Navigating Ideation Space: Decomposed Conceptual Representations for Positioning Scientific Ideas

Anonymous ACL submission

Abstract

Scientific discovery is a cumulative process and requires new ideas to be situated within an ever-expanding landscape of existing knowledge. An emerging and critical challenge is how to identify conceptually relevant prior work from rapidly growing literature, and assess how a new idea differentiates from existing research. Current embedding approaches typically conflate distinct conceptual aspects into single representations and cannot support fine-grained literature retrieval; meanwhile, LLM-based evaluators are subject to sycophancy biases, failing to provide discriminative novelty assessment. To tackle these challenges, we introduce the **Ideation Space**, a structured representation that decomposes scientific knowledge into three distinct dimensions, i.e., *research problem, methodology*, and *core findings*, each learned through contrastive training. This framework enables principled measurement of conceptual distance between ideas, and modeling of ideation transitions that capture the logical connections within a proposed idea. Building upon this representation, we propose a **Hierarchical Sub-Space Retrieval** framework for efficient, targeted literature retrieval, and a **Decomposed Novelty Assessment** algorithm that identifies which aspects of an idea are novel. Extensive experiments demonstrate substantial improvements, where our approach achieves Recall@30 of 0.329 (16.7% over baselines), our ideation transition retrieval reaches Hit Rate@30 of 0.643, and novelty assessment attains 0.37 correlation with expert judgments. In summary, our work provides a promising paradigm for future research on accelerating and evaluating scientific discovery.

1 Introduction

Scientific discovery (Wang et al., 2023; Lu et al., 2024; Wang et al., 2024; Si et al., 2025) is an inherently cumulative endeavor, where each new idea needs to be situated within an ever-expanding landscape of existing knowledge. The pace of this expansion is astonishing, particularly in the field of computer science and AI. For instance, arXiv alone publishes over 20,000 papers monthly[1] in 2025 compared with around 6,000 monthly papers in 2010, and the emerging new literature may bear direct relevance to a specific research direction. When developing a new research idea, a critical challenge emerges: *how to identify prior work that is conceptually relevant from a massive amount of literature, and measure how well the idea differentiates the existing works*?

Existing approaches to scientific literature retrieval and assessment suffer from three fundamental limitations. **First**, current embedding-based methods encode scientific papers into monolithic representations (Cohan et al., 2020; Singh et al., 2023). While effective for topical matching, such representations conflate distinct conceptual aspects, e.g., a paper's problem formulation, its methodological approach, and its empirical findings are collapsed into a single vector, obscuring which aspect should drive the similarity used in retrieval. **Second**, while recent work has turned to large language models (LLMs) as evaluators for scientific hypotheses, LLM-based assessment often suffers from sycophancy biases (Sharma et al., 2024; Ye et al., 2025). The models tend toward positive or intermediate ratings regardless of actual novelty, and consequently produce scores that lack discriminative power and actionable insight. **Third**, existing methods usually do not evaluate which part of this idea is conceptually novel. A new idea may introduce a genuinely new research question while employing standard methodology, or apply an established technique to an unprecedented domain. Without distinguishing these orthogonal aspects, researchers will gain little guidance on how to strengthen their work.

[1] https://arxiv.org/stats/monthly_submissions

To address these challenges, we introduce the **Ideation Space**, a structured representation that enables principled conceptual similarity measurement across decomposed dimensions of scientific contribution (§2). Our key motivation is that scientific knowledge can be decomposed into conceptually orthogonal dimensions, each capturing a distinct aspect of research contribution. Rather than representing papers as a single embedding, we construct three specialized sub-spaces, i.e., the *problem* being addressed, the *method* proposed, and the core *findings* expected to discover, where each is learned through contrastive training to leverage the conceptual relationships in the citation network for each aspect. Based on the notion of ideation space, we utilize the three sub-space encoders to measure the *conceptual distance* among scientific ideas. When the encoded embeddings of two ideas are closer in the ideation space, it indicates that these ideas are more conceptually similar. In addition, we propose to model the *ideation transition* by computing the vector across the sub-spaces to capture the logical connections within a proposed idea, such as how the methodology is derived from the research problem, and what expected findings would be discovered from the method.

Empowered by our ideation space, we propose a **Hierarchical Sub-Space Retrieval** framework that organizes literature into a structured database supporting efficient, targeted retrieval of conceptually relevant prior work for each dimension (§3.1). Building upon retrieval, we propose a graph-based algorithm for **Decomposed Novelty Assessment**. Since an idea's novelty is inversely related to how closely it resembles retrieved prior work, we measure novelty by quantifying differentiation from retrieved evidence along each conceptual dimension, identifying which specific aspects of an idea are novel (§3.2). Through extensive experiments, we demonstrate substantial improvements over existing methods. For retrieving conceptually similar literature, our approach achieves a Recall@30 of 0.329, outperforming the baseline by 16.7%. Our ideation transition method further enables retrieval of relevant reasoning patterns, achieving Hit Rate@30 of 0.643. In addition, our novelty assessment method shows a promising correlation of 0.37 with human judgments, while providing grounded evidence for evaluation decisions. Our contributions are summarized as follows:

- We introduce the concept of *Ideation Space* and propose a framework that learns decomposed representations of scientific knowledge through citation-contextualized contrastive learning, enabling the improved measurement of conceptual distance and ideation transition.
- We propose a large-scale hierarchical sub-space retrieval framework that organizes scientific literature by conceptual function, which enables efficient and targeted retrieval with up to 16.7% improvement in recall over strong baselines.
- We develop a decomposed novelty evaluation pipeline that provides fine-grained assessments identifying where and why an idea is novel, and achieve promising correlation with expert judgment comparable to frontier LLMs.

2 Ideation Space

We introduce **Ideation Space**, a structured framework for representing scientific knowledge that enables principled measurement of conceptual similarity across distinct dimensions of research contribution. Our key insight is that scientific ideas are not monolithic; instead, they comprise separable components serving different intellectual functions.

Our framework consists of three major steps. **First**, we partition scientific knowledge into three functionally distinct sub-spaces, i.e., *problem*, *method*, and *findings*, each capturing a different aspect of contribution (§2.1). **Second**, we leverage the citation network to construct training data reflecting how researchers perceive conceptual relationships, using citation context to determine which dimension connects two papers (§2.2). **Third**, we train specialized encoders via contrastive learning to produce embeddings where proximity reflects genuine conceptual similarity within each dimension (§2.3). The resulting ideation space supports both fine-grained similarity measurement and modeling of *ideation transitions*, i.e., the logical connections within a proposed idea (§2.4).

2.1 Sub-Space Decomposition

We partition the collection of scientific literature $\mathcal{L}$ into three functionally distinct sub-spaces based on the conceptual function:

$$\mathcal{L} = S_{\text{problem}} \cup S_{\text{method}} \cup S_{\text{findings}} \quad (1)$$

where S_{problem} captures research problems and motivations, S_{method} represents methodological approaches, and S_{findings} indicates the expected results and conclusions. This decomposition can

provide more comprehensive representations of scientific knowledge, such as a shared problem definition, a relevant methodology, or comparisons against reported findings. Each paper is embedded into these three spaces by specialized encoders E_k ($k \in \{\text{problem}, \text{method}, \text{findings}\}$), enabling fine-grained conceptual distance measurement.

2.2 Training Data Construction

Building effective sub-space encoders requires training data reflecting genuine conceptual relationships within each dimension. We construct such data through a two-stage pipeline: identifying structurally related paper pairs from the citation graph, then classifying their functional relationships to assign them to appropriate sub-spaces. Our dataset comprises 259,340 papers from arXiv spanning 1993 to 2024, covering four major AI domains (cs.AI, cs.LG, cs.CV, and cs.CL), which ensures broad generalization.

Candidate Pair Identification. We extract candidate positive pairs based on structural proximity within a 5-year publication window. Two papers (p, p^+) form a candidate pair if they have a direct citation link and satisfy at least one structural criterion: *bibliographic coupling* (citing $\geq$10 common references) or *co-citation* (co-cited $\geq$10 times). These thresholds ensure substantive intellectual connections, as papers meeting them typically engage with overlapping research questions or methodological traditions.

Function Classification. The key challenge is determining *which* conceptual dimension connects a pair of related papers. We employ an LLM-based zero-shot classifier that analyzes structured information about both papers (titles, research questions, methods, key findings) and produces confidence scores for three categories: `[Research Problem]`, `[Method Approach]`, and `[Key Findings]`. Pairs may be assigned to multiple sub-spaces when their relationship spans dimensions; those fitting no category are labeled `[Irrelevant]` and excluded. This distributes candidate pairs into three sub-graphs G_{problem}, G_{method}, and G_{findings}, each encoding dimension-specific conceptual relationships. More details of function classification are included in Appendix A.2.

Negative Sampling. Effective contrastive learning requires carefully constructed negatives. We employ two complementary strategies. **In-batch negatives** sample from papers in the same batch that lack citation, co-citation, or bibliographic coupling relationships with the anchor, ensuring structural independence. **Hard conceptual negatives** exploit cross-dimensional relationships: we select pairs where the classifier assigns high scores to one category but low scores to another. For example, papers with high `[Research Problem]` but low `[Method Approach]` scores serve as hard negatives for E_{method}, forcing encoders to distinguish papers highly related in one dimension but less relevant in another.

2.3 Contrastive Encoder Training

We train three specialized encoders (E_k, $k \in \{\text{problem}, \text{method}, \text{findings}\}$) independently on their corresponding sub-graphs. Each encoder initializes from SPECTER2 (Singh et al., 2023) to inherit general scientific knowledge, then fine-tunes to specialize in its designated dimension.

To ensure each encoder focuses on dimension-relevant content, we use an LLM to extract sub-space-specific text, i.e., research questions for S_{problem}, methodological descriptions for S_{method}, and key findings for S_{findings}. Each encoder trains exclusively on its corresponding text type. For an anchor paper p with positive set P^+ and negative set P^-, encoder E_k minimizes:

$$\mathcal{L}_k = -\frac{1}{|P^+|} \sum_{p^+ \in P^+} \log \frac{e^{s(p,p^+)/\tau}}{\sum_{x \in P^+ \cup P^-} e^{\ell(p,x)}} \quad (2)$$

where $s(\cdot,\cdot)$ denotes cosine similarity, τ is temperature, and:

$$\ell(p,x) = \begin{cases} s(p,x)/\tau, & x \in P^+, \\ s(p,x)/\tau + \log w(x), & x \in P^-. \end{cases} \quad (3)$$

The weighting function $w(x) = ((s(p,x)+1)/2)^\gamma$ emphasizes hard negatives by increasing their contribution to normalization.

2.4 Conceptual Similarity and Ideation Transition

The trained encoders enable two complementary forms of analysis. For **conceptual similarity**, given a new idea i and prior work p, we measure their similarity in sub-space k as $\text{sim}_k(i,p) = \cos(E_k(i), E_k(p))$. Higher similarity indicates greater conceptual overlap along dimension k, suggesting reduced novelty in that aspect.

Beyond static similarity, we capture the **ideation transition**, i.e., the reasoning structure within a paper, through transition vectors:

$$T_{\text{PM}}(p) = E_{\text{method}}(p) - E_{\text{problem}}(p) \quad (4)$$

$$T_{\text{MF}}(p) = E_{\text{findings}}(p) - E_{\text{method}}(p) \quad (5)$$

The problem-to-method transition T_{PM} captures how methodology addresses the research problem, while method-to-findings T_{MF} represents how methodology leads to conclusions. These vectors enable the retrieval of papers with analogous reasoning patterns across different topics.

3 Hierarchical Retrieval and Novelty Assessment

Having established the ideation space as a structured representation of scientific knowledge, we now demonstrate its utility through two applications: retrieving conceptually relevant prior work and evaluating ideation novelty. These tasks are intrinsically interconnected: an idea's novelty is inversely related to how closely retrieved prior work resembles it. The more relevant the retrieved work, the less novel the idea. By leveraging the decomposed ideation space, we first retrieve prior work that is genuinely relevant along each conceptual dimension (§3.1), then assess novelty by measuring how much a new idea differentiates from the retrieved evidence (§3.2). This retrieval-grounded approach ensures that novelty scores reflect conceptual distance from the most relevant existing work, rather than surface-level textual similarity.

3.1 Hierarchical Sub-Space Retrieval

Traditional retrieval treats papers as atomic units, returning results based on overall topical similarity. However, researchers often seek prior work relevant to *specific aspects*, e.g., papers addressing similar problems, employing related methods, or reporting comparable findings. Our framework addresses this need through structured databases supporting targeted, dimension-specific queries.

Database Construction. We construct five specialized databases from our corpus of 259,340 arXiv papers. Three **node databases** store sub-space-specific embeddings: for each paper, we extract dimension-specific text and encode it using the corresponding encoder (E_{problem}, E_{method}, E_{findings}). Two **transition databases** store the vectors $T_{\text{PM}}(p)$ and $T_{\text{MF}}(p)$ for each paper, enabling retrieval based on reasoning patterns rather than topical content.

Multi-Database Retrieval. Given a query paper, we perform parallel retrieval across all five databases using cosine similarity. Node retrieval identifies papers with similar problems, methods, or findings individually, while transition retrieval surfaces papers with analogous reasoning structures Together, these provide a comprehensive view of conceptually related prior work.

3.2 Decomposed Novelty Assessment

Beyond retrieval, we leverage ideation space for fine-grained novelty assessment. Our approach identifies *which specific aspects* are novel and *which* build upon existing work.

Reasoning Graph Extraction. We decompose the input idea I into a reasoning graph $\mathcal{G}_I = (V_I, E_I)$ using an LLM with structured prompting. Unlike the simpler extraction for database construction, this produces a detailed graph representing the logical flow of the hypothesis. Nodes are classified into five types based on argumentative role: Background Context (BG) for established facts, Research Problem (RP) for limitations being addressed, Reasoning Insight (RI) for intermediate reasoning steps, Proposed Approach (PA) for concrete methodology, and Intended Contribution (CO) for expected outcomes. Edges represent logical dependencies, capturing argumentative flow from motivation through method to contribution.

Idea Encoding and Retrieval. We map each node to the appropriate sub-space based on its role: BG and RP nodes use E_{problem}; RI and PA nodes use E_{method}; CO nodes use E_{findings}. For each node $v \in V_I$, we query the corresponding database to retrieve similar prior claims, obtaining fine-grained evidence of which aspects have precedent.

Graph-Based Novelty Aggregation. We synthesize retrieval results using our novelty evaluation algorithm, which weights nodes by structural importance, where central arguments connecting multiple claims matter more than peripheral details. For each node v, we define $s(v)$ as the maximum similarity to any retrieved prior work. The importance weight combines two centrality measures:

$$w(v) = \frac{1}{2}\big(C_{\text{degree}}(v) + C_{\text{betweenness}}(v)\big) \quad (6)$$

where $C_{\text{degree}}(v) = \frac{d(v)}{|V_I|-1}$ measures connectivity, and $C_{\text{betweenness}}(v) = \sum_{u \neq v \neq t} \frac{\sigma_{ut}(v)}{\sigma_{ut}}$ quantifies bridging importance (σ_{ut} counts shortest paths from u to t; $\sigma_{ut}(v)$ counts those through v). Weights are rescaled to $[0.5, 2.0]$ for balanced contribution. The overall novelty score is:

$$N(I) = 1 - \frac{\sum_{v \in V_I} s(v) \cdot \tilde{w}(v)}{\sum_{v \in V_I} \tilde{w}(v)} \quad (7)$$

This formulation ensures that novelty in structurally critical components, i.e., the intellectual core of the contribution, dominates the assessment. Beyond the aggregate score, for each node, we provide its similarity to the closest prior work, and the specific retrieved evidence, enabling researchers to understand not just *how novel* their idea is, but *where* novelty lies and *what prior work* to engage with.

4 Experiments

Our experimental evaluation targets two fundamental questions: Can our framework accurately retrieve the specific prior work relevant to a new idea's components? And does it generate novelty assessments that align with expert human judgment? We answer these through two complementary evaluations: **Hierarchical Retrieval Performance** and **Novelty Evaluation**.

4.1 Experimental Setup

Evaluation Datasets. For retrieval evaluation, we use 500 papers randomly selected from ICLR 2025 submissions as the test query set. For each test paper, we extract its components and compute five query vectors (three node-level, two transition-level), measuring retrieval performance against ground-truth relevant papers identified through citation relationships. For novelty meta-evaluation, we use 93 research ideas from the AI-Researcher dataset (Si et al., 2025), which includes expert-annotated novelty scores.

Ground Truth Construction for Retrieval. For retrieval tasks, we employ GPT-5.2 (OpenAI, 2025) to curate ground truth by processing both references and OpenReview expert reviews. We extract all references cited by each test paper and use an LLM to classify their relevance for novelty assessment. The LLM is provided with structured information about both papers (title, research question, method, key findings) and determines whether the cited work is pertinent to evaluating novelty. Additionally, we extract papers mentioned in human reviews as being relevant to the novelty assessment. Our extraction process identifies papers relevant to three dimensions and papers that share similar Reasoning Patterns between adjacent nodes. After filtering, 466 out of 500 papers have valid annotations and are used for retrieval evaluation. More details are in Appendix A.3, A.4 and A.5.

Retrieval Baselines. We compare against strong sparse and dense retrieval models, including (1) **BM25** (Robertson et al., 2009), a traditional sparse keyword retrieval method; (2) **E5$_{\text{base}}$-v2** (Wang et al., 2022), a general-purpose semantic embedding model; (3) **SPECTER2** (Singh et al., 2023), a scientific document embedding model; and (4) **SciNCL** (Ostendorff et al., 2022), a contrastive model optimized for citation graphs. For node retrieval, each baseline constructs its own database using three input configurations: paper titles only, abstracts only, and concatenated title+abstract. For transition retrieval, baselines concatenate research problem and method approach texts to form transition representations.

Novelty Evaluation Baselines. For the novelty alignment task, we compare against two categories of baselines. First is **LLM-as-a-Judge**, where we prompt a set of large language models to assess novelty using the same evaluation rubric as expert annotators, including Qwen3-8B (Yang et al., 2025), Llama3.1-8B (Dubey et al., 2024), DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), and GPT-5 Mini and GPT-5 Nano (OpenAI, 2025), proprietary models representing compact, production-grade language models. The second is **Embedding-based Retrieval.** We use E5$_{\text{base}}$-v2 (Wang et al., 2022) embeddings with different input configurations (title only, summary only, and concatenated) to retrieve the top-1 most similar paper from the corpus, and compute novelty as $1 -$ similarity, where lower similarity indicates higher novelty.

Implementation Details. We implement our framework using PyTorch and train all models on NVIDIA A40 GPUs. For the encoder, we initialize from SPECTER2 (Singh et al., 2023) and fine-tune using the AdamW optimizer with a learning rate of 1×10^{-5}, employing a linear warmup for 700 steps followed by linear decay. We train for 10 epochs with a batch size of 32. For contrastive learning, we sample 2 positive and 8 negative examples per anchor, balancing memory efficiency and gradient stability. We use a temperature parameter of 0.1 for

Subspace	Model	Pos Mean	Neg Mean	Margin
Research Problem	Base	0.942	0.836	0.106
	Adapter	0.933	0.828	0.105
	Fine-tuned	0.792	0.058	**0.735**
Method Approach	Base	0.938	0.844	0.095
	Adapter	0.932	0.838	0.093
	Fine-tuned	0.654	0.010	**0.644**
Key Findings	Base	0.942	0.840	0.102
	Adapter	0.936	0.838	0.098
	Fine-tuned	0.697	0.020	**0.678**

Table 1: Embedding space statistics on validation set: Positive/Negative means and margins across subspaces.

the InfoNCE loss and set the focal loss gamma to 2.0 to emphasize hard negatives. All experiments use the same hyperparameters across the three subspaces to ensure fair comparison.

4.2 Embedding Space Analysis

To understand the separation capabilities of our trained encoders, we analyze the positive and negative sample means and margins on the validation set across different subspaces. As shown in Table 1, the fine-tuned models significantly increase the margin compared to base and adapter models, indicating superior discriminative power. The fine-tuned encoders achieve margins of 0.735, 0.644, and 0.678 for the three subspaces respectively, compared to 0.1 for base models, demonstrating that our contrastive training effectively separates functionally relevant papers from irrelevant ones.

4.3 Results of Hierarchical Retrieval

We evaluate retrieval performance on two distinct tasks: **Node Retrieval** (identifying specific scientific entities, i.e., Problems, Methods, and Findings) and **Transition Retrieval** (identifying analogous reasoning patterns between these entities). We report **Recall@K**, **NDCG@K**, and **Hit Rate@K** to measure retrieval quality.

Node Retrieval Results. Table 2 presents the performance for retrieving scientific entities. Our Ideation Space consistently achieves the highest performance across all metrics. For instance, our **Recall@90** reaches **0.483**, significantly outperforming the best baseline (SPECTER2 concat: 0.423). Even our individual subspaces (RP, MA, KF) often outperform the baselines, highlighting the effectiveness of decomposing scientific content into functional units rather than treating documents as monolithic blocks. The strong performance across different K values demonstrates that our approach retrieves relevant prior work at both high precision (low K) and high coverage (high K).

Transition Retrieval Results. Table 3 presents the results for retrieving reasoning patterns. Identifying logical leaps is a more complex task than entity matching. While $E5_{base}$-v2 shows competitive ranking performance (NDCG), our **Transition Space** excels in **Recall@90 (0.275)** and **Hit Rate@90 (0.762)**. This demonstrates that explicitly modeling the vector transition between concepts (T_{PM} and T_{MF}) is more effective at capturing diverse reasoning paths than static embeddings, providing a critical foundation for assessing methodological novelty.

4.4 Results of Novelty Evaluation

We evaluate whether our framework generates novelty assessments that align with expert human judgment. We measure the correlation between predicted novelty scores and ground-truth expert scores using both Pearson and Spearman correlation coefficients. Table 5 shows the correlation results. Our method achieves the highest correlation with expert judgment (Pearson: **0.370**, p=0.0003; Spearman: **0.384**, p=0.0002), outperforming all baselines including GPT-5 Mini (Pearson: **0.369**). Among embedding-based retrieval baselines, SPECTER2 performs best (Pearson: **0.333** for concat), followed by $E5_{base}$-v2 (Pearson: **0.287** for concat), while SciNCL shows the weakest correlation (Pearson: **0.157** for concat). The strong performance of our approach demonstrates that our decomposed ideation space, combined with graph-theoretic importance weighting, captures the multi-dimensional nature of novelty better than both LLM-based and simple embedding-based approaches, while being more computationally efficient than large language models.

4.5 Retrieval Case Study

To demonstrate the effectiveness of our multi-dimensional retrieval approach, we conduct a qualitative analysis on the paper *LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models*. Table 4 presents a concise comparison between our method and the baseline.

Our approach captures nuanced aspects that align closely with the query paper's contributions. For the **Research Problem** dimension, both methods retrieve SEED-Bench-2, recognizing the need for comprehensive MLLM evaluation. However,

Model	Recall @ K			NDCG @ K			Hit Rate @ K		
	9	30	90	9	30	90	9	30	90
Baselines									
BM25 (title)	0.087	0.159	0.232	0.140	0.152	0.180	0.481	0.622	0.719
BM25 (abstract)	0.143	0.246	0.365	0.230	0.236	0.281	0.674	0.794	0.856
BM25 (concat)	0.157	0.276	0.412	0.251	0.261	0.314	0.710	0.811	0.891
$E5_{base}$-v2 (title)	0.110	0.180	0.264	0.169	0.175	0.207	0.573	0.682	0.779
$E5_{base}$-v2 (abstract)	0.134	0.239	0.372	0.204	0.218	0.268	0.635	0.773	0.873
$E5_{base}$-v2 (concat)	0.144	0.252	0.389	0.216	0.230	0.280	0.678	0.788	0.891
SPECTER2 (title)	0.093	0.156	0.233	0.145	0.148	0.178	0.491	0.622	0.725
SPECTER2 (abstract)	0.145	0.259	0.393	0.215	0.230	0.281	0.644	0.764	0.878
SPECTER2 (concat)	0.160	0.282	0.423	0.242	0.257	0.310	0.667	0.811	0.893
SciNCL (title)	0.090	0.149	0.231	0.133	0.138	0.169	0.470	0.588	0.719
SciNCL (abstract)	0.136	0.256	0.392	0.206	0.227	0.279	0.622	0.779	0.869
SciNCL (concat)	0.157	0.281	0.427	0.233	0.250	0.305	0.697	0.788	0.886
Ours									
Subspace: RP	0.160	0.300	0.450	0.238	0.264	0.322	0.678	0.831	0.897
Subspace: MA	0.167	0.299	0.453	**0.251**	0.270	0.328	0.697	0.835	0.901
Subspace: KF	**0.172**	0.304	0.444	**0.251**	0.270	0.324	0.695	0.820	0.886
Ideation Space	0.169	**0.329**	**0.483**	0.250	**0.277**	**0.335**	**0.702**	**0.861**	**0.927**

Table 2: Node Retrieval Performance Comparison (Recall, NDCG, Hit Rate). K values are scaled by 3x (Displaying K=9, 30, 90 corresponding to data K=3, 10, 30). The best results for each metric are highlighted in bold.

Model	Recall @ K			NDCG @ K			Hit Rate @ K		
	9	30	90	9	30	90	9	30	90
Baselines (RP+MA)									
BM25	0.076	0.135	0.229	0.140	0.135	0.169	0.432	0.593	0.739
$E5_{base}$-v2	0.095	0.160	0.247	**0.168**	0.160	0.192	**0.492**	0.627	0.730
SPECTER2	0.090	0.153	0.236	0.162	0.157	0.189	0.461	0.602	0.719
SciNCL	0.090	0.154	0.236	0.157	0.152	0.182	0.467	0.602	0.708
Ours									
Subspace: P2M	0.094	0.168	0.267	0.157	0.160	0.199	0.449	0.611	0.744
Subspace: M2K	0.078	0.148	0.237	0.131	0.135	0.169	0.393	0.548	0.710
Transition Space	**0.099**	**0.180**	**0.275**	0.159	**0.166**	**0.203**	0.465	**0.643**	**0.762**

Table 3: Transition Retrieval Performance (Recall, NDCG, Hit Rate). K values are scaled by 3x (Displaying K=9, 30, 90 corresponding to data K=3, 10, 30). The best results for each metric are highlighted in bold.

our method's strength becomes evident in the **Method Approach** and **Key Findings** dimensions. Specifically, we retrieve *AMBER*, which shares the critical "LLM-free" design constraint mentioned in LongHalQA, and *HaELM*, which focuses on hallucination evaluation, directly relevant to the query paper's core objective. In contrast, the baseline retrieves general MLLM benchmarks (MME, MLLM-Bench) that assess broad multimodal capabilities but lack the specific focus on hallucination evaluation and LLM-free design. This demonstrates our method's ability to identify methodologically similar papers rather than merely topically related ones. The detailed retrieval results with full text excerpts are provided in Appendix B.

5 Related Work

AI-Assisted Scientific Discovery and Novelty Assessment. Recent advances in LLMs have catalyzed interest in AI-assisted scientific discovery (Du et al., 2024; Kulkarni et al., 2025; Ghafarollahi and Buehler, 2025; Qi et al., 2025) and the associated evaluation (Zheng et al., 2023; Jin et al., 2024; Feng et al., 2025; Baek et al., 2025). Lu et al. (2024) introduced the AI Scientist for automated research, while Si et al. (2025) conducted the first large-scale study comparing LLM and human-expert ideas. Wang et al. (2024) proposed SciMON to iteratively refine hypotheses for novelty, and Radensky et al. (2024) developed Scideator for human-LLM collaborative ideation with retrieval-

Query Paper	
LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models	
Our Multi-Dimensional Retrieval	
Research Problem	SEED-Bench-2: Benchmarking Multimodal Large Language Models
Method Approach	An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation
Key Findings	Evaluation and Analysis of Hallucination in Large Vision-Language Models
Baseline (Title + Abstract Based Retrieval)	
Rank 1	SEED-Bench-2: Benchmarking Multimodal Large Language Models
Rank 2	MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models
Rank 3	MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria

Table 4: Concise retrieval comparison for *LongHalQA*. Our multi-dimensional approach identifies specific hallucination benchmarks (AMBER, HaELM) while the baseline focuses on general MLLM evaluation frameworks. See Appendix B for detailed text excerpts.

Method	**Pearson**	**Spearman**
LLM-as-a-Judge		
Qwen3-8B	0.218	0.222
Llama3.1-8B	0.053	0.089
DeepSeek-R1-Distill-Llama-8B	0.223	0.233
GPT-5 Mini	0.369	0.385
GPT-5 Nano	0.168	0.211
Embedding-based Retrieval: $E5_{base}$-v2		
Title	0.089	0.032
Summary	0.236	0.245
Concat	0.287	0.309
Embedding-based Retrieval: SciNCL		
Title	0.018	0.003
Summary	0.169	0.144
Concat	0.157	0.122
Embedding-based Retrieval: SPECTER2		
Title	0.114	0.096
Summary	0.316	0.357
Concat	0.333	0.359
IdeationEval (Ours)	**0.370**	**0.384**

Table 5: Correlation between different novelty estimation methods and ground-truth expert scores.

augmented novelty evaluation. However, LLM-based evaluation typically suffers from sycophancy biases (Sharma et al., 2024), where models tend toward positive ratings regardless of actual quality, producing scores that lack discriminative power (Ye et al., 2025). Furthermore, existing methods treat novelty as unidimensional without distinguishing which aspects of an idea are novel. To address these limitations, we measure conceptual distance from retrieved evidence rather than relying on potentially biased LLM judgments, and our algorithm provides fine-grained assessments that identify where novelty lies along each dimension of scientific contribution.

Scientific Knowledge Representations. Learning effective representations of scientific knowledge (Beltagy et al., 2019; Zhang et al., 2023) is fundamental to literature retrieval, similarity measurement, etc. Cohan et al. (2020) introduced SPECTER, training Transformer encoders using citation-informed contrastive learning. Ostendorff et al. (2022) proposed SciNCL with neighborhood contrastive learning over citation graph embeddings, while Singh et al. (2023) developed SPECTER2 with task-specific adapters for multi-format representation. However, these methods encode papers into monolithic representations that conflate distinct conceptual aspects into a single vector, limiting their utility for fine-grained retrieval and novelty assessment. Our Ideation Space explicitly decomposes scientific knowledge into three functionally distinct sub-spaces, each learned through specialized contrastive training that leverages citation context to determine which conceptual dimension connects related papers.

6 Conclusion

We introduced the Ideation Space, a structured framework that decomposes scientific knowledge into three conceptually orthogonal dimensions, i.e., research problem, methodology, and core findings, enabling improved measurement of conceptual distance and ideation transitions. Building upon this representation, our Hierarchical Sub-Space Retrieval achieves up to 16.7% improvement in recall, while our Decomposed Novelty Assessment attains 0.37 correlation with expert judgments. Our work establishes a promising paradigm for navigating the rapidly expanding scientific landscape, with potential applications in accelerating literature review, supporting peer review processes, and guiding researchers toward more impactful contributions.

Limitations

Our framework focuses on the AI and machine learning domain, with training data drawn from four arXiv categories (cs.AI, cs.LG, cs.CV, and cs.CL). While this scope ensures high-quality, domain-specific representations, exploring how well our decomposed ideation space transfers to other scientific disciplines, such as biology, physics, or social sciences, remains an interesting direction for future work. Besides, our three-dimensional decomposition captures the core functional aspects of most scientific contributions, though certain interdisciplinary or theory-heavy works may exhibit more nuanced structures that could benefit from finer-grained representations.

Additionally, our evaluation relies on citation relationships and LLM-assisted ground truth construction, which, while providing scalable and reproducible benchmarks, may not capture all dimensions of conceptual relevance that human experts perceive. The novelty meta-evaluation uses 93 research ideas from the AI-Researcher dataset, which, though expert-annotated, represents a specific subset of research styles. Extending our evaluation to larger-scale human studies and diverse research communities would further validate the generalizability of our approach.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. 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.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tao Feng, Yihang Sun, and Jiaxuan You. 2025. Graph-eval: A lightweight graph-based LLM framework for idea evaluation. In *The Thirteenth International Conference on Learning Representations*.

Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint arXiv:2505.04651*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

OpenAI. 2025. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingyuan Qi, Zian Jia, Minqian Liu, Wangzhi Zhan, Junkai Zhang, Xiaofei Wen, Jingru Gan, Jianpeng Chen, Qin Liu, Mingyu Derek Ma, Bangzheng Li, Haohui Wang, Adithya Kulkarni, Muhao Chen, Dawei Zhou, Ling Li, Wei Wang, and Lifu Huang. 2025. MetaScientist: A human-AI synergistic framework for automated mechanical metamaterial design. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 404–436, Albuquerque, New Mexico. Association for Computational Linguistics.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Kai Zhang, Kaisong Song, Yangyang Kang, and Xiaozhong Liu. 2023. Content- and topology-aware representation learning for scientific multi-literature. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7490–7502, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A More Experimental Details

A.1 More Details of Structured Extraction

We use Qwen3-8B (model path: `Qwen/Qwen3-8B`) to extract structured representations from paper abstracts. The model decomposes each abstract into three atomic dimensions: Research Problem, Method Approach, and Key Findings, with the following configuration: temperature=0.0 for deterministic outputs, max_tokens=512 to accommodate detailed three-sentence responses, top-p=0.95 for controlled randomness, and repetition_penalty=1.1 to prevent redundant phrasing. Below is the complete prompt used for extraction:

`You are an expert academic writer and editor tasked with creating precise, informative structured summaries of research paper abstracts. Your goal is to distill the essence of complex research into a clear and concise summary that captures the paper's unique contributions and distinguishes it from related work.`

Task Instructions:
`Transform the following academic paper abstract into exactly THREE sentences. Each sentence should address one of the required aspects and be self-contained and clear to understand for researchers in related fields. Each sentence should be both concise and informative, maintaining the key information while being brief and not verbose.`

Required Elements to Include (One sentence each):

1. **Research Question/Problem**: What specific challenge, gap, or question does this work uniquely address that distinguishes it from existing research?
    - Focus on the novel problem formulation or unique angle the paper takes
    - Identify what specific gap in current literature this work fills
    - Emphasize the distinctive aspect of the research question that sets it apart
2. **Method/Approach**: What novel technique, algorithm, methodology, or experimental design is introduced that sets this work apart from previous approaches?
    - Highlight the innovative technical contribution or methodological advancement
    - Describe the unique algorithmic approach, architecture, or experimental setup
    - Emphasize what makes the method distinct from existing solutions
3. **Key Findings/Insights/Contributions**: What are the main empirical results, theoretical insights, or practical contributions that represent the paper's unique value and impact?
    - Focus on concrete results, performance improvements, or theoretical breakthroughs
    - Include quantitative metrics when available (accuracy, efficiency, etc.)
    - Emphasize the practical or theoretical significance of the findings

Quality Guidelines:

- **Structure**: Exactly 3 sentences, each addressing one required element in order
- **Clarity**: Use precise, concrete language with specific technical terms when necessary
- **Specificity**: Include quantitative results when available (e.g., "achieves 95% accuracy", "reduces computation by 40%")
- **Uniqueness**: Emphasize what makes this paper distinct from related work in the same field
- **Distinction**: Ensure each aspect captures non-overlapping information to maximize novelty evaluation
- **Accessibility**: Minimize technical jargon; when used, ensure it's essential and well-defined
- **Completeness**: Each sentence should stand alone and be informative without requiring additional context
- **Novelty Focus**: Prioritize novel contributions over standard or incremental improvements

Writing Style:

- Use active voice when possible for clarity and directness
- Start each sentence with the main point for that element to improve readability
- Ensure logical flow: problem → method → findings
- Focus on novel contributions rather than standard approaches or background information
- Use specific, concrete language rather than vague or generic descriptions
- Maintain academic tone while being accessible to researchers in related fields

What to Avoid:

- Generic phrases like "this paper presents" or "we propose" without specific details
- Repetitive information already implied by context or common knowledge
- Unnecessary background information available elsewhere in the literature
- More or fewer than exactly 3 sentences
- Overly technical terms without clear necessity or explanation
- Sentences that don't clearly address their assigned element
- Overlapping content between the three aspects (each should capture distinct information)

- Generic descriptions that could apply to multiple papers in the same field
- Vague statements without concrete details or specific contributions
- Overly broad claims without supporting evidence or specificity

Evaluation Criteria for Novelty:

- **Research Question**: Does it address a unique problem or take a novel angle on an existing problem?
- **Method**: Does it introduce a genuinely new technique, algorithm, or approach?
- **Findings**: Do the results represent significant advancement or unique insights?

Abstract to Summarize:
`{abstract}`

Your Three-Sentence Structured Summary:

1. **Research Question/Problem**: [Sentence addressing the unique challenge or question that distinguishes this work from existing research in the fields]
2. **Method/Approach**: [Sentence describing the novel technique, algorithm, or methodology that sets this work apart from previous approaches]
3. **Key Findings/Insights/Contributions**: [Sentence stating the unique empirical results, theoretical insights, or practical contributions with specific metrics when available]

A.2 More Details of Citation Edge Classification

We use Qwen3-4B-Instruct-2507 (model path: `Qwen/Qwen3-4B-Instruct-2507`) to classify citation relationships between paper pairs. The model evaluates three dimensions of citation relationships: Research Problem Relationship, Method Approach Relationship, and Key Findings Relationship, with the following configuration: max_new_tokens=1024 to accommodate detailed analysis, do_sample=False for deterministic outputs, and `enable_thinking=False` to suppress internal reasoning tokens. Below is the complete prompt used for this classification:

`You are an expert scientific researcher analyzing citation relationships between academic papers. Your task is to evaluate WHY a citing paper (P2) references a cited paper (P1) by assessing three distinct dimensions of their relationship.`

Input Structure:
For both papers, you will receive:

- **Title**
- **Research Question:** The core problem or question the paper addresses
- **Method:** The approach, techniques, or methodology used
- **Key Findings:** The main results, conclusions, or contributions
- **Citing Paper (P2):** The paper making the citation
- **Cited Paper (P1):** The paper being cited

Your Task:
Analyze the citation relationship across three dimensions and assign scores from 0.0 to 1.0 (in 0.2 increments) for each dimension. Consider what role P1 plays in P2's work.

Dimension 1: Research Problem Relationship
Compare how P2's research question relates to P1's research question.

- **1.0 - Direct Extension/Continuation:** P2 explicitly builds upon, extends, or is the next step in P1's research problem. P2 directly states it is addressing limitations, gaps, or natural progressions of P1's problem. The problems form a clear lineage.
  - Examples: "extending X to handle Y", "addressing the limitation of X in scenarios Z", "generalizing X's problem to broader contexts"
- **0.8 - Direct Competition (Same Problem):** P2 and P1 tackle the exact same research problem but propose different solutions. They are direct competitors addressing identical challenges. P2 likely compares its approach directly against P1 as a primary baseline.
  - Examples: Both papers aim to solve "few-shot image classification on ImageNet" or "neural machine translation for low-resource languages"
- **0.6 - Sub-problem or Specialization:** P2 addresses a specific sub-problem, special case, or narrower aspect of P1's broader problem. P1 tackles a general problem while P2 focuses on a particular instantiation or component of it.
  - Examples: P1 addresses "dialogue systems" while P2 focuses specifically on "turn-taking in task-oriented dialogues"
- **0.4 - Same Domain, Different Problems:** P2 and P1 work within the same research domain or application area but address distinctly different problems. They share high-level context but diverge in their specific research questions.
  - Examples: Both in "object detection" but P1 focuses on "small object detection" while P2 addresses "3D bounding box estimation"
- **0.2 - Tangential/Peripheral Connection:** Very weak connection. Papers share only broad field membership or abstract conceptual links. The research questions are fundamentally different but might share very high-level themes.

- Examples: Both in "computer vision" but P1 does "image segmentation" while P2 does "video captioning"

- **0.0 - No Connection:** Research questions are unrelated. P1 is cited for purely tangential reasons (e.g., citing a standard dataset paper, a general survey, or methodological tool unrelated to the core research question).

Dimension 2: Method Approach Relationship
Compare how P2's methodology relates to P1's methodology.

- **1.0 - Direct Use/Adaptation as Core Component:** P2 directly implements, adapts, extends, or modifies P1's method as a central, essential component of its approach. P1's method is foundational to P2's technical contribution. Removing P1's method would fundamentally change P2's approach.
  - Examples: "We adopt the X architecture from P1", "Building on P1's Y algorithm, we modify...", "Using P1's framework as our backbone..."
- **0.8 - Primary Baseline Comparison:** P2 implements P1's method specifically to compare against it as a main baseline or competitor. P1's method represents the state-of-the-art or primary alternative that P2 aims to outperform. Substantial discussion of methodological differences.
  - Examples: "Compared to P1's approach which uses X, our method employs Y..."
- **0.6 - Methodological Inspiration/Similar Framework:** P2 is conceptually inspired by P1's methodological approach or uses a similar framework/paradigm, but with significant modifications. The connection is more than superficial but P2 doesn't directly implement P1's method.
  - Examples: "Inspired by P1's use of attention mechanisms...", "Following a similar encoder-decoder structure to P1..."
- **0.4 - Minor Component or Tool:** P2 uses a specific technique, module, or tool from P1 as one small piece of its overall method. P1's contribution is auxiliary or peripheral to P2's main methodological approach.
  - Examples: Using P1's data augmentation technique, loss function, or evaluation protocol as one component among many
- **0.2 - Abstract/High-Level Similarity:** Very weak methodological connection. Papers might both use the same general class of methods (e.g., "both use transformers" or "both use reinforcement learning") but in substantively different ways. The connection is at the level of paradigm rather than specific technique.
- **0.0 - No Methodological Connection:** Methods are completely unrelated. P1 might be cited for its findings, problem formulation, or dataset, but not for its methodology.

Dimension 3: Key Findings Relationship
Compare how P2's findings relate to and engage with P1's findings.

- **1.0 - Foundational Premise/Dependency:** P2 explicitly relies on P1's findings as a foundational premise, assumption, or starting point for its own work. P1's findings are accepted as established facts that justify or motivate P2's approach. P2 builds its argument on P1's conclusions.
  - Examples: "Given P1's finding that X is crucial, we propose...", "P1 demonstrated that Y, therefore we investigate..."
- **0.8 - Direct Validation/Contradiction/Discussion:** P2 directly engages with P1's specific findings—either validating, contradicting, refining, or extensively discussing them in its own experimental results section. P1's findings are a central point of discussion in P2's analysis.
  - Examples: "We confirm P1's finding that...", "Contrary to P1's results...", "Our experiments show P1's conclusion holds only when..."
- **0.6 - Direct Improvement/Advancement:** P2 reports findings that directly improve upon, advance, or supersede the specific quantitative or qualitative results reported in P1. Direct performance comparison on the same metrics, datasets, or evaluation criteria.
  - Examples: "We improve upon P1's accuracy of X% to Y%", "While P1 achieved Z, our approach attains..."
- **0.4 - Background Context:** P2 mentions P1's findings only as general background, related work, or literature context. P1's findings are acknowledged but not central to P2's own results or analysis. Superficial citation in related work section.
  - Examples: "Previous work (P1) found that...", "P1 reported...", typically appearing only in introduction or related work
- **0.2 - Minimal/Tangential Relevance:** P1's findings are barely relevant to P2. Perhaps a single finding is mentioned in passing, or the connection is extremely indirect. P1 might be cited in a list of references without specific findings being discussed.
- **0.0 - No Engagement with Findings:** P1's findings are not discussed, not relevant, or P1 is cited purely for its method, problem formulation, or as a generic reference. The findings themselves play no role in P2.

Output Requirements:
Step 1: Provide detailed reasoning for your scores.
Step 2: After your reasoning, provide a final JSON output.
YOUR ENTIRE RESPONSE MUST BE A SINGLE JSON OBJECT.
The reasoning and analysis MUST be included as fields within the JSON.
Produce a single JSON object in exactly this format:

```
{
"research_problem_score": [score between 0.0-1.0],
"method_approach_score": [score between 0.0-1.0],
"key_findings_score": [score between 0.0-1.0],
"reasoning": "[One concise sentence summarizing
the primary reason for the citation]",
"research_problem_analysis": "[2-4 sentences
comparing P1's research question to P2's research
question. Explicitly identify the relationship type
and justify your score with specific evidence]",
"method_approach_analysis": "[2-4 sentences
comparing P1's method to P2's method. Explicitly
identify the relationship type and justify your
score with specific evidence]",
"key_findings_analysis": "[2-4 sentences comparing
P1's key findings to P2's key findings. Explicitly
identify the relationship type and justify your
score with specific evidence]"
}
```

IMPORTANT: You must respond with a single JSON object and nothing else. Do not include any preamble, explanation, or markdown formatting before or after the JSON. Your entire response must be the JSON object itself, starting with '{' and ending with '}'.

Paper Information:

```
<CITING PAPER (P2)>
Title: {citing_title}
Research Question: {citing_rq}
Method: {citing_method}
Key Findings: {citing_findings}

<CITED PAPER (P1)>
Title: {cited_title}
Research Question: {cited_rq}
Method: {cited_method}
Key Findings: {cited_findings}
```

A.3 More Details of Building Ground Truth for Node Retrieval from Human Reviews

We use GPT-5.2 (via OpenAI Responses API) to extract novelty-related references from human review comments. The model analyzes reviewer feedback and categorizes cited works into three dimensions: Research Problem Overlap, Method/Approach Overlap, and Key Findings Overlap, with the following configuration: max_output_tokens=2048 to accommodate detailed structured output, reasoning effort set to `'medium'` and text verbosity set to `'medium'`. Below is the complete prompt used for this extraction:

You are acting as an assistant to an academic reviewer. Your task is to extract and organize references from a reviewer's comments that are relevant to evaluating the **novelty** of a submission. Novelty in this context means how the submission's contribution compares to what has been done before. A reviewer often notes prior work to highlight redundancy or overlap: perhaps the underlying **research problem** has already been studied, the **method or approach** is borrowed, or the **key findings** have already been reported. Capturing these citations systematically will help assess where the new work is or is not novel.

Task Instructions:
To do this effectively, read the entire review carefully and identify every piece of prior work that the reviewer explicitly mentions or implicitly references when questioning novelty. For each identified work, decide which of the following three categories best describes why the reviewer brought it up:

Category Definitions:

1. **Research problem overlap** – the prior work poses or solves the *same research problem* as the submission. For example, if the reviewer writes "this paper revisits the same scheduling problem studied in *A Unified Approach to Fair Scheduling*," you would assign that citation to this category. Here we care about overlap in the *problem statement* or *objective*, regardless of how it was approached.
2. **Method/approach overlap** – the prior work uses the *same algorithm, technique, or methodology* as the submission. For instance, if a reviewer says "the authors' method is essentially the Transformer architecture introduced by *Attention is All You Need*," that reference belongs here. This category focuses on whether the procedure or algorithmic contribution is already known, even if the problem differs.
3. **Key findings overlap** – the prior work reports *results or conclusions* similar to those claimed by the submission. An example would be "*Deep Residual Learning for Image Recognition* already demonstrated that skip connections improve accuracy on ImageNet," when the reviewer criticizes the novelty of the submission's experimental findings. This category applies when the novelty concern is about the novelty of outcomes or insights.

Extraction Guidelines:
Only extract references that are actually tied to a novelty concern. Do **not** list general background citations, standard datasets (unless the review specifically notes missing comparisons), or passing mentions of common techniques that are not framed as novelty issues. The goal is to build a set of citations that the reviewer uses to question or contextualize the submission's novelty, so err on the side of excluding irrelevant or purely contextual references.

Required Fields for Each Reference:

- **title** (Required): Use the most specific paper title you can. If the reviewer explicitly names the paper, reproduce the title verbatim. If only an acronym or method name is given (e.g., "ResNet"), infer the likely full title (e.g., "Deep Residual Learning for Image Recognition"). If you cannot confidently infer the full title, include the exact span mentioned in the review instead.
- **title_completeness** (Required): Indicate whether the title you provided is complete (`"complete"`) or a partial/inferred title (`"incomplete"`).
- **authors** (Optional): Include authors only if the reviewer explicitly mentions them (e.g., "Smith et al."). Do not guess or infer authors.
- **venue** (Optional): Include the publication venue (conference, journal, etc.) only if it is explicitly mentioned by the reviewer.
- **year** (Optional): Include the publication year only if explicitly mentioned in the review.
- **context** (Required): Summarize why this reference is relevant for novelty evaluation in the chosen category—explain briefly whether it overlaps in problem, method, or findings.
- **original_mention** (Required): Quote one or two sentences from the review that mention this reference so that another reader can verify the context.
- **novelty_concern** (Required): Paraphrase the specific novelty issue raised, such as "previous work addressed the same problem," "uses the same method," or "reports similar results."
- **confidence** (Optional): If you wish, include a qualitative confidence level (`"high"`, `"medium"`, `"low"`) in your extraction.

Output Format (JSON only, no other text):

```
{
"research_problem": [
{
"title": "Full Paper Title Here",
"title_completeness": "complete",
"authors": ["Author1", "Author2"],
"venue": "Conference/Journal",
"year": "2020",
"context": "Why this work addresses the
same problem",
"original_mention": "Sentence(s) from the
review
mentioning the reference",
"novelty_concern": "What novelty issue is
raised
(problem overlap)",
"confidence": "high"
}
],
"method_approach": [
{
"title": "Full Paper Title Here",
"title_completeness": "complete",
"authors": ["Author1", "Author2"],
"venue": "Conference/Journal",
"year": "2020",
"context": "Why this work uses the same
method or
approach",
"original_mention": "Sentence(s) from the
review
mentioning the reference",
"novelty_concern": "What novelty issue is
raised
(method overlap)",
"confidence": "medium"
}
],
"key_findings": [
{
"title": "Full Paper Title Here",
"title_completeness": "incomplete",
"authors": [],
"venue": null,
"year": null,
"context": "Why this work reports similar
findings",
"original_mention": "Sentence(s) from the
review
mentioning the reference",
"novelty_concern": "What novelty issue is
raised
(finding overlap)",
"confidence": "low"
}
]
}
```

If a particular category has no relevant citations, output an empty array for that category. If no novelty-relevant references are found at all, output:

```
{
"research_problem": [],
"method_approach": [],
"key_findings": []
}
```

Review text to analyze:
`{REVIEW_TEXT}`

A.4 More Details of Building Ground Truth for Node Retrieval from References

We use GPT-5.2 (via OpenAI Responses API) to filter paper references based on novelty relevance. The model compares citing and cited papers across three dimensions: Research Problem Overlap, Method/Approach Overlap, and Key Findings Overlap, with the following configuration: max_output_tokens=1024 for structured boolean classification, reasoning effort configurable as texttt'medium'), and text verbosity set to `'medium'` for balanced output. Below is the complete prompt used for this filtering:

```
You are an expert scientific researcher
analyzing novelty relationships between
academic papers. Your task is to
```

evaluate whether a cited reference poses a novelty concern for a citing paper by assessing three distinct dimensions of their relationship.

Input Structure:
For both papers, you will receive:

- **Title**
- **Research Question:** The core problem or question the paper addresses
- **Method/Approach:** The approach, techniques, or methodology used
- **Key Findings:** The main results, conclusions, or contributions
- **Citing Paper (Current Paper):** The paper whose novelty we are evaluating
- **Cited Paper (Reference):** A paper that the current paper cites

Your Task:
Analyze the novelty relationship across three dimensions and assign boolean values (true/false) for each dimension. A reference poses a "novelty concern" if it significantly overlaps with the citing paper's contributions in any dimension, potentially undermining the citing paper's novelty claims.

Dimension 1: Research Problem Overlap
Does the cited paper address the same or very similar research problem/question as the citing paper?

- **True (poses novelty concern):** The cited paper tackles the exact same research problem, or a nearly identical problem that makes the citing paper's problem formulation unoriginal. The citing paper is essentially revisiting or only slightly modifying a problem that has already been studied in the cited work.
  - Examples: Both papers address "federated learning with heterogeneous data distributions" or "neural architecture search for mobile devices"
- **False (no novelty concern):** The cited paper addresses a different research problem, even if there are some conceptual connections or shared domain knowledge. The problems are fundamentally distinct despite potential thematic overlap.
  - Examples: One paper studies "image classification" while the other studies "object detection", even though both are computer vision tasks

Dimension 2: Method/Approach Overlap
Does the cited paper use the same or very similar methodology/approach as the citing paper?

- **True (poses novelty concern):** The cited paper employs essentially the same algorithmic approach, technical method, or methodological framework as the citing paper. The citing paper's methodological contribution is largely anticipated or directly borrowed from the cited work.
  - Examples: Both papers use "contrastive learning with momentum encoders" or "transformer-based sequence-to-sequence models with multi-head attention"
- **False (no novelty concern):** The cited paper uses different methods or approaches, even if there are some shared techniques or general paradigms. The methodological approaches are substantively different.
  - Examples: One uses "reinforcement learning" while the other uses "supervised learning", despite both being machine learning approaches

Dimension 3: Key Findings Overlap
Does the cited paper report similar or identical key findings/results as the citing paper?

- **True (poses novelty concern):** The cited paper demonstrates or concludes similar key results, insights, or empirical findings as the citing paper. The citing paper's main contributions in terms of results or conclusions have already been established in the cited work.
  - Examples: Both papers show that "adding skip connections improves gradient flow in deep networks" or "self-attention mechanisms outperform RNNs on long sequences"
- **False (no novelty concern):** The cited paper reports different findings or results, even if there are some related observations or shared datasets. The core conclusions and insights are distinct.
  - Examples: One paper finds "method A achieves 85% accuracy" while another finds "method B achieves 87% accuracy" on the same dataset – the specific findings differ

Output Requirements:
A reference poses a "novelty concern" if ANY of the three dimensions shows significant overlap (true). If a reference shows overlap in at least one dimension, it should be considered novelty-relevant and retained for further novelty analysis.
Step 1: Provide detailed analysis for each dimension with specific evidence from the provided texts.
Step 2: After your analysis, provide a final JSON output.
YOUR ENTIRE RESPONSE MUST BE A SINGLE JSON OBJECT.
The analysis and comparisons MUST be included as fields within the JSON.
Produce a single JSON object in exactly this format:

```
{
"is_novelty_relevant": true/false,
"reasoning":    "[One   concise   sentence
summarizing whether
this reference poses a novelty concern for
the citing
paper]",
"research_problem_overlap": true/false,
"method_approach_overlap": true/false,
"key_findings_overlap": true/false,
"research_problem_comparison":       "[2-4
```

sentences comparing the citing paper's research question to the cited paper's research question. Explicitly state whether there is overlap and justify your assessment with specific evidence from the provided texts.]", "method_approach_comparison": "[2-4 sentences comparing the citing paper's method/approach to the cited paper's method/approach. Explicitly state whether there is overlap and justify your assessment with specific evidence from the provided texts.]", "key_findings_comparison": "[2-4 sentences comparing the citing paper's key findings to the cited paper's key findings. Explicitly state whether there is overlap and justify your assessment with specific evidence from the provided texts.]", "overall_analysis": "[2-3 sentences explaining the overall novelty relationship between the papers and why this reference is or isn't relevant to the citing paper's novelty claims.]"
}

IMPORTANT: You must respond with a single JSON object and nothing else. Do not include any preamble, explanation, or markdown formatting before or after the JSON. Your entire response must be the JSON object itself, starting with '{' and ending with '}'.

Paper Information:

<CITING PAPER (Current Paper)>
Title: {citing_title}
Research Question: {citing_research_question}
Method/Approach: {citing_method_approach}
Key Findings: {citing_key_findings}

<CITED PAPER (Reference)>
Title: {cited_title}
Research Question: {cited_research_question}
Method/Approach: {cited_method_approach}
Key Findings: {cited_key_findings}

A.5 More Details of Building Ground Truth for Transition Retrieval from References

We use GPT-5.2 (via OpenAI Responses API) to filter paper references based on novelty relevance in transition spaces. The model analyzes reasoning transitions between concepts across two dimensions: Research Problem → Method/Approach (P2M) and Method/Approach → Key Findings (M2K), with the following configuration: max_output_tokens=4096 for detailed transition analysis, reasoning effort configurable as 'medium'), and text verbosity set to 'medium' for balanced output. Below is the complete prompt used for this filtering:

You are an expert scientific researcher analyzing novelty relationships between academic papers in terms of their reasoning transitions. Your task is to evaluate whether a cited reference poses a novelty concern for a citing paper by assessing two distinct transition dimensions.

Input Structure:
For both papers, you will receive:

- **Research Question**
- **Method/Approach**
- **Key Findings**
- **Citing Paper (Current Paper):** The paper whose novelty we are evaluating
- **Cited Paper (Reference):** A paper that the current paper cites

Your Task:
Analyze the novelty relationship by examining whether the cited paper anticipates or establishes similar reasoning transitions as the citing paper. A reference poses a "novelty concern" if it demonstrates similar logical connections between concepts, potentially undermining the citing paper's novelty claims.

Dimension 1: P2M Transition Similarity (Research Problem → Method/Approach)
Based on the research problem and method/approach described, does the cited paper appear to follow a similar logical reasoning pattern when transitioning from problem to method as the citing paper?

- **True (poses novelty concern):** The cited paper appears to connect its research problem to methodological choices through similar logical reasoning as the citing paper. The relationship between problem formulation and method selection suggests analogous intellectual approaches, even if the specific problems and methods differ.
  - Examples: Both papers show similar problem-to-method reasoning like connecting gradient-related issues with attention mechanisms, or linking memory retention problems with replay-based solutions
- **False (no novelty concern):** The cited paper appears to connect problem and method through fundamentally different logical reasoning. The relationship between problem formulation and method selection suggests distinct intellectual approaches, even if there might be superficial similarities in terminology.
  - Examples: One paper connects error analysis problems with architectural innovations while another connects theoretical convergence questions with optimization algorithms

Dimension 2: M2K Transition Similarity (Method/Approach → Key Findings)
Based on the method/approach and key findings described, does the cited paper appear to follow a similar logical reasoning pattern when transitioning from method application to deriving findings as the citing paper?

- **True (poses novelty concern):** The cited paper appears to connect methodological application to key findings through similar logical reasoning as the citing paper. The relationship between method execution and result interpretation suggests analogous analytical approaches, even if the specific methods and findings differ.
  - Examples: Both papers show similar method-to-findings reasoning like connecting attention mechanism analysis with syntactic insights, or linking representation analysis with contextual learning strategies
- **False (no novelty concern):** The cited paper appears to connect method and findings through fundamentally different logical reasoning. The relationship between method execution and result interpretation suggests distinct analytical approaches, even if there might be superficial similarities in experimental design.
  - Examples: One paper connects component ablation with efficiency insights while another connects mathematical analysis with theoretical understanding

Output Requirements:
Return ONLY a single JSON object wrapped in a markdown code block (“‘json ... “‘).
Do NOT output any additional text before or after the JSON.
Keep all string fields concise to avoid truncation.

```
“‘json
{
"is_novelty_relevant": true/false,
"reasoning": "[One concise sentence summarizing whether
this reference poses a novelty concern]",
"p2m_transition_similarity": true/false,
"m2k_transition_similarity": true/false,
"p2m_transition_analysis": "[1-2 sentences. Compare the
intellectual bridge from problem formulation to method
choice.]",
"m2k_transition_analysis": "[1-2 sentences. Compare the
intellectual bridge from method application to
findings/claims.]",
"overall_analysis": "[1-3 sentences. Overall novelty
relationship summary.]"
}
“‘
```

Paper Information:

```
<CITING PAPER (Current Paper)>
Research Question:
{citing_research_question}
Method/Approach: {citing_method_approach}
Key Findings: {citing_key_findings}

<CITED PAPER (Reference)>
Research Question:
{cited_research_question}
Method/Approach: {cited_method_approach}
Key Findings: {cited_key_findings}
```

B Qualitative Analysis: Retrieval Case Studies

Table 6: Detailed retrieval comparison for *LongHalQA*. The Query Paper details are presented first. For each dimension, we display the top retrieved paper's Title and relevant Text. Our retrieval surfaces specific hallucination benchmarks (AMBER, HaELM) while the baseline focuses on general MLLM benchmarks.

Query Paper Details		
Title: LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models		
Research Problem: This work addresses the gap in existing benchmarks for evaluating hallucinations in multimodal large language models (MLLMs), which rely on **simplistic discriminative questions** or **computationally intensive LLM evaluators**.		
Method Approach: The paper introduces LongHalQA, an **LLM-free benchmark** featuring **6K long and complex hallucination texts** generated by GPT4V, along with two unified tasks—hallucination discrimination and completion.		
Key Findings: Extensive experiments on recent MLLMs reveal **significant challenges in handling long and complex hallucination texts**, highlighting the necessity for advanced benchmarks like LongHalQA.		
Dimension	**Retrieved Paper (Title & Text)**	**Score**
Our Multi-Dimensional Retrieval (Top 1 per Dimension)		
Research Problem	**Title:** SEED-Bench-2: Benchmarking Multimodal Large Language Models **Text:** This work addresses the gap in existing MLLM benchmarks by identifying the need for a **comprehensive evaluation framework** that assesses the **hierarchical multimodal capabilities** of models beyond single-image-text comprehension.	0.893
Method Approach	**Title:** An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation **Text:** The paper introduces AMBER, an **LLM-free, multi-dimensional benchmark** that efficiently evaluates **hallucination types** across a wide range of tasks, using a **low-cost and scalable evaluation pipeline**.	0.826
Key Findings	**Title:** Evaluation and Analysis of Hallucination in Large Vision-Language Models **Text:** HaELM achieves approximately 95% performance comparable to ChatGPT, demonstrating strong effectiveness in **evaluating hallucination** while providing practical benefits such as **local deployment**.	0.835
Problem → Method	**Title:** Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion **Text:** This work addresses the unique challenge of evaluating image-context reasoning in large visual-language models (LVLMs) by focusing on the **nuanced detection of hallucination and visual illusion**.	0.660
Method → Findings	**Title:** An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation **Text:** AMBER demonstrates significant practical value by providing a comprehensive assessment of mainstream MLLMs like GPT-4V, revealing **performance disparities** and offering **actionable guidelines to mitigate hallucinations**.	0.682
Baseline Title & Abstract Based Retrieval (Top 5)		
Rank 1	**Title:** SEED-Bench-2: Benchmarking Multimodal Large Language Models **Text:** This work addresses the gap in existing MLLM benchmarks by identifying the need for a **comprehensive evaluation framework** that assesses the **hierarchical multimodal capabilities** of models.	0.893
Rank 2	**Title:** MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models **Text:** This work addresses the gap in comprehensive evaluation of Multimodal Large Language Models (MLLMs), as existing case studies fail to provide a **holistic assessment of their perception and cognition capabilities**.	0.879
Rank 3	**Title:** MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria **Text:** This work addresses the gap in evaluating multimodal large language models (MLLMs) by focusing on the challenge of assessing creative and associative tasks through **user experience-driven metrics**.	0.877
Rank 4	**Title:** SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension **Text:** This work addresses the lack of comprehensive evaluation frameworks for generative comprehension in multimodal large language models (MLLMs), focusing on the unique challenge of assessing both **image and video understanding**.	0.626

Continued on next page...

Table 6: (continued)

Table 6 – *Continued from previous page*		
Query Paper: *LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models*		
Dimension	**Retrieved Paper (Title & Text)**	**Score**
Rank 5	**Title:** TinyLVLM-eHub: Towards Comprehensive and Efficient Evaluation for Large Vision-Language Models **Text:** This work addresses the gap in evaluating Large Vision-Language Models (LVLMs) by focusing on a comprehensive assessment of their multimodal capabilities, particularly in areas such as **object hallucination**.	0.625