wang et al., 2025c)*
5	**ChannelMamba**: An end-to-end channel prediction model utilizing the Mamba architecture for linear-complexity temporal modeling.	*ChannelMamba: A Mamba-Driven Selective State-Space Model for Channel Prediction of High-Mobility MIMO in 6G IoT(Shi et al., 2026)*
6	**FM-Channel Estimator**: A pilot-free channel inference framework using Flow Matching to map multimodal sensing data to channel matrices.	*Environment-Aware Channel Inference via Cross-Modal Flow: From Multimodal Sensing to Wireless Channels(Liang et al., 2025)*
7	**C2C-SemCom**: A direct communication paradigm where LLMs exchange KV-Cache states instead of tokenized text.	*Cache-to-Cache: Direct Semantic Communication Between Large Language Models(Fu et al., 2026)*
8	**KGRAG-SC**: A semantic communication scheme using Retrieval-Augmented Generation with Knowledge Graphs for error correction.	*KGRAG-SC: Knowledge Graph RAG-Assisted Semantic Communication(Fan et al., 2025)*
9	**SpikACom**: A neuromorphic framework using Spiking Neural Networks for power-efficient beamforming and channel estimation.	*SpikACom: A Neuromorphic Computing Framework for Green Communications(Liu et al., 2026)*
10	**T3NPM**: A hybrid MAC protocol framework combining token-based LLM adaptation with lightweight neural distillation.	*Resilient LLM-Empowered Semantic MAC Protocols via Zero-Shot Adaptation and Knowledge Distillation(Kim et al., 2025)*
11	**Meta-GFlowNet**: A security framework using meta-learning and GFlowNets for rapid adaptation of secure beamforming in mobile scenarios.	*Meta-Learning-Driven GFlowNets for 3D Directional Modulation in Mobile Wireless Systems(Tao and Petropulu, 2025)*
12	**LNN-Beam Tracking**: A robust beam tracking method using Liquid Neural Networks to handle noisy, continuous-time channel dynamics.	*Robust Continuous-Time Beam Tracking with Liquid Neural Network(Zhu et al., 2024)*
13	**WiKAN**: An indoor localization model replacing MLPs with Kolmogorov-Arnold Networks to capture non-linear RSS-distance mappings.	*WiKAN: Lightweight Kolmogorov-Arnold Networks for Accurate Indoor WiFi Localization(Gu et al., 2025)*
14	**WirelessAgent++**: A self-evolving generative agent framework that designs and refines task-oriented workflows for specialized wireless tasks.	*WirelessAgent++: Automated Agentic Workflow Design and Benchmarking for Wireless Networks(Tong et al., 2026)*
15	**Fed-PEFT**: A federated parameter-efficient fine-tuning paradigm for adapting large models under distributed learning settings.	*Federated Large Language Model: Solutions, Challenges and Future Directions(Hu et al., 2025)*
16	**xLSTM-IMU-UWB**: An indoor localization framework using xLSTM for temporal modeling and multimodal fusion of UWB and IMU data.	*An Advanced Indoor Localization Method Based on xLSTM and Residual Multimodal Fusion of UWB/IMU Data(Wang et al., 2025a)*
17	**Llama-Recipe**: A Llama-based platform for 5G network-slice orchestration and cloud-native service deployment.	*Llama-Recipe — Fine-Tuned Meta's Llama LLM, PBOM and NFT Enabled 5G Network-Slice Orchestration and End-to-End Supply-Chain Verification Platform(Bandara et al., 2025)*
18	**Ret-TS**: A traffic prediction model utilizing Retentive Networks (RetNet) to optimize Elastic Optical Network (EON) resources.	*Retentive Time Series: A Scalable Machine Learning Model for Traffic Prediction in Elastic Optical Networks(Khosravi and Shadaram, 2025)*
19	**QML-Wireless Sensing**: A framework utilizing quantum machine learning models to process wireless signals for sensing applications in AIoT.	*Wireless Sensing in Artificial Intelligence of Things: A General Quantum Machine Learning Framework(Liao et al., 2025)*
20	**Physics-GenAI**: A physics-grounded generative AI framework for mitigating hallucination in wireless communication tasks.	*Wireless Hallucination in Generative AI-enabled Communications: Concepts, Issues, and Solutions(Wang et al., 2025b)*

Table 7: Source papers used for dataset construction. The table lists the literature from which the dataset was built. The composite concepts are standardized concept formulations distilled from these papers for dataset construction and analysis; their names and definitions are derived from the source literature through abstraction and standardization rather than reproduced verbatim.