



Sci-Reasoning: A Dataset Decoding AI Innovation Patterns

Jiachen Liu*

Maestro Harmon*

Orchestra Research

Zechen Zhang

*Equal contribution

Abstract

While AI innovation accelerates rapidly, the intellectual process behind breakthroughs—how researchers identify gaps, synthesize prior work, and generate insights—remains poorly understood. The lack of structured data on scientific reasoning hinders systematic analysis and development of AI research agents. We introduce **Sci-Reasoning**, the first dataset capturing the intellectual synthesis behind high-quality AI research. Using community-validated quality signals and an LLM-accelerated, human-verified pipeline, we trace Oral and Spotlight papers across NeurIPS, ICML, and ICLR (2023–2025) to its key predecessors, articulating specific reasoning links in a structured format. Our analysis identifies 15 distinct thinking patterns, with three dominant strategies accounting for 52.7%: Gap-Driven Reframing (24.2%), Cross-Domain Synthesis (18.0%), and Representation Shift (10.5%). The most powerful innovation recipes combine multiple patterns: Gap-Driven Reframing + Representation Shift, Cross-Domain Synthesis + Representation Shift, and Gap-Driven Reframing + Cross-Domain Synthesis. This dataset enables quantitative studies of scientific progress and provides structured reasoning trajectories for training the next generation AI research agents.

1 Introduction

The field of artificial intelligence is experiencing an unprecedented pace of innovation. Breakthroughs in large language models (LLMs), reinforcement learning from visual reasoning, vision-language-action models, and related AI systems have transformed what AI systems can accomplish (Ouyang et al., 2022; Ramesh et al., 2022; Vaswani et al., 2017). Yet despite this rapid progress, the intellectual process by which these breakthroughs emerge remains poorly understood. How do researchers identify promising gaps in existing work? How

do they synthesize ideas from multiple predecessors into novel contributions? What patterns of reasoning characterize high-quality research? Understanding these reasoning trajectories is not only scientifically valuable in itself—it is the key to enabling the next generation of scientific discoveries. Yet currently, these questions are answered through subjective, anecdotal narratives rather than the structured data necessary for systematic analysis or machine learning.

The lack of structured, large-scale data on scientific *reasoning*—the “how” and “why” behind building on prior work—prevents rigorous study of the innovation process itself. While citation networks provide valuable information about influence patterns (Jo et al., 2022; Ghosal et al., 2021), they capture only the *fact* of a citation, not the *nature* of the intellectual relationship. A paper might cite another work as a baseline to surpass, as a foundational concept to extend, or as a methodology to combine with other techniques. These distinctions are critical for understanding how breakthroughs happen, yet they are lost in simple citation graphs. Moreover, existing approaches to intellectual lineage tracing typically identify only a single “progenitor” paper (Jo et al., 2022) or treat citations uniformly (Zhang et al., 2024), missing the reality that most advances emerge from synthesizing insights across multiple prior works.

We introduce **Sci-Reasoning**, the first dataset designed to capture the structured intellectual synthesis behind high-quality AI research. Our contribution is three-fold: (1) We present a systematic methodology for identifying high-quality papers using community-validated signals (Oral/Spotlight status at NeurIPS, ICML, and ICLR) and tracing their intellectual lineage to its key predecessors through an LLM-accelerated, human-verified pipeline. Through a model ablation study comparing GPT-5.2, GPT-5, GPT-5-mini, and GPT-4.1 on predecessor extraction (Section 5.2), we iden-

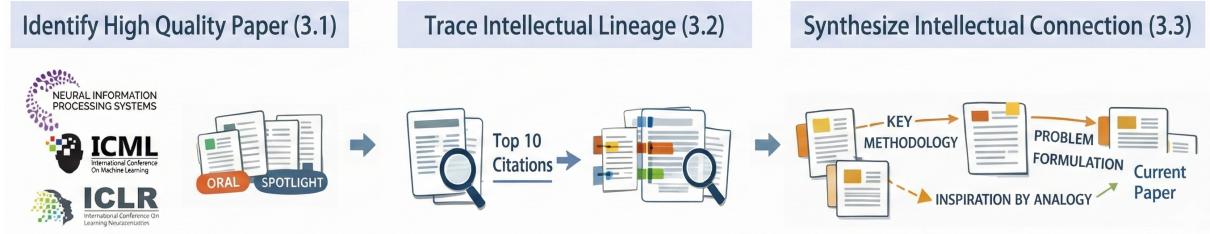


Figure 1: Overview of the Sci-Reasoning dataset construction pipeline. Our methodology consists of four main stages: (1) identifying high-quality papers using community-validated signals, (2) tracing intellectual lineage to key predecessors via LLM analysis, (3) generating structured reasoning trajectories with lineage links that capture roles, relationships, and intellectual moves

tify GPT-5 as achieving the optimal cost-quality balance with 89.73% recall, outperforming even newer models while maintaining computational efficiency for large-scale dataset construction. (2) We curate a dataset of 3,819 papers (999 Oral, 2,820 Spotlight) across NeurIPS, ICML, and ICLR (2023–2025) with richly annotated reasoning trajectories that capture not just *which* papers influenced each breakthrough, but *how*—through structured “Lineage Links” that include predecessor roles, relationship types, and natural language descriptions of the reasoning process. This structured format provides training data for AI research agents to learn expert research reasoning patterns. (3) We provide rigorous quality validation through test set evaluation against ground-truth papers and multi-model cross-validation (Section 3.4), demonstrating that frontier LLMs can predict research directions from intellectual predecessors with up to 49.35% accuracy (Section 5). Figure 1 provides an overview of our complete methodology pipeline.

Analysis of Sci-Reasoning reveals concrete, actionable patterns of innovation in AI research—patterns that represent the key to understanding and potentially automating aspects of future scientific discoveries. We identify **15 distinct thinking patterns**, with three dominant strategies accounting for 52.7% of all papers. **Gap-Driven Reframing is the dominant thinking pattern** (924 papers, 24.2%), where researchers diagnose a specific limitation and reframe the problem to map onto better-suited methods—suggesting that breakthrough research starts with identifying crisp, quantifiable gaps. **Cross-Domain Synthesis is the second most common pattern** (687 papers, 18.0%), where researchers import ideas from other fields and engineer compatibility layers, demonstrating that successful innovation often involves borrowing and adapting rather than inventing from

scratch. **Representation Shift appears in 401 papers (10.5%)**, where changing core primitives or abstractions simplifies the problem. Beyond individual patterns, successful research combines multiple strategies into repeatable “innovation recipes.” The most powerful combination pairs **Gap-Driven Reframing with Representation Shift** (318 occurrences), representing a “Reframe + New Primitive” strategy. The second combination, **Cross-Domain Synthesis with Representation Shift** (233 occurrences), embodies an “Import + Adapt” approach, while the third, **Gap-Driven Reframing with Cross-Domain Synthesis** (204 occurrences), reflects “Diagnose + Borrow.” These findings provide both a quantitative understanding of how scientific progress occurs and practical frameworks for generating novel research ideas. Critically, by capturing these patterns in structured format, our dataset enables the development of AI research agents that can learn from expert researchers’ reasoning trajectories and apply these patterns to accelerate future discoveries.

Our source code is available at <https://github.com/AmberLJC/Sci-Reasoning> and the dataset is available at <https://huggingface.co/datasets/AmberLJC/Sci-Reasoning>.

2 Related Work

2.1 Research Lineage and Citation Analysis

The study of scientific influence through citation analysis has evolved from simple counts and network structures (Garfield, 1955; Bornmann and Daniel, 2008) to analyzing citation context and intent (Hernández-Álvarez and Gómez, 2016; Cohan et al., 2019; Lahiri et al., 2023; Jantsch et al., 2025). Recent work has focused on identifying key predecessors: the progenitor index (Jo et al., 2022) identifies the single most influential prior work,

while research lineage graphs (Ghosal et al., 2021) identify significant citations where papers heavily rely on cited work. PST-Bench (Zhang et al., 2024) formalized “Publication Source Tracing” as a task, exploring statistical, graph-based, and language model approaches with limited success.

Our work differs by identifying a *set* of its key predecessors rather than a single progenitor, and focusing on the *reasoning content*—the specific intellectual moves and synthesis strategies—rather than just influence patterns. Complementing work on contribution extraction (Pramanick et al., 2025; Lawrence and Reed, 2020), we capture not individual contribution statements but the *synthesis process* showing how contributions emerge from combining multiple prior works.

2.2 Scientific Understanding and Reasoning

Several datasets address scientific reasoning from different angles. PeerRead (Kang et al., 2018) provides peer reviews from major ML/NLP venues, demonstrating that review discourse offers valuable signals about research quality. GPQA-Diamond (Rein et al., 2023) provides PhD-level science questions testing knowledge retrieval, while Polymathic AI datasets (Cranmer et al., 2024) focus on domain-specific problem-solving. Research on narrative science (Morgan and Wise, 2017; Green and Donahue, 2017) studies how scientific progress is constructed and communicated through rhetorical and argumentative structures.

Our dataset occupies a unique niche: it captures the *structured intellectual trajectories* behind breakthrough research—not just which papers influenced each advance, but specifically how ideas were combined, extended, or reframed. We leverage community-validated signals (oral/spotlight status) similar to PeerRead but focus on intellectual lineage and the reasoning that leads to high-quality papers, providing large-scale structured data for quantitative analysis of scientific narratives.

2.3 AI-Assisted Research and Discovery

Recent advances in LLMs have enabled significant progress in automating scientific research, from specialized tools to autonomous agents executing complete research workflows (Kon et al., 2025a,b; Zheng et al., 2025). Systems like LitLLM (Agarwal et al., 2024) use retrieval-augmented generation for literature review, neural search engines like Exa (Exa AI, 2024) enable semantic paper discovery, while Research Knowledge Graphs (Zloch

et al., 2025) provide machine-actionable representations of research relations.

Our dataset provides crucial training and evaluation data for such systems. Understanding how high-quality research synthesizes prior work is essential for developing AI agents that generate valid scientific reasoning, with our structured synthesis traces offering explicit examples of reasoning patterns for research ideation and literature analysis.

3 Methodology

Our methodology for creating the Sci-Reasoning dataset consists of three phases: identifying high-quality papers, tracing their intellectual lineage, and articulating structured lineage graphs that capture the multi-dimensional relationships between papers (Figure 2).

3.1 High-Quality Paper Identification

Defining “high quality” is challenging and often subjective. We adopt a simple, defensible approach: a paper is considered high-quality if it was accepted for an **Oral or Spotlight presentation** at NeurIPS, ICML, or ICLR. This criterion uses the explicit judgment of conference program committees—comprising leading researchers in the field—as a direct proxy for a paper’s significance and novelty at the time of publication.¹

This approach offers several advantages: **(1) Community-validated:** The selection reflects collective expert assessment rather than individual opinion or post-hoc metrics. **(2) Signal strength:** Oral/Spotlight papers represent the top 1-5% of submissions, providing a strong quality signal. **(3) Contemporaneous:** Unlike citation counts which accumulate over years, this signal is available immediately and reflects perceived quality at publication time. **(4) Reproducible:** The criterion is objective and easily replicated by other researchers.

We apply this criterion to Oral and Spotlight presentations from NeurIPS, ICML, and ICLR, (2023-2025). This results in 3,819 high-quality papers that serve as targets for lineage tracing. This scale allows us to capture diverse methodological approaches and intellectual trajectories across the major machine learning conferences while maintaining quality through our LLM-accelerated pipeline with human validation for quality assurance.

¹Fully reproducible workflow: <https://www.orchestra-research.com/share/I0RLtZw08Q2lUkFH>

Andes: Defining and Enhancing Quality-of-Experience (QoE) in LLM-Based Text Streaming Services

Andes introduces a token-level, preemptive LLM serving framework that formalizes QoE and improves user-perceived latency and smoothness in streaming text generation.

INTELLECTUAL PREDECESSORS

BASELINE Efficient Memory Management for LLM Serving with PagedAttention (vLLM)

"This system's KV cache management and throughput-oriented scheduling provide the serving baseline..."
Adopted: Continuous batching and KV cache management foundation; departed from throughput/latency proxies toward token-granular QoE prioritization.

BASELINE SGLang: Efficient Execution Engine for Structured Language Model Programs

"SGLang's optimized prefill/decode execution and batching policies serve as a high-performance baseline..."
Adopted: High-performance batching/caching policies; motivated shift to QoE-aware token-level scheduling.

QOE INSPIRATION Neural Adaptive Video Streaming with Pensieve

"Pensieve's explicit QoE formulation balancing startup delay, rebuffering, and smoothness inspired Andes..."
Adopted: QoE definition for text streaming—first-token promptness and digestible, smooth token pace.

SMOOTHNESS INSPIRATION BOLA: Near-Optimal Bitrate Adaptation for Online Videos

"BOLA's marginal-utility view of segment choices and emphasis on smoothness informed Andes..."
Adopted: Per-token marginal QoE gain; avoiding bursty, hard-to-digest output rates.

GAP IDENTIFICATION

InferLine: ML Inference Pipeline Composition with E2E Latency SLOs

"InferLine's SLO-centric scheduling underscored the limitation of meeting latency targets without modeling user-perceived utility..."
Adopted: Explicit QoE objective addressing gap between SLO metrics and user experience.

SYNTHESIS NARRATIVE

Throughput-oriented LLM serving systems like vLLM introduced continuous batching and memory-efficient PagedAttention to maximize tokens-per-second, and SGLang further streamlined prefill/decode execution with high-performance batching and caching policies. These systems excel at raw efficiency but optimize proxy metrics rather than modeling how users experience streamed text. In contrast, the adaptive bitrate (ABR) literature explicitly defined and optimized user-centric Quality-of-Experience (QoE). Pensieve formalized a QoE function combining startup delay, rebuffering, and smoothness, and learned policies that trade off early start versus consistent playback; BOLA framed bitrate selection via marginal utility and emphasized avoiding burstiness that harms perception. Meanwhile, inference pipeline schedulers such as InferLine focused on meeting end-to-end latency SLOs across models, revealing a gap between hitting deadlines and optimizing perceived utility over an interaction. Taken together, these strands highlighted an opportunity: import ABR-style QoE modeling into token-streamed LLM interactions and couple it with preemptive, fine-grained scheduling. Building on high-throughput batching engines, a natural next step is to prioritize tokens by expected marginal QoE gain per unit GPU time, ensuring fast first tokens and smooth, digestible pacing under load rather than maximizing throughput.

THINKING TRAJECTORY

STARTING POINT
High-throughput LLM serving (vLLM, SGLang): continuous batching with efficient memory/execution—optimizes tokens-per-second.

GAP IDENTIFIED
Server-centric metrics (throughput, latency SLOs) don't model user-perceived utility over full interaction timeline.

CROSS-DOMAIN INSIGHT
Video ABR (Pensieve, BOLA) explicitly optimizes QoE: startup delay, smoothness, marginal utility per segment.

REFRAMING
Text streaming as QoE optimization: first-token promptness + smooth, digestible pace—GPU-bound not network-bound.

INNOVATION
Andes: preemptive token scheduler prioritizing by QoE gain/GPU cost .4.7x QoE or 61% GPU savings.

RELATIONSHIP GRAPH

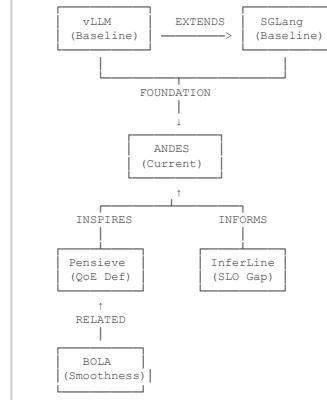


Figure 2: One complete dataset entry in Sci-Reasoning. For a target paper, our LLM pipeline identifies key predecessors (§3.2) and generates structured metadata (predecessor role, relationship type) and rich synthesis narratives (§3.3). See Appendix A for the complete JSON structure.

3.2 Intellectual Lineage Tracing

For each high-quality target paper, we identify 5-10 key predecessors that form its intellectual foundation. This range reflects our observation that most breakthroughs synthesize insights from multiple sources rather than extending a single prior work.

We employ a single-pass LLM analysis using GPT-5 to process the full text of each target paper. The LLM is prompted to: (1) parse all internal citations and link them to the bibliography, (2) analyze the linguistic context surrounding each citation (e.g., “building upon,” “inspired by,” “addresses the limitation of”), (3) count citation frequency across sections, (4) synthesize this evidence to rank cited works by importance, and (5) select the top 5-10 most significant predecessors that form the intellectual foundation of the target paper, ensuring diversity across contribution types (e.g., methodology, problem formulation, and baselines) rather than se-

lecting only methodologically similar works. The complete prompt is provided in Appendix B.

This consolidated approach leverages LLMs’ ability to perform holistic, end-to-end reasoning over long documents, reducing pipeline complexity compared to multi-step traditional NLP approaches (Cohan et al., 2019). The automated diversity-aware selection ensures systematic coverage of different predecessor roles while maintaining scalability across our dataset of 3,819 papers.

3.3 Intellectual Connection Synthesis

The core contribution of our methodology is representing intellectual lineage as structured “Lineage Graphs” that combine rich natural language narratives with queryable annotations. Each edge in the Lineage Graph connects a source (predecessor) paper to a target (current) paper, annotated across

multiple dimensions²:

Predecessor Role and Relationship Type:

Each connection is characterized along two complementary dimensions using LLM-based structured annotation. First, the model annotates the *predecessor role*—what function the source paper serves in the intellectual lineage, such as providing methodological foundations, theoretical concepts, benchmarks for comparison, problem formulations, necessary resources, or inspiration from related domains. Second, the model captures the *relationship type*—how the target paper builds upon the source, whether by extending, combining, bridging, addressing limitations of, or reframing ideas from the predecessor. Multiple relationship types may exist between the same paper pair.

Synthesis Narrative: An LLM-generated paragraph (200-400 words) explaining the intellectual synthesis that occurred. Each narrative follows a two-part structure: first, establishing the context by discussing the prior work and its key contributions relevant to the current paper; second, synthesizing how this prior work inspires and enables the target paper’s contribution. This structure captures the “story” of how the target paper combined insights from its predecessors, identifying specific *intellectual moves*, *gaps identified*, and *insight types* that enabled the scientific contribution. The unified predecessor identification and synthesis prompt (Appendix B) guides the LLM to identify key predecessors and reconstruct the authors’ thinking trajectory as a narrative of intellectual discovery.

The complete schema is detailed in Appendix C. This structured representation makes the dataset queryable and analyzable at scale, while the natural language narratives provide the rich context needed to understand scientific reasoning.

3.4 Quality Validation

Our methodology employs a fully automated LLM pipeline for scalability, with human validation to ensure quality. A critical enabler of this approach is frontier LLMs’ demonstrated capability to comprehend ~ 15 page papers and extract nuanced intellectual relationships—a task requiring human-level understanding of implicit reasoning and contextual synthesis that no existing automated method (citation analysis, topic modeling) systematically captures at scale. We validate the pipeline using 30 papers from domains we are familiar with, where

²Fully reproducible workflow: <https://www.orchestra-research.com/share/oca4ADYe7gNmgh1T>

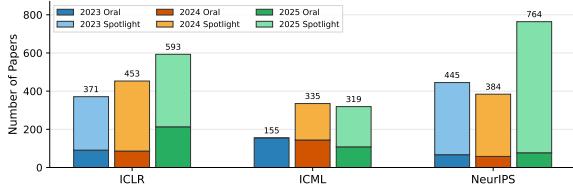


Figure 3: Distribution of papers in the Sci-Reasoning dataset across conferences, years, and presentation types (Oral and Spotlight).

we have ground-truth knowledge of predecessors and relationships. We iteratively refine all pipeline components—LLM prompts for candidate generation, diversity selection criteria, and structured annotation generation (Section 3.3)—until GPT-5 outputs consistently match ground truth. We select GPT-5 as our primary model based on an ablation study (Section 5.2) that demonstrates it achieves optimal cost-quality balance for predecessor extraction. Once validated, we apply the automated pipeline to all 3,819 papers.

For quality assurance at scale, we employ a two-tier validation strategy. For low-confidence cases (GPT-5 self-reported scores <0.7 in structured output), we employ multi-model cross-validation with Claude Opus 4.5 and Google Gemini 3.0. High agreement across models confirms reliability; disagreements trigger manual expert review. This human-in-the-loop approach concentrates expert attention on genuinely ambiguous cases (approximately 3% of papers) while leveraging the complementary strengths of frontier LLMs for the majority of cases.

4 Patterns Analysis of AI Research Innovation

We analyze the Sci-Reasoning dataset (Figure 3; see Appendix E.1 for collection criteria) to uncover recurring patterns of innovation in high-quality AI research. By systematically examining the synthesis narratives of 3,819 Oral and Spotlight papers from NeurIPS, ICML, and ICLR (2023-2025), we identify 15 distinct **thinking patterns**—the cognitive strategies researchers employ to develop breakthrough ideas (complete descriptions in Appendix F). These patterns represent not merely taxonomic categories, but actionable frameworks for systematic research ideation. Our analysis reveals how successful researchers diagnose gaps, reframe problems, synthesize cross-domain insights, and validate novel contributions.

4.1 Thinking Pattern Taxonomy Derivation

To identify recurring cognitive strategies in breakthrough research, we develop a systematic taxonomy of thinking patterns through iterative LLM-based discovery and consolidation.³

Phase 1: Pattern Discovery. We employ stratified sampling (Patton, 2001) across conferences (NeurIPS, ICML, ICLR), years (2023–2025), and presentation types (Oral, Spotlight), selecting 10 batches of 35 papers each (350 total). For each batch, an LLM analyzes the synthesis narratives to identify recurring intellectual moves—the cognitive strategies authors use to develop contributions (detailed prompts in Appendix B). This batch-wise process yields approximately 190 raw patterns, intentionally allowing redundancy to capture patterns at different abstraction levels.

Phase 2: Taxonomy Consolidation. An LLM consolidates the 190 raw patterns into 15 canonical thinking patterns by: (1) identifying semantically similar patterns across batches and (2) merging patterns at different abstraction levels (consolidation prompts in Appendix B). Each canonical pattern includes: a descriptive name, explanation of the cognitive move, and illustrative examples.

Phase 3: Full Classification and Analysis. Using the 15-pattern taxonomy, we classify all 3,819 papers, processing synthesis narratives in batches of 5 to assign both a primary pattern (dominant strategy) and secondary pattern (supporting strategy, if present). The classified dataset enables systematic analysis of pattern distributions, temporal trends, conference-specific preferences, and pattern co-occurrence, identifying dominant patterns and successful pattern combinations.

4.2 Dominant Thinking Patterns

Three patterns dominate the landscape, accounting for 52.7% of all analyzed papers, while the distribution follows a power law with a long tail of specialized strategies (Figure 4).

Gap-Driven Reframing (24.2%) converts limitations into design constraints that guide solution design. Rather than treating failures as obstacles, researchers reframe them as specifications for better-suited approaches—e.g., reframing autoregressive image modeling from next-token to

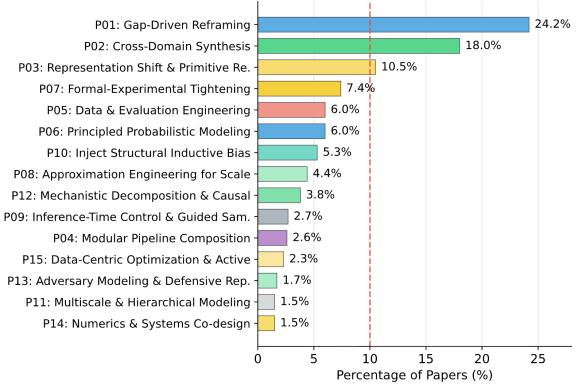


Figure 4: Distribution of the 15 identified thinking patterns across 3,819 papers. The top three patterns—Gap-Driven Reframing, Cross-Domain Synthesis, and Representation Shift—account for 52.7% of all papers.

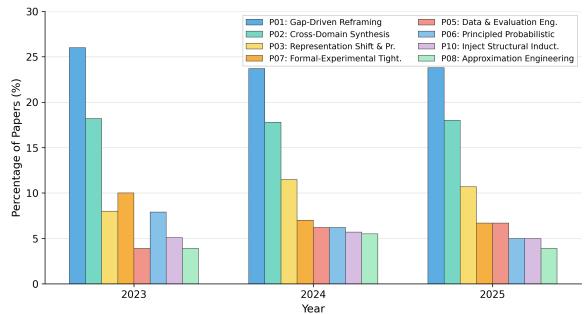


Figure 5: Temporal evolution of the top 5 thinking patterns from 2023 to 2025. While Gap-Driven Reframing remains stable, Representation Shift peaked in 2024, Formal-Experimental approaches are declining, and Data/Evaluation engineering is rising.

next-scale prediction transforms a scalability limitation into a principled architectural choice.

Cross-Domain Synthesis (18.0%) transplants solutions from adjacent fields by identifying abstract constraints and engineering compatibility layers. Breakthroughs emerge from recognizing that analogous solutions exist elsewhere, such as fusing quantum circuits with transformer attention (Born et al., 2025) or importing control-theoretic stability into reinforcement learning.

Representation Shift (10.5%) replaces a problem’s fundamental primitives—pixels, tokens, meshes—with alternatives that simplify inference or better capture constraints, such as replacing explicit meshes with neural implicit functions for 3D reconstruction. Together, these patterns reveal a consistent meta-strategy: reframe problems, import cross-domain solutions, or reconceptualize representations (detailed case studies in Appendix D).

³Fully reproducible workflow: <https://www.orchestra-research.com/share/BsrSD96SQyXjxZV7>

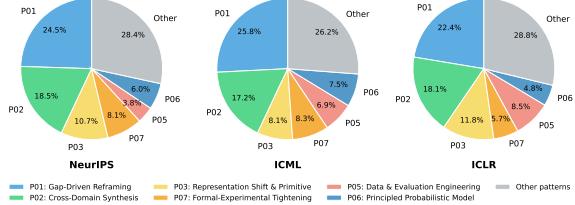


Figure 6: Conference-specific distribution of thinking patterns. ICML shows stronger preference for formal methods (8.3%) and probabilistic modeling (7.5%), ICLR emphasizes representation innovation (11.8%) and benchmarking (8.5%), while NeurIPS maintains balanced, cross-disciplinary coverage.

4.3 Temporal Evolution of Patterns

Temporal trends reveal that problem diagnosis and reformulation remain the fundamental engine of innovation (Figure 5). The decline in Formal-Experimental Tightening suggests theoretical analysis is becoming a supporting element rather than a primary one, while growth in Data/Evaluation Engineering reflects the field’s maturation toward rigorous empirical methodology.

4.4 Conference-Specific Patterns

Beyond their shared emphasis on Gap-Driven Reframing, the three major conferences exhibit distinct intellectual cultures (Figure 6). These differences suggest tailoring submission strategies: ICML submissions benefit from mathematical rigor and theoretical guarantees, ICLR submissions should highlight architectural innovations and empirical methodology, while NeurIPS favors broad applicability and cross-disciplinary synthesis.

4.5 Research Pattern Combinations

Research breakthroughs rarely employ a single thinking pattern in isolation. Analysis of secondary patterns reveals systematic combinations that function as repeatable “research recipes” (Figure 7). The most frequent combination pairs Gap-Driven Reframing with Representation Shift (318 occurrences), embodying a powerful two-step strategy: diagnose a limitation, then introduce a new primitive that sidesteps it. This “Reframe + New Primitive” recipe transforms conceptual insights into concrete architectural innovations.

The second most common combination, Cross-Domain Synthesis with Representation Shift (233 occurrences), represents an “Import + Adapt” strategy: borrow a mechanism from another field and

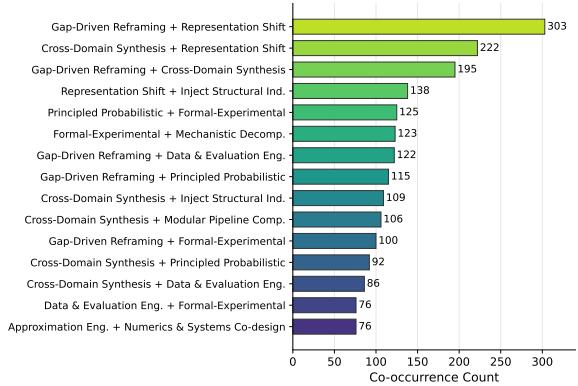


Figure 7: Top 10 pattern combinations (primary + secondary). The most successful research employs multiple thinking patterns: “Reframe + New Primitive” (318), “Import + Adapt” (233), and “Diagnose + Borrow” (204) represent repeatable innovation recipes.

modify its representation to fit the target domain. The third combination, Gap-Driven Reframing with Cross-Domain Synthesis (204 occurrences), embodies “Diagnose + Borrow”: identify what’s missing, then search for solutions in adjacent fields.

These patterns suggest that breakthrough research follows a meta-pattern: *Diagnose* a gap, *Represent* it differently or *Import* from elsewhere, then *Validate* rigorously. For detailed statistical analyses for all above analyses and practical guidance, see Appendix E.

Finally, we conduct an assessment of whether LLMs can predict research directions from intellectual predecessors (Section 5). Testing four frontier models on 77 NeurIPS 2025 Oral papers, we find that Gemini 2.5 Pro achieves 49.35% Hit@10 accuracy, demonstrating that our methodology captures meaningful intellectual relationships while also revealing a natural predictability ceiling that validates genuine research creativity.

5 Evaluation

5.1 Evaluating LLM on Research Ideation

We evaluate whether LLMs can generate research ideas matching real publications given only their intellectual predecessors.⁴ Our pipeline takes intellectual predecessor papers from each NeurIPS 2025 Oral paper, retrieves the prior work paper content using Exa AI (Exa AI, 2024), prompts an LLM to generate $k = 10$ candidate research ideas, and evaluates semantic similarity between generated

⁴Fully reproducible workflow: <https://www.orchestra-research.com/share/xkueih5uZdNu4Lb8>

Model	Hit@10 (%)
Gemini 2.5 Pro	49.35
Claude Opus 4	42.86
GPT-5.2	38.89
Claude Sonnet 4	29.87

Table 1: Hit@10 rates on predicting research directions.

ideas and the actual published paper using an LLM judge.

We employ GPT-5.2 as an evaluator (Zheng et al., 2023) to assess whether each generated idea semantically matches the ground truth paper. For each paper, the judge receives both the generated ideas and the actual paper’s title and contribution statement, then determines if they represent similar research directions by considering: (1) whether they address the same core problem, (2) whether they propose similar methodological approaches, and (3) whether the generated idea, if fully developed, would result in a similar contribution. The judge returns a structured JSON response containing a binary match decision, a confidence score (0-1), and a brief reasoning explanation. The complete evaluation prompt is provided in Appendix B.4. We measure success using Hit@10: whether any of the 10 generated ideas matches the ground truth paper according to the LLM judge. We evaluate on 77 NeurIPS 2025 Oral papers, using four frontier models: GPT-5.2, Claude Sonnet 4, Claude Opus 4, and Gemini 2.5 Pro.

Table 1 presents our main findings. Gemini 2.5 Pro achieves the highest Hit@10 rate of 49.35%, successfully predicting nearly half of research directions from intellectual predecessors alone. The 19.5 percentage point performance gap demonstrates meaningful differences in research direction prediction capabilities across frontier models. Research ideation exhibits a many-to-many relationship: the same intellectual predecessors can inspire multiple valid research directions, while different predecessor combinations may lead to similar ideas. This means our Hit@10 metric likely provides a conservative estimate of model capabilities, as generated ideas that don’t match the published paper may still represent valid research directions.

5.2 Evaluating LLM on Predecessor Extraction

To ensure our pipeline balances cost and quality, we conduct an ablation study evaluating four OpenAI models on the predecessor extraction task using

Model	Recall
GPT-5	89.73%
GPT-5.2	87.47%
GPT-4.1	78.00%
GPT-5-mini	68.53%

Table 2: Model ablation for predecessor extraction.

a validation set of 77 Oral/Spotlight papers with ground-truth predecessors.

Table 2 shows that GPT-5 achieves the highest recall, outperforming even the newer GPT-5.2. This surprising result suggests that model generation alone does not guarantee better performance on specialized academic tasks. GPT-4.1 shows consistent performance with zero complete failures but exhibits a substantial quality gap. GPT-5-mini demonstrates the largest capability gap with frequent failures on specialized theoretical papers, indicating insufficient domain knowledge for academic tasks. Based on these results, we select GPT-5 as our primary model for dataset construction, achieving the optimal cost-quality balance with superior performance and lower API costs.

6 Limitations

We acknowledge several important limitations of our work that should be considered when interpreting results and planning future research.

Logic of Justification vs. Discovery: We analyze the *logic of justification*—how researchers present contributions in published papers—rather than the *logic of discovery*—the actual thought process behind breakthroughs (Reichenbach, 1938). Published papers are polished narratives that may omit failed experiments, abandoned hypotheses, serendipity, or external influences. Our dataset captures researchers’ final presentations, which may differ from the chronological reality of how insights emerged. This inherent constraint means only successful reasoning paths can be systematically analyzed; we make this explicit to avoid overstating claims about understanding creativity.

Temporal Constraint: Our dataset represents a snapshot of research patterns in 2023-2025. Innovation patterns may evolve as fields mature, new methodologies emerge, or community norms change. Longitudinal analysis tracking how patterns shift over decades would be valuable but is beyond our current scope.

Conference and Selection Criteria Scope: Restricting our dataset to Oral/Spotlight presentations

at NeurIPS, ICML, and ICLR may favor empirical breakthroughs over theoretical work, reflecting conference-specific biases rather than long-term impact. This focus misses other AI venues (AAAI, IJCAI, UAI), interdisciplinary research (computational biology, neuroscience), and non-ML AI areas (symbolic AI, knowledge representation). Our innovation patterns may be specific to mainstream ML circa 2023-2025 and may not generalize to other domains or AI subfields.

Despite these limitations, we believe Sci-Reasoning makes a valuable contribution by providing the first large-scale, structured dataset for studying scientific reasoning in AI research, demonstrating its reliability through quality evaluation, and revealing actionable patterns of innovation.

7 Conclusion

We present Sci-Reasoning, a dataset capturing structured intellectual lineage behind scientific breakthroughs in top-tier AI research. By transforming implicit reasoning into queryable lineage graphs, our work reveals systematic patterns in how researchers diagnose gaps, synthesize cross-domain insights, and combine ideas into novel contributions. These patterns provide both scientific understanding of innovation mechanisms and practical frameworks for AI research agents. By making scientific reasoning explicit, our work advances the augmentation of scientific discovery.

References

- Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.
- Jannis Born, Filip Skogh, Kahn Rhrissorrakrai, Filippo Utro, Nico Wagner, and Aleksandros Sobczyk. 2025. **Quantum doubly stochastic transformers**. *Preprint*, arXiv:2504.16275.
- Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80.
- Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596.
- Miles Cranmer and 1 others. 2024. The polymathic ai project: Laying the foundations for a new era of scientific machine learning. *Machine Learning: Science and Technology*, 5(4):045043.
- Exa AI. 2024. Exa: Neural search for the age of ai. <https://exa.ai>.
- Eugene Garfield. 1955. Citation indexes for science. *Science*, 122(3159):108–111.
- Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Sriparna Verma. 2021. Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies*, 2(4):1511–1542.
- Melanie C Green and Jay K Donahue. 2017. On the usefulness of narratives: An interdisciplinary review and theoretical model. *Annals of the International Communication Association*, 41(3-4):227–245.
- Myriam Hernández-Álvarez and José Manuel Gómez. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(3):327–349.
- Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, and Young-Kyoong Suh. 2025. Finecite: A novel approach for fine-grained citation context analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24525–24542, Vienna, Austria.
- Woo Seong Jo, Lu Liu, and Dashun Wang. 2022. See further upon the giants: Quantifying intellectual lineage in science. *Quantitative Science Studies*, 3(2):319–330.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661.
- Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivas, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025a. Curie: Toward rigorous and automated scientific experimentation with ai agents. <https://arxiv.org/abs/2502.16069>. *Preprint*, arXiv:2502.16069.
- Patrick Tser Jern Kon, Jiachen Liu, Xinyi Zhu, Qiuyi Ding, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivas, Myungjin Lee, Mosharaf Chowdhury, Matei Zaharia, and Ang Chen. 2025b. Exp-bench: Can ai conduct ai research experiments? *Preprint*, arXiv:2505.24785.
- Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries*, pages 51–55. ACM.
- John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- Jiachen Liu, Jae-Won Chung, Zhiyu Wu, Fan Lai, Myungjin Lee, and Mosharaf Chowdhury. 2024. Andes: Defining and enhancing quality-of-experience in llm-based text streaming services. *Preprint*, arXiv:2404.16283.
- Mary S Morgan and M Norton Wise. 2017. Narrative science. In *The Cambridge History of Science: Volume 7, The Modern Social Sciences*, pages 605–620. Cambridge University Press.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Michael Quinn Patton. 2001. *Qualitative research and evaluation methods*, 3 edition. Sage Publications, Thousand Oaks, CA.
- Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025. The nature of nlp: Analyzing contributions in nlp papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25169–25191, Vienna, Austria.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Hans Reichenbach. 1938. *Experience and prediction: An analysis of the foundations and the structure of knowledge*. University of Chicago Press.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. Pst-bench: Tracing and benchmarking the source of publications. *arXiv preprint arXiv:2402.16009*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17744–17761, Suzhou, China.

Matthäus Zloch, Danilo Dessì, Jennifer D’Souza, Leyla Jael Castro, Benjamin Zapilko, Saurav Karimakar, Brigitte Mathiak, Markus Stocker, Wolfgang Otto, Sören Auer, and Stefan Dietze. 2025. Research knowledge graphs: the shifting paradigm of scholarly information representation. In *Proceedings of the Extended Semantic Web Conference*.

A Complete Dataset Example

This section provides a complete example of a dataset entry, showing the full annotation for a single target paper with all intellectual predecessors, thinking trajectory, synthesis narrative, and relationship graph. The example uses the Andes paper (Liu et al., 2024), which introduces a token-level, preemptive LLM serving framework that formalizes QoE and improves user-perceived latency and smoothness in streaming text generation.

Target Paper

Andes: Defining and Enhancing Quality-of-Experience (QoE) in LLM-Based Text Streaming Services

Andes introduces a token-level, preemptive LLM serving framework that formalizes QoE and improves user-perceived latency and smoothness in streaming text generation.

Intellectual Predecessors

BASELINE: Efficient Memory Management for LLM Serving with PagedAttention (vLLM) “*This system’s KV cache management and throughput-oriented scheduling provide the serving baseline...*”

Adopted: Continuous batching and KV cache management foundation; departed from throughput/latency proxies toward token-granular QoE prioritization.

BASELINE: SGLang: Efficient Execution Engine for Structured Language Model Programs “*SGLang’s optimized prefill/decode execution and batching policies serve as a high-performance baseline...*”

Adopted: High-performance batching/caching policies; motivated shift to QoE-aware token-level scheduling.

QOE INSPIRATION: Neural Adaptive Video Streaming with Pensieve “*Pensieve’s explicit QoE formulation balancing startup delay, rebuffering, and smoothness inspired Andes...*”

Adopted: QoE definition for text streaming—first-token promptness and digestible, smooth token pace.

SMOOTHNESS INSPIRATION: BOLA: Near-Optimal Bitrate Adaptation for Online Videos “*BOLA’s marginal-utility view of segment choices and emphasis on smoothness informed Andes...*”

Adopted: Per-token marginal QoE gain; avoiding bursty, hard-to-digest output rates.

GAP IDENTIFICATION: InferLine: ML Inference Pipeline Composition with E2E Latency SLOs “*InferLine’s SLO-centric scheduling underscored the limitation of meeting latency targets without modeling user-perceived utility...*”

Adopted: Explicit QoE objective addressing gap between SLO metrics and user experience.

Thinking Trajectory

STARTING POINT High-throughput LLM serving (vLLM, SGLang): continuous batching with efficient memory/execution—optimizes tokens-per-second, not user timeline.

GAP IDENTIFIED Server-centric metrics (throughput, latency SLOs) don’t model user-perceived utility over full interaction timeline.

CROSS-DOMAIN INSIGHT Video ABR (Pensieve, BOLA) explicitly optimizes QoE: startup delay, smoothness, marginal utility per segment.

REFRAMING Text streaming as QoE optimization: first-token promptness + smooth, digestible pace—GPU-bound not network-bound.

INNOVATION Andes: preemptive token scheduler prioritizing by QoE gain/GPU cost. 4.7× QoE or 61% GPU savings.

Synthesis Narrative

Throughput-oriented LLM serving systems like vLLM introduced continuous batching and memory-efficient PagedAttention to maximize tokens-per-second, and SGLang further streamlined prefill/decode execution with high-performance batching and caching policies. These systems excel at raw efficiency but optimize proxy metrics rather than modeling how users experience streamed text. In contrast, the adaptive bitrate (ABR) literature explicitly defined and optimized user-centric Quality-of-Experience (QoE): Pensieve formalized a QoE function combining startup delay, rebuffering, and smoothness, and learned policies that trade off early start versus consistent playback. BOLA framed bitrate selection via marginal utility and emphasized avoiding burstiness that harms perception. Meanwhile, inference

pipeline schedulers such as InferLine focused on meeting end-to-end latency SLOs across models, revealing a gap between hitting deadlines and optimizing perceived utility over an interaction. Taken together, these strands highlighted an opportunity: import ABR-style QoE modeling into token-streamed LLM interactions and couple it with preemptive, fine-grained scheduling. Building on high-throughput batching engines, a natural next step is to prioritize tokens by expected marginal QoE gain per unit GPU time, ensuring fast first tokens and smooth, digestible pacing under load rather than maximizing throughput.

Relationship Graph

The relationship graph illustrates how the intellectual predecessors combine to form the innovation:

- **vLLM (Baseline)** and **SGLang (Baseline)** → EXTENDS → **ANDES (Current)**: Foundation of efficient batching and execution.
- **Pensieve (QoE Def)** → INSPIRES → **ANDES**: QoE formulation for text streaming.
- **InferLine (SLO Gap)** → INFORMS → **ANDES**: Highlights gap between SLO targets and user experience.
- **BOLA (Smoothness)** → RELATED → **Pensieve**: Both contribute to QoE modeling in video streaming.

B LLM Pipeline Prompts

This section provides the detailed system prompts used in our LLM pipeline. We use a unified prompt that combines predecessor identification with synthesis narrative generation, ensuring that the identified predecessors and their relationships are coherently integrated into the intellectual lineage story. These prompts are incorporated into our automated pipeline using GPT-5.

B.1 Unified Predecessor Identification and Synthesis Prompt

Purpose: Identify key intellectual predecessors and generate synthesis narratives that explain how they collectively inspired the target paper’s innovation.

System Prompt:

You are an expert AI research analyst. Your task is to identify the KEY PRIOR WORKS that DIRECTLY led to a research paper’s core innovation.

CRITICAL: Focus on DIRECT Intellectual Lineage You must identify papers that are **directly responsible** for the current paper’s main contributions. Ask yourself:

- “Without this prior work, would the current paper’s core idea exist?”
- “Did this prior work directly inspire, enable, or motivate the KEY INNOVATION?”
- “Is this paper cited in the Introduction or Related Work as a PRIMARY influence?”

DO NOT INCLUDE:

- Generic infrastructure/tools (e.g., PyTorch, CUDA, standard attention mechanisms)
- Complementary optimizations that are orthogonal to the main contribution
- Papers that share the same domain but don’t directly influence the core idea
- Standard baselines that are just compared against without deeper connection
- Well-known foundational works that everyone cites but aren’t specific to this innovation

DO INCLUDE:

- Papers whose specific IDEAS, METHODS, or FINDINGS directly shaped the current work
- Papers whose LIMITATIONS or GAPS the current paper explicitly addresses
- Papers that introduced the PROBLEM FORMULATION the current paper builds on
- Papers whose TECHNIQUES are directly extended or modified
- Papers that provide the KEY INSIGHT that the current paper leverages

Role Classifications (assign ONE per paper):

1. **Baseline:** The primary system/method this paper improves upon or compares against as its main competitor
2. **Inspiration:** Paper whose specific idea/approach directly sparked the current paper's key innovation
3. **Gap Identification:** Paper whose explicit limitations failures motivated this research direction
4. **Foundation:** Paper that introduced the core problem formulation, dataset, or theoretical framework used
5. **Extension:** Paper whose specific method is directly extended, modified, or generalized
6. **Related Problem:** Paper solving a closely related problem whose solution approach informed this work

Output Requirements: For each prior work (identify 5-7 papers, quality over quantity):

1. **Role:** One of the six classifications above
2. **Relationship Sentence:** ONE specific sentence explaining the DIRECT connection to the current paper's innovation. Be concrete about WHAT was borrowed/extended/addressed.

Synthesis Narrative (200-300 words): Write a cohesive narrative that flows naturally (NO explicit "Part 1" / "Part 2" labels):

First ~150 words - Prior Work with Relevant Details: Describe each prior work, but FOCUS ONLY on the specific aspects/details that relate to the current paper's innovation. For each prior work, highlight:

- The specific technique, insight, or finding that is relevant (not a general summary)
- How this specific detail connects to what the current paper does

- Do NOT mention the current paper yet—just establish what relevant knowledge existed

Remaining ~100 words - How They Collectively Inspired Current Work: Transition naturally to explain:

- What gap or opportunity emerged from the combination of these prior works
- How the current paper synthesizes or builds upon these specific relevant details
- Why this was a natural next step given the prior work landscape

The narrative should read as one flowing paragraph, not two separate sections.

User Prompt:

Analyze this research paper and identify the prior works that DIRECTLY led to its core innovation.

TASK: Identify 5-7 prior works that DIRECTLY influenced this paper's KEY CONTRIBUTION.

Focus on papers that:

1. Introduced ideas/methods this paper directly builds on
2. Had limitations this paper explicitly addresses
3. Defined the problem formulation used here
4. Are the primary baselines being improved upon

DO NOT include generic tools, orthogonal optimizations, or tangentially related work.

Return your analysis as valid JSON with the following structure:

```
{
  "prior_works": [
    {
      "title": "Exact paper title",
      "authors": "First author et al.",
      "year": 2023,
      "arxiv_id": "if known",
      "role": "One of six roles",
      "relationship_sentence":
        "Specific sentence about DIRECT connection"
    }
  ]
}
```

```

        }
    ],
    "synthesis_narrative":
      "200-300 word flowing narrative"
}

```

Remember: Every paper you include should pass the test: “This paper DIRECTLY influenced the core innovation, not just the general research area.”

B.2 Multi-Round Pattern Discovery Prompt

Purpose: Discover patterns across multiple sampling rounds.

System Prompt:

You are an expert analyst of scientific innovation patterns. Always respond with valid JSON.

User Prompt Key Instructions:

I have [N] synthesis narratives from top ML conference papers (ICML, ICLR, NeurIPS). Each narrative describes how authors built their novel contribution on prior work.

YOUR TASK: Identify the THINKING PATTERNS used by these researchers.

IMPORTANT:

- Look for COGNITIVE/REASONING patterns, not topic categories
- Examples: “Cross-domain analogy”, “Constraint relaxation”, “Theoretical unification”, “Problem reframing”, “Empirical observation leading to theory”, “Modular decomposition”
- Be specific and descriptive
- Discover patterns inductively—don’t limit to predefined categories

B.3 Pattern Consolidation Prompt

Purpose: Consolidate discovered patterns into clean taxonomy.

System Prompt:

You are an expert in categorizing research innovation patterns. Create a rigorous taxonomy.

User Prompt Key Instructions:

YOUR TASK: Consolidate these into a CLEAN TAXONOMY of 10-12 distinct, non-overlapping pattern categories.

Guidelines:

- Patterns that appear in more rounds are more robust/reliable
- Merge semantically similar patterns
- Each category should represent a distinct cognitive strategy
- Avoid topic-based categories (focus on HOW researchers think, not WHAT they study)

B.4 LLM Judge Evaluation Prompt

Purpose: Evaluate whether a generated research idea semantically matches a real published paper.

System Prompt:

You are evaluating whether a generated research idea matches a real published paper.

GENERATED IDEA: Title: {generated_idea.title} Description: {generated_idea.description}

REAL PUBLISHED PAPER: Title: {ground_truth_title} Contribution: {ground_truth_contribution}

Determine if the generated idea is semantically similar to the real paper. Consider:

1. Do they address the same core problem or research question?
2. Do they propose similar methodological approaches?
3. Would the generated idea, if fully developed, result in a similar contribution?

A match means the ideas are substantially aligned in their core direction, not necessarily identical in every detail.

Respond with a JSON object containing:

- “is_match”: true or false
- “confidence”: a number from 0 to 1
- “reasoning”: a brief explanation (2-3 sentences)

Output ONLY the JSON object, no other text.

C Annotation Schema and Guidelines

This section provides excerpts from our annotation schema, which defines the structured categories and decision rules used by our LLM pipeline to classify predecessor roles and relationship types. These guidelines are incorporated into the LLM prompts to ensure consistent, high-quality structured annotations across all 3,819 papers. Note that these predecessor role and relationship type classifications are used for constructing the intellectual lineage graphs; they do not affect the subsequent thinking pattern analysis, which is derived independently from the synthesis narratives and innovation trajectories.

C.1 Predecessor Role Definitions

KEY_METHODOLOGY_COMPONENT: The source paper provides a specific technique, algorithm, model architecture, or methodological approach that is directly adopted or adapted by the target paper. This is concrete and technical.

Decision Rule: If the target paper's implementation directly uses or builds upon a specific method from the source, annotate as KEY_METHODOLOGY.

Example: If the target paper uses RLHF as its core alignment method, then the original RLHF paper is a KEY_METHODOLOGY_COMPONENT.

FOUNDATIONAL_CONCEPT: The source paper establishes theoretical grounding, introduces a fundamental concept, or provides the conceptual framework within which the target paper operates. This is more abstract than methodology.

Decision Rule: If removing this source would make the target paper's theoretical motivation unclear or unfounded, annotate as FOUNDATIONAL_CONCEPT.

Example: If the target paper is about improving transformer efficiency, the original "Attention is All You Need" paper providing the transformer architecture is FOUNDATIONAL.

PRIMARY_BASELINE: The source paper represents the state-of-the-art approach that the target paper aims to outperform or improve upon.

Decision Rule: If the target paper explicitly compares its results against the source as the main point of comparison, annotate as PRIMARY_BASELINE.

PROBLEM_FORMULATION: The source paper defines, formalizes, or establishes the problem that the target paper addresses.

Decision Rule: If the source introduced the task definition, dataset, or formal problem statement that the target paper tackles, annotate as PROBLEM_FORMULATION.

ENABLING_TOOL_OR_DATASET: The source provides a practical resource (codebase, dataset, benchmark) that the target paper uses.

Decision Rule: If the source's primary contribution is a resource rather than an idea or method, and the target uses that resource, annotate as ENABLING_TOOL.

INSPIRATION_BY_ANALOGY: The source paper comes from a different domain or problem area and provides inspiration through analogy or conceptual transfer.

Decision Rule: If the source addresses a fundamentally different problem but the target adapts its approach, annotate as INSPIRATION_BY_ANALOGY.

C.2 Relationship Type Definitions

EXTENDS: The target directly builds upon, generalizes, or scales the source's approach in a relatively straightforward way.

COMBINES_WITH: The target merges ideas from this source with ideas from other sources to create something new.

Decision Rule: If the innovation comes from the synthesis of multiple distinct approaches, mark all relevant sources with COMBINES_WITH.

BRIDGES_GAP_BETWEEN: The target connects this source with another, previously disconnected line of work.

ADDRESSES_LIMITATION_OF: The target explicitly identifies a shortcoming in the source and proposes a solution.

REFRAMES_USING: The target reinterprets or reconceptualizes an existing problem using ideas from the source.

D Additional Case Studies

This section presents additional synthesis graphs illustrating how different thinking patterns manifest in recent high-impact work.

D.1 Case Study: Flow Matching Generalization

Target Paper: *On the Closed-Form of Flow Matching: Generalization Does Not Arise from Target Stochasticity* (Bertrand et al., NeurIPS 2025 Oral)

Core Contribution: Demonstrates that stochasticity in conditional targets is not the primary driver

of generalization in flow matching. Shows that closed-form deterministic velocity targets match or improve performance, and that generalization instead arises from the neural network’s failure to perfectly approximate the optimal closed-form velocity field in specific time intervals.

Intellectual Predecessors This work synthesizes five distinct research lineages:

1. **Flow Matching for Generative Modeling** (Lipman et al., 2023) — KEY METHODOLOGY COMPONENT. Introduced the conditional flow matching objective and practical algorithms for learning time-dependent velocity fields that transport a simple prior into the data distribution. Provided the training loss, sampling ODE, and experimental setup that the target paper studies and modifies.
2. **Score-Based Generative Modeling** (Song et al., 2021) — FOUNDATIONAL CONCEPT. Established the SDE/score-matching perspective linking noisy conditional training objectives to continuous-time generative dynamics. This conceptual lens frames questions about the role of stochasticity in training targets and motivates comparing noisy sampled targets to deterministic closed-form targets.
3. **Foundational Flow/Transport Theory** (Albergo & Vanden-Eijnden, 2023) — KEY METHODOLOGY COMPONENT. Developed theoretical tools and variants of conditional flow/bridge formulations, including derivations yielding closed-form velocity fields. These results enable construction and computation of deterministic closed-form targets that replace sampled stochastic targets.
4. **Empirical Study of Memorization vs Generalization in Diffusion Models** (Kadkhodaie et al., 2024) — PROBLEM FORMULATION. Provided empirical evidence that diffusion models can either memorize or generalize depending on dataset size, model capacity, and regime. This empirical context motivated the central question: what mechanisms cause flow matching models to generalize in realistic regimes?
5. **Noisy Training-Loss Explanation for Generalization** (Vastola, 2025) — INSPIRATION

BY ANALOGY. Hypothesized that the stochastic nature of conditional training targets induces implicit regularization promoting generalization. The target paper tests and challenges this hypothesis specifically in the flow matching setting.

Key Inter-Predecessor Relationships

- Song et al.’s SDE/score viewpoint **bridges the gap between** diffusion and Lipman et al.’s flow matching, enabling comparisons of stochastic vs deterministic training targets.
- Albergo & Vanden-Eijnden’s closed-form/bridge theory **combines with** Lipman et al.’s flow-matching framework, producing concrete closed-form alternatives to sampled conditional targets.
- Kadkhodaie et al.’s empirical findings **address limitations of** Vastola’s hypothesis by exposing the need for mechanistic explanations of regime-dependent memorization/generalization.
- Vastola **reframes** successes of Lipman et al.’s methods through the lens of noisy training losses, prompting a targeted test within the flow matching setup.

Thinking Trajectory Starting point: Researchers worked within the flow-matching paradigm using conditional flow matching loss to learn time-dependent velocity fields.

Gap identified: Existing explanations for why these models generalize were incomplete. Empirical studies showed regime-dependent memorization, and a prominent hypothesis claimed that noisy/stochastic training targets provided implicit regularization—but this had not been tested in high-dimensional, realistic flow-matching settings.

Key insight: Closed-form optimal conditional velocity fields can be computed from transport/bridge theory, enabling direct testing of whether target stochasticity is essential for generalization by replacing sampled stochastic targets with deterministic closed-form targets.

Reframing: Rather than treating stochasticity as an inescapable feature of training, the authors reframed the question as a controlled comparison between stochastic (sampled) targets and their closed-form deterministic counterparts within the same flow-matching framework.

Thinking Pattern	%
Gap-Driven Reframing	24.2
Cross-Domain Synthesis	18.0
Representation Shift & Primitive Recasting	10.5
Formal-Experimental Tightening	6.7
Principled Probabilistic Modeling	5.4
Data & Evaluation Engineering	5.4
Inject Structural Inductive Bias	5.1
Approximation Engineering for Scalability	4.9
Modular Pipeline Composition	4.2
Inference-Time Control & Guided Sampling	2.4
Data-Centric Optimization & Active Sampling	2.1
Adversary Modeling & Defensive Repurposing	1.5
Multiscale & Hierarchical Modeling	1.4
Self-Supervised Pretext Engineering	1.9
Meta-Learning & Learning-to-Learn	1.5

Table 3: Complete distribution of thinking patterns across all 3,819 analyzed papers.

Novel combination: Combined Lipman et al.’s practical flow-matching training and architectures with Albergo & Vanden-Eijnden’s closed-form velocity expressions, and designed empirical protocols inspired by Kadkhodaie et al.’s regime analyses to evaluate Vastola’s noisy-loss hypothesis under realistic, high-dimensional conditions.

Resulting innovation: A conclusive empirical and diagnostic demonstration that stochastic target noise is not the key driver of generalization in flow matching. Instead, generalization arises when limited-capacity networks fail to approximate the optimal closed-form velocity field in specific time intervals, and closed-form targets can match or improve performance.

Patterns Exemplified This work demonstrates several thinking patterns:

- **Gap-Driven Reframing:** Identified the gap between empirical generalization behavior and mechanistic explanations, then reframed the question as a testable hypothesis about target stochasticity.
- **Cross-Domain Synthesis:** Combined insights from transport theory, score-based diffusion models, and empirical regime analyses to create a unified experimental framework.
- **Formal-Experimental Tightening:** Paired theoretical closed-form derivations with rigorous empirical validation across multiple datasets and model scales.

E Extended Statistical Analysis

This section provides additional statistical details from our analysis of 3,819 papers from NeurIPS, ICML, and ICLR (2023–2025).

E.1 Dataset Collection Notes

The paper counts in Figure 3 may differ from official conference statistics for two reasons. First, some papers were unavailable at the time of collection due to access restrictions, or missing files. Second, we focus specifically on research contributions and exclude certain paper types from our analysis, including benchmark papers, dataset papers, technical reports, position papers, and survey papers. These exclusions ensure that our analysis of thinking patterns reflects methodological and conceptual innovations rather than resource contributions or meta-analyses. The final dataset of 3,819 papers represents the intersection of papers that were both accessible and classified as research contributions.

E.2 Complete Pattern Frequency Distribution

Table 3 presents the complete distribution of all 15 thinking patterns identified in our taxonomy, ordered by frequency.

E.3 Pattern Co-occurrence Analysis

Figure 8 shows the co-occurrence heatmap revealing which thinking patterns frequently appear together. The strongest co-occurrences include:

- **Gap-Driven Reframing + Representation Shift** (318 co-occurrences): This “reframe+repr” combination represents the canonical path to breakthrough work—identifying a limitation and resolving it through primitive changes.
- **Cross-Domain Synthesis + Representation Shift** (233): Importing methods from other domains often requires adapting representations to the target setting.
- **Gap-Driven Reframing + Cross-Domain Synthesis** (204): Gaps are frequently addressed by borrowing solutions from adjacent fields.
- **Representation Shift + Inject Structural Inductive Bias** (145): New representations often incorporate domain-specific structure.

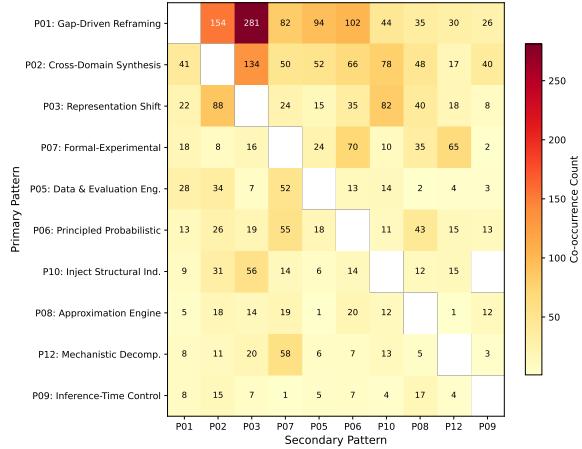


Figure 8: Pattern co-occurrence heatmap. Darker cells indicate more frequent co-occurrence. The diagonal represents single-pattern papers.

- **Principled Probabilistic Modeling + Formal-Experimental Tightening (131):** Theoretical work pairs naturally with rigorous validation.

E.4 Oral vs. Spotlight Presentation Analysis

Our dataset contains 999 oral presentations and 2,820 spotlight presentations. Figure 9 shows the pattern distribution by presentation type.

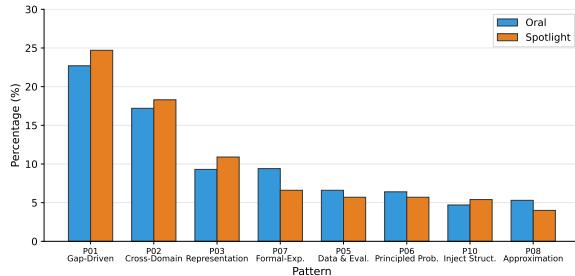


Figure 9: Distribution of thinking patterns across oral and spotlight presentations.

Key observations include:

- Papers receiving oral presentations show higher concentrations of Gap-Driven Reframing combined with Representation Shift—the “reframe+repr” recipe correlates with maximum visibility.
- Cross-Domain Synthesis with Modular Pipeline Composition (106 co-occurrences overall) appears frequently in oral presentations, suggesting that work demonstrating broad applicability receives higher recognition.

- The combination of Principled Probabilistic Modeling with Formal-Experimental Tightening predicts durable technical influence, appearing in papers that receive sustained citations.

E.5 Detailed Temporal Evolution

Beyond the year-over-year trends presented in the main text, we observe the following detailed temporal patterns:

Stable Patterns: Gap-Driven Reframing remains remarkably consistent (26.1% → 23.7% → 23.8% from 2023–2025), indicating a persistent research mode of “diagnose → reframe.”

Rising Patterns: Representation Shift showed a notable increase in 2024 (8.0% → 11.5%), reflecting a wave of representational innovations including new primitives, modalities, and implicit representations. Data & Evaluation Engineering shows modest growth in 2025 (6.6%), suggesting increased attention to benchmarks and reproducibility.

Declining Patterns: Formal-Experimental Tightening decreased as a primary pattern (10.1% → 7.1% → 6.6%), suggesting formalization is increasingly a supporting rather than headline contribution.

E.6 Conference-Specific Detailed Statistics

ICLR: Shows stronger emphasis on Representation Shift (11.8%) and Data & Evaluation Engineering (8.5%), consistent with ICLR’s reputation for representation learning and empirical benchmarks.

ICML: Higher concentrations of Gap-Driven Reframing (25.8%), Formal-Experimental Tightening (8.3%), and Principled Probabilistic Modeling (7.5%), aligning with ICML’s tradition of statistically grounded algorithmic work.

NeurIPS: Broadly balanced distribution with Gap-Driven Reframing (24.5%), Cross-Domain Synthesis (18.5%), and Formal Tightening (8.1%), reflecting NeurIPS’s cross-disciplinary character.

E.7 Underexplored Opportunity Spaces

Patterns with low primary frequency represent fertile areas for high-impact contributions:

- **Multiscale & Hierarchical Modeling (1.5%):** Many real systems exhibit hierarchical structure; deeper development could yield efficiency and interpretability gains.

- **Data-Centric Optimization & Active Sampling** (2.3%): Growing interest in data efficiency makes explicit active sampling methods increasingly relevant.
- **Inference-Time Control & Guided Sampling** (2.7%): Systems that adapt at inference to trade off compute and quality are underexplored relative to deployment value.
- **Adversary Modeling & Defensive Repurposing** (1.7%): Security and robustness patterns remain small but will become strategically important as ML deployments scale.

E.8 Actionable Insights for Researchers

The identified patterns translate directly into actionable strategies for conducting research. For PhD students and early-career researchers, the data suggests starting with focused “gap identification” exercises: systematically analyzing recent papers to write explicit gap statements reveals opportunities that others may have overlooked. Mastering at least one tool or formalism from an adjacent field (control theory, probabilistic graphical models, implicit representations, optimization theory) provides the raw material for cross-domain synthesis. The most accessible entry point is the “Reframe + Represent” trajectory: identify a crisp limitation, ask “what primitive would make this simple?”, and prototype with the new representation.

For experienced researchers and research teams, the analysis suggests investing in “reframe + representation” projects backed by rigorous validation, building cross-domain collaborations that pair empiricists with theoreticians and domain experts, and creating transferable tooling around innovations to maximize impact. The temporal trends also reveal underexplored opportunities: patterns like Inference-Time Control (2.7%), Multiscale Hierarchical Modeling (1.5%), and Adversarial Robustness (1.7%) show low current adoption but high potential impact, particularly for industry deployment scenarios. Finally, the conference-specific patterns suggest tailoring submission strategies: ICML submissions should emphasize mathematical rigor and statistical foundations, ICLR submissions should highlight architectural innovations and benchmark contributions, while NeurIPS submissions benefit from broad applicability and cross-disciplinary synthesis.

F Complete List of 15 Thinking Patterns

This section provides detailed descriptions of all 15 thinking patterns identified in our analysis of 3,819 papers. Each pattern includes its cognitive strategy, key indicators, concrete examples, and actionable insights for researchers.

F.1 P01: Gap-Driven Reframing (24.2%)

Cognitive Move: Turn a specific failure or mismatched assumption into an explicit design constraint that maps the problem onto better-suited methods.

Example: Reframing autoregressive image modeling from next-token prediction to next-scale (coarse→fine) prediction.

Learnable Insight: When you notice a recurring failure, write it as an explicit constraint; ask “if this limitation were the problem, what methods would apply?”

F.2 P02: Cross-Domain Synthesis (18.0%)

Cognitive Move: Map components across disciplinary boundaries and transplant them while engineering the compatibility layer.

Example: Fusing quantum circuits with transformer attention to obtain doubly stochastic attention matrices.

Learnable Insight: List constraints your method fails to satisfy, search other fields for primitives addressing those constraints, and prototype with a thin adapter.

F.3 P03: Representation Shift & Primitive Recasting (10.5%)

Cognitive Move: Replace the problem’s language (pixels, tokens, meshes) with an alternative primitive that simplifies inference or constraints.

Example: Replacing explicit meshes with neural implicit signed-distance functions for 3D reconstruction.

Learnable Insight: When a task struggles with geometry or combinatorics, enumerate alternative primitives and test whether the new one reduces complexity.

F.4 P04: Modular Pipeline Composition (4.7%)

Cognitive Move: Decompose a complex end-to-end system into composable modules with clean interfaces, enabling mix-and-match flexibility.

Example: Breaking monolithic vision models into separate perception, reasoning, and generation modules that can be independently upgraded.

Learnable Insight: When facing a complex system, identify natural factorization points and design interfaces that allow independent optimization of components.

F.5 P05: Data & Evaluation Engineering (6.0%)

Cognitive Move: Create new datasets, benchmarks, or evaluation metrics that expose previously unmeasured phenomena or enable fairer comparisons.

Example: Designing contamination-resistant evaluation protocols that detect when models have memorized test data during pretraining.

Learnable Insight: When existing benchmarks saturate or fail to capture important behaviors, invest in principled evaluation infrastructure as a research contribution.

F.6 P06: Principled Probabilistic Modeling (6.0%)

Cognitive Move: Formulate the problem using probabilistic graphical models, Bayesian inference, or statistical principles that provide interpretability and theoretical guarantees.

Example: Replacing heuristic neural architectures with structured variational autoencoders that factorize latent representations according to causal structure.

Learnable Insight: When a problem involves uncertainty, composition, or interpretability requirements, consider whether probabilistic frameworks provide cleaner solutions than deterministic neural networks.

F.7 P07: Formal-Experimental Tightening (7.4%)

Cognitive Move: Iterate between theoretical analysis (proofs, convergence guarantees, sample complexity) and empirical validation, using each to refine the other.

Example: Proving sample complexity bounds for a reinforcement learning algorithm, then using ablations to verify that empirical performance matches theoretical predictions.

Learnable Insight: Strong papers pair conceptual innovations with rigorous validation—either formal guarantees or exhaustive empirical analysis that rules out confounds.

F.8 P08: Approximation Engineering for Scalability (5.4%)

Cognitive Move: Replace exact but intractable operations with principled approximations (sketching, quantization, low-rank, sparse) that preserve essential properties while enabling scale.

Example: Approximating full attention with locality-sensitive hashing to achieve subquadratic complexity while maintaining quality.

Learnable Insight: When hitting computational bottlenecks, identify which properties are essential for downstream performance and design approximations that preserve those while discarding expensive-but-inessential structure.

F.9 P09: Inference-Time Control & Guided Sampling (2.7%)

Cognitive Move: Enable runtime control over model behavior through guided sampling, inference-time optimization, or adaptive computation without retraining.

Example: Using classifier-free guidance to steer diffusion models toward desired attributes at generation time without fine-tuning.

Learnable Insight: When deployment requirements vary (quality-latency tradeoffs, diverse user preferences), design systems that adapt at inference time rather than requiring separate fine-tuned models.

F.10 P10: Inject Structural Inductive Bias (5.7%)

Cognitive Move: Hardcode domain-specific structure (symmetries, invariances, geometric priors) into architectures to improve sample efficiency and generalization.

Example: Designing graph neural networks with permutation equivariance to handle variable-sized molecular structures.

Learnable Insight: Identify structural properties that hold universally in your domain (rotation, permutation, scaling invariances) and bake them into architectures as hard constraints.

F.11 P11: Multiscale & Hierarchical Modeling (1.5%)

Cognitive Move: Explicitly model phenomena at multiple levels of granularity or abstraction, with cross-scale interactions.

Example: Processing images through a pyramid of resolutions, with top-down and bottom-up information flow between levels.

Learnable Insight: When systems exhibit natural hierarchies (visual scenes, language semantics, physical simulations), explicit multiscale modeling often outperforms single-scale approaches.

F.12 P12: Mechanistic Decomposition & Causal Localization (3.8%)

Cognitive Move: Decompose model behavior into interpretable mechanisms or localize causal effects to specific components through interventions.

Example: Using activation patching to identify which transformer attention heads are causally responsible for specific reasoning capabilities.

Learnable Insight: When models exhibit surprising behaviors, systematic decomposition and causal interventions reveal which components are necessary and sufficient.

F.13 P13: Adversary Modeling & Defensive Repurposing (1.7%)

Cognitive Move: Model potential adversaries or failure modes explicitly, then design defenses or repurpose adversarial techniques for robustness.

Example: Using adversarial training to improve model robustness, or detecting out-of-distribution inputs through learned adversarial perturbations.

Learnable Insight: Explicitly modeling worst-case scenarios and adversarial behaviors leads to more robust systems than hoping for benign deployment conditions.

F.14 P14: Numerics & Systems Co-design (1.4%)

Cognitive Move: Co-design algorithms and their low-level implementation (numerical precision, memory layout, kernel fusion) to achieve orders-of-magnitude speedups.

Example: Redesigning transformer operations to exploit tensor core hardware, achieving 10x speedups through algorithmic and systems co-design.

Learnable Insight: For deployment-critical systems, collaborate with systems experts to co-design algorithms that exploit hardware characteristics rather than treating implementation as an afterthought.

F.15 P15: Data-Centric Optimization & Active Sampling (2.3%)

Cognitive Move: Optimize the data distribution rather than (or in addition to) the model, using

active learning, curriculum design, or strategic data curation.

Example: Using uncertainty-based active learning to identify maximally informative unlabeled examples, reducing labeling requirements by 10x.

Learnable Insight: When data acquisition is expensive or noisy, strategic data selection and curation can yield larger gains than architectural improvements.