Prompt Engineering as Epistemic Instrumentation: Multi-Lens Frameworks for Bias Detection in AI Research Evaluation

Brendan Malloy
Independent Researcher
github.com/bmalloy-224

October 22, 2025

Abstract

Large language models increasingly mediate research discovery, yet their evaluative criteria remain implicit and unexamined. I introduce a prompt-engineered framework that operationalizes research evaluation as multi-dimensional epistemic mapping. I tasked three frontier AI systems with independently assessing 2,548 scientific papers under four distinct evaluative lenses: Cold (baseline prestige), Implementation (12-month deployability), Transformative (paradigm-shifting potential), and Toolmaker (methodological enabling). My results reveal near-zero overlap (0-6%) between Implementation and Transformative selections despite variable intra-lens consensus (20-78% across lenses), indicating orthogonal value dimensions. Systems exhibited persistent architectural signatures: Model A favored mathematical formalism, Model B prioritized systems-level integration, Model C emphasized capability creation. Critically, papers achieving universal consensus under prestige-based evaluation disappeared entirely under implementation criteria, demonstrating that dominant and deployable research occupy disjoint spaces. My framework also reveals that evaluation biases are fundamentally *temporal*: each lens operates on a distinct horizon from immediate deployment to timeless infrastructure. I propose that systematic prompt variation functions as epistemic instrumentation—a reproducible method for surfacing latent value hierarchies in AI systems and enabling explicit value alignment in research assessment.

Keywords: prompt engineering, epistemic instrumentation, research evaluation, bias detection, temporal bias, value alignment, meta-science

1 Introduction

Artificial intelligence systems now routinely participate in research triage, literature review, and funding recommendations. Yet these systems inherit evaluative biases from their training data—implicit hierarchies that privilege certain research types over others. Traditional bibliometric approaches collapse research value into single dimensions: citation counts, journal prestige, or novelty metrics. This compression obscures the fundamentally multi-dimensional and temporally situated nature of scientific contribution.

I demonstrate that *prompt engineering can function as epistemic instrumentation*—a controlled method for probing and measuring the latent value structures within AI reasoning systems. By systematically varying evaluative criteria through prompt design, I surface distinct dimensions of research value that would otherwise remain entangled. Furthermore, I demonstrate that these value dimensions correspond to fundamentally different temporal scales of impact, from immediate deployment to timeless methodological infrastructure.

This paper reports results from a controlled experiment in which I tasked three frontier AI systems[1] with independently evaluating a shared corpus of 2,548 recent scientific papers under four distinct evaluative frameworks. My findings reveal that research value is not unidimensional but comprises at least four orthogonal axes, each surfacing fundamentally different work.

2 Related Work

Bibliometrics and Research Evaluation. Traditional approaches to research assessment rely on citation-based metrics [1, 2], which systematically favor established fields and cumulative knowledge over emerging domains and disruptive innovation. Recent work has documented how these metrics embed cultural and linguistic biases [3].

AI Evaluation and Value Alignment. The challenge of aligning AI systems with human values has been studied primarily in the context of harmful content generation [4]. However, less attention has been paid to implicit value hierarchies in evaluation tasks. Liang et al. [5] demonstrated that language model performance varies dramatically across evaluation frameworks, suggesting that assessment itself is value-laden.

Prompt Engineering as Methodology. Recent work has shown that prompts can elicit distinct reasoning modes from language models [6]. I extend this observation by treating prompt variation as a systematic measurement technique—what I term "epistemic instrumentation."

[1] Models evaluated: Model A (OpenAI o1-class reasoning model), Model B (Anthropic Claude Sonnet 4-class), Model C (DeepSeek R1-class). Specific versions withheld to emphasize architectural patterns over version-specific artifacts.

3 Methodology

3.1 Experimental Design

I designed a four-lens evaluation framework where each lens operationalizes a distinct dimension of research value:

1. **Cold Lens** (Baseline): "Select the most interesting papers." Minimal structure to reveal default preferences.

2. **Implementation Lens**: "Identify research with highest implementation-to-citation potential." Prioritizes work deployable by practitioners within 12 months.

3. **Transformative Lens**: "Identify research that changes possibility space." Focuses on paradigm-shifting potential and cascade effects.

4. **Toolmaker Lens**: "Identify research that enables other research." Emphasizes methodological infrastructure and generative capability.

Each lens included explicit inclusion criteria, exclusion rules, and required justifications. Critically, lenses were designed to be conceptually orthogonal: Implementation explicitly excluded theoretical frameworks, while Transformative explicitly excluded incremental optimization.

3.2 Corpus and Procedure

The corpus comprised 2,548 multidisciplinary research papers from Q4 2024 preprint servers, spanning physics, computer science, biology, social science, and engineering. This scale provides sufficient statistical power to detect systematic evaluation patterns across domains and enables robust estimation of effect sizes for inter-lens divergence and intra-lens consensus. Papers were presented to each model with titles only, simulating real-world research discovery scenarios.

Each of three models independently produced Top 10 selections for each lens, yielding 120 total evaluations (3 models $\times$ 4 lenses $\times$ 10 papers) drawn from a comprehensive evaluation of 30,576 paper-lens assessments. Models received no information about other models' selections and processed lenses in randomized order to prevent ordering effects.

3.3 Metrics

Intra-Lens Consensus: Jaccard similarity of Top 10 selections across models for the same lens.

Inter-Lens Divergence: Overlap between lenses within each model.

Domain Distribution: Manual categorization into research domains to detect systematic bias.

4 Results

Our multi-lens analysis reveals that AI systems evaluate research through orthogonal value dimensions with varying degrees of consensus. The near-zero overlap between Implementation and Transformative selections (0-6% across models) demonstrates that deployable and paradigm-shifting research occupy fundamentally disjoint spaces in AI evaluation. This orthogonality persists despite variable agreement within lenses, ranging from high consensus on methodological infrastructure (Toolmaker: 78% mean overlap) to highly subjective notions of prestige (Cold: 20% consensus). These findings indicate that AI models have internalized distinct, non-overlapping criteria for different forms of research value.

4.1 Orthogonal Value Dimensions

Cross-lens analysis revealed near-zero overlap between Implementation and Transformative selections (0-6% across models), confirming these dimensions are functionally orthogonal. In contrast, Cold and Transformative lenses showed moderate overlap (19-25%), reflecting shared emphasis on novelty and theoretical ambition.

Table 1: Cross-Lens Overlap (Model B - Anthropic, Jaccard Similarity)

	Cold	**Impl.**	**Trans.**
Cold	1.00	0.05	0.22
Implementation	0.05	1.00	0.00
Transformative	0.22	0.00	1.00
Toolmaker	0.08	0.09	0.11

Notably, papers achieving universal consensus under the Cold lens (e.g., "A Theory of Everything," "Aberrant E-I Balance in Major Depressive Disorder") were *entirely absent* from Implementation selections. This demonstrates that prestige and deployability occupy disjoint evaluative spaces.

4.2 Strong Intra-Lens Consensus with Lens-Specific Variation

Models exhibited variable agreement within lenses (mean overlap 48%, range 20-78%). The Toolmaker lens showed highest consensus (78% mean overlap), with Models A and C achieving perfect 10/10 agreement, suggesting clear shared understanding of methodological infrastructure. In contrast, the Cold lens showed lowest consensus (20%), reflecting more subjective and culturally-contingent notions of "interestingness."

Six papers achieved 3/3 consensus under the Transformative lens:

- Semantic Generalization of Shannon's Information Theory
- Diagram-Hilbert-Space Framework for Physics
- Brain Cell Type Resource via Multi-Agent LLMs
- Cascade Combustion Ion Technology (Interstellar Propulsion)
- Five-Dimensional Delay Field Cosmology Model
- A Theory of Everything

This convergence suggests these papers genuinely satisfy transformative criteria across different reasoning architectures.

4.3 Systematic Domain Bias

Cold lens selections over-represented theoretical physics and mathematics (30% of selections) relative to corpus representation (approximately 14%). Conversely, social science fieldwork comprised 23% of corpus but 0% of Cold selections. Implementation lens corrected this bias: social science rose to 31% of selections while physics dropped to 8%.

4.4 Persistent Architectural Signatures

Cross-lens analysis revealed model-specific evaluation patterns:

Table 2: Model Architectural Signatures

Model	**Signature**	**Characteristic Picks**
A	Mathematical Formalist	Bayesian inference, distributional RL, pattern recovery geometry
B	Systems Synthesizer	Disaster response, infrastructure modeling, comprehensive surveys
C	Capability Creator	Bioremediation tools, diagnostic methods, screening platforms

These tendencies persisted across all four lenses, suggesting they reflect training-data-derived value preferences rather than lens-specific responses.

5 Analysis

5.1 Prompt Engineering as Measurement

My results demonstrate that systematic prompt variation acts as an *epistemic probe*. Rather than introducing noise, controlled prompt manipulation surfaces latent dimensions of value that models learned from training data but cannot articulate explicitly.

This framing positions prompt engineering not as interface design but as *instrumentation*—analogous to how spectroscopy reveals chemical composition by systematically varying electromagnetic frequency. Different prompts activate different regions of a model's learned value landscape, making implicit preferences measurable.

5.2 Shared Training Epoch Effects

The strong intra-lens consensus likely reflects shared training epochs rather than independent discovery of objective quality. Like researchers educated in the same academic tradition, these models internalized overlapping citation structures, narrative conventions, and disciplinary hierarchies from temporally adjacent corpora.

This suggests that convergence indicates *cultural synchronization* more than objective correctness. A model trained on 1990s literature, or on practitioner blogs rather than academic papers, would likely produce different consensus sets.

5.3 Emergent Value Taxonomy

The four lenses reveal at least four distinct dimensions of research contribution:

- **Readiness (R):** Immediate practical deployment
- **Generativity (G):** Ability to spawn tools and methods
- **Catalyticity (C):** Potential to redefine disciplines
- **Prestige (P):** Alignment with dominant academic values

Each lens primarily projects onto a different axis (Implementation→R, Toolmaker→G, Transformative→C, Cold→P). This multi-dimensional structure explains why single-metric evaluation systematically fails: excellent work in one dimension may be invisible in another.

5.4 Temporal Architecture of Research Value

The four evaluative lenses can also be understood as probing distinct temporal horizons:

- **Cold Lens:** Present-moment prestige, reflecting what is "hot" now.

- **Implementation Lens:** Near-future deployability (12-month horizon).
- **Transformative Lens:** Long-term paradigm shifts (10+ years).
- **Toolmaker Lens:** Timeless infrastructure enabling research across eras.

The near-zero overlap between Implementation and Transformative selections (0-6%) reflects these distinct temporal orientations. Universal consensus picks under Cold but absent in Implementation illustrate how temporal positioning constrains model evaluation. The exceptionally high Toolmaker consensus (78%) suggests that identification of methodological infrastructure is less temporally contingent than other forms of value assessment. These findings suggest that AI evaluation biases are not only disciplinary but fundamentally temporal.

6 Discussion

6.1 Implications for Research Evaluation

Multi-lens evaluation enables *portfolio-based research assessment*. Rather than seeking the single "best" paper, evaluators can consciously balance:

- Near-term deliverables (Implementation)
- Long-term paradigm shifts (Transformative)
- Methodological infrastructure (Toolmaker)
- Field recognition (Cold)

Funding agencies could set explicit targets: 40% Implementation, 30% Transformative, 20% Toolmaker, 10% high-risk Cold selections. This makes value trade-offs transparent rather than implicit.

6.2 AI Value Alignment

This framework provides a diagnostic tool for AI alignment research. By measuring lens-dependent behavior, developers can:

- Detect unintended value biases in training data
- Test whether alignment interventions affect all value dimensions or just visible ones
- Design multi-objective training that balances competing goods

6.3 Limitations and Future Work

This study analyzed 2,548 papers from predominantly English-language sources. The scale provides robust statistical power for the main findings, but future work should:

- Test across broader temporal ranges and languages
- Include additional model families and architectures
- Validate long-term: do "Transformative" picks actually transform fields?
- Develop quantitative metrics for lens quality and orthogonality

7 Ethical Considerations

This framework surfaces but does not eliminate value biases. Practitioners must recognize that multi-lens evaluation reveals *which* preferences are active, not whether they are justified. Additionally, the models analyzed share Western, English-language training dominance; results may not generalize to systems trained on linguistically or culturally diverse corpora. I recommend pairing computational lens analysis with human expert panels representing diverse epistemic traditions.

8 Conclusion

I have demonstrated that prompt engineering can function as epistemic instrumentation—a reproducible method for measuring latent value structures in AI systems. By systematically varying evaluative criteria across a corpus of 2,548 papers, I revealed that research assessment comprises at least four orthogonal dimensions across distinct temporal horizons, each surfacing fundamentally different work.

This framework transforms research evaluation from implicit hierarchy to explicit multi-dimensional mapping. This enables conscious portfolio balancing, bias detection, and value alignment in AI-assisted scientific assessment. Most critically, it provides a methodology for making AI value judgments transparent, measurable, and accountable.

Future work should extend this approach beyond research evaluation to any domain where AI systems make quality judgments: hiring, content moderation, medical diagnosis, policy analysis. Wherever implicit values shape outcomes, epistemic instrumentation can make those values visible—and thus, governable.

References

[1] S. Fortunato et al., "Science of science," *Science*, vol. 359, no. 6379, 2018.

[2] L. Bornmann and H.-D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45-80, 2008.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM FAccT*, 2021, pp. 610-623.

[4] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[5] P. Liang et al., "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2023.

[6] E. Perez et al., "Discovering language model behaviors with model-written evaluations," in *Proc. ACL*, 2023, pp. 13387-13415.

A Complete 120-Selection Matrix

Table 3: Top 10 selections per model per lens (full matrix)

Lens/Model	Top Selections
Cold - A	1. Aberrant E-I Balance and Brain Criticality in Major Depressive Disorder 2. Absolute Abstraction: A Renormalisation Group Approach 3. A*-Thought: Efficient Reasoning via Bidirectional Compression 4. A Theory of Everything 5. A Diagram-Hilbert-Space Framework: Mathematical Foundations 6. A Five-Dimensional Delay Field Model with Mellin Integrals 7. A Survey of Scientific Large Language Models 8. A Systematic Review of Polyherbal Plant Combinations 9. A Bayesian Framework for Symmetry Inference in Chaotic Attractors 10. A Unified Framework for Pattern Recovery in Penalized Estimation
Cold - B	1. A Theory of Everything 2. A High-Coverage Genome from a 200,000-Year-Old Denisovan 3. A Survey of Scientific Large Language Models 4. A Comprehensive Survey on Reinforcement Learning-based Agentic Search 5. A Brain Cell Type Resource Created by Large Language Models 6. Aberrant E-I Balance and Brain Criticality in Major Depressive Disorder 7. A Framework for Hyper-Velocity Interplanetary Propulsion 8. A Theoretical Study on Bridging Internal Probability and Self-Consistency for LLM Reasoning 9. A Deep Learning Approach for Rational Affinity Maturation of Anti-VEGF Nanobodies

Continued on next page

Table 3 – continued from previous page

Lens/Model	Top Selections
	10. A Cocktail of SARS-CoV-2 Spike Stem Helix Domain and Receptor Binding Domain Human Monoclonal Antibodies
Cold - C	1. A Theory of Everything 2. A High-Coverage Genome from a 200,000-Year-Old Denisovan 3. A Brain Cell Type Resource Created by Large Language Models 4. A Human Alveolus-on-Chip Recapitulates SARS-CoV-2-mediated Lung Injury 5. Aberrant E-I Balance and Brain Criticality in Major Depressive Disorder 6. A Five-Dimensional Delay Field Model with Mellin Integrals 7. A Novel GPT-Based Framework for Anomaly Detection in System Logs 8. A Weakly Supervised Transformer for Rare Disease Diagnosis from EHRs 9. A Diagram-Hilbert-Space Framework: Mathematical Foundations 10. The Nash Code: Awareness, Freedom, and Ethical Game Strategy
Implementation - A	1. "No culture stops me from taking tablets": Exploring community-level factors influencing pregnant women's IPTp-SP uptake in Nigeria 2. A Vantagem do Trabalho Remoto na Administração Pública: Eficiência, Qualidade de Vida e Inovação na Gestão 3. A Recommendation System for Stress Management at the Workplace Using RAG-based LLM 4. A Real-Time BCI for Stroke Hand Rehabilitation Using Latent EEG Features from Healthy Subjects 5. A Spatio-Temporal Dynamic Approach to Modelling the Space-Time Dynamics of Poverty in Central Sulawesi 6. A Multi-Layered India-Specific Hate Speech Detection and Censorship System for Social Media 7. A Yeast-Based High-Throughput Screening Platform for the Discovery of Novel pre-mRNA Splicing Modulators 8. A Multimodal Lightweight Approach to Fault Diagnosis of Induction Motors in High-Dimensional Dataset 9. A Bacillus Sp. Strain Bgsc1 That Efficiently Degrades p-Nitrophenol 10. A Study on the Construction and Application of PBL Teaching Model of Pathology Driven by Metacognitive Strategies
Implementation - B	1. "No culture stops me from taking tablets": Exploring community-level factors influencing pregnant women's IPTp-SP uptake in Nigeria 2. A Recommendation System for Stress Management at the Workplace Using RAG-based LLM 3. A Real-Time BCI for Stroke Hand Rehabilitation Using Latent EEG Features from Healthy Subjects 4. A STUDY ON PERCEPTION OF RETAILERS TOWARDS DIGITAL PAYMENTS WITH REFERENCE TO SIRA TOWN 5. A Multi-Stage Hybrid CNN-Transformer Network for Automated Pediatric Lung Sound Classification

Continued on next page

Table 3 – continued from previous page

Lens/Model	Top Selections
	6. A Newly Designed GABA-AT Inactivator, (S)-MeCPP-115, Suppresses Paclitaxel-Induced Neuropathic Pain in Mice 7. A Vantagem do Trabalho Remoto na Administração Pública: Eficiência, Qualidade de Vida e Inovação na Gestão 8. A Study on Maneuverability of a Ship Equipped with the CFRP Propeller by Means of Simulation Model Including Engine Dynamics 9. A Storm-Centric 250 m NEXRAD Level-II Dataset for High-Resolution ML Nowcasting 10. A Social Context-Aware Graph-Based Multimodal Attentive Learning Framework for Disaster Content Classification During Emergencies
Implementation - C	1. "No culture stops me from taking tablets": Exploring community-level factors influencing pregnant women's IPTp-SP uptake in Nigeria 2. A STUDY ON PERCEPTION OF RETAILERS TOWARDS DIGITAL PAYMENTS WITH REFERENCE TO SIRA TOWN 3. A Vantagem do Trabalho Remoto na Administração Pública: Eficiência, Qualidade de Vida e Inovação na Gestão 4. A Bacillus Sp. Strain Bgsc1 That Efficiently Degrades p-Nitrophenol 5. A Real-Time BCI for Stroke Hand Rehabilitation Using Latent EEG Features from Healthy Subjects 6. A Newly Designed GABA-AT Inactivator, (S)-MeCPP-115, Suppresses Paclitaxel-Induced Neuropathic Pain in Mice 7. A Spatio-Temporal Dynamic Approach to Modelling the Space-Time Dynamics of Poverty in Central Sulawesi 8. A Multi-Layered India-Specific Hate Speech Detection and Censorship System for Social Media 9. A STUDY ON URBANIZATION AND SOCIAL CHANGE IN INDIA 10. A Recommendation System for Stress Management at the Workplace Using RAG-based LLM
Transformative - A	1. A Semantic Generalization of Shannon's Information Theory and Applications 2. A Diagram-Hilbert-Space Framework: Mathematical Foundations, Projection Mechanism and Emergence of the Standard Model 3. A Brain Cell Type Resource Created by Large Language Models and a Multi-Agent AI System for Collaborative Community Annotation 4. A Framework for Hyper-Velocity Interplanetary Propulsion: The Cascade Combustion Ion Technology (CComIT) Drive 5. A Five-Dimensional Delay Field Model with Mellin Integrals: Unifying Cosmology and Quantum Resonances 6. Absolute Abstraction: A Renormalisation Group Approach 7. A Pure Hypothesis Test for Inhomogeneous Random Graph Models Based on a Kernelised Stein Discrepancy

Continued on next page

Table 3 – continued from previous page

Lens/Model	Top Selections
	8. A Novel GPT-Based Framework for Anomaly Detection in System Logs 9. A Weakly Supervised Transformer for Rare Disease Diagnosis and Subphenotyping from EHRs with Pulmonary Case Studies 10. A Theory of Everything
Transformative - B	1. A Diagram-Hilbert-Space Framework: Mathematical Foundations, Projection Mechanism and Emergence of the Standard Model 2. A Theory of Everything 3. A Semantic Generalization of Shannon's Information Theory and Applications 4. A Brain Cell Type Resource Created by Large Language Models and a Multi-Agent AI System for Collaborative Community Annotation 5. A Framework for Hyper-Velocity Interplanetary Propulsion: The Cascade Combustion Ion Technology (CComIT) Drive 6. A Primer on Kolmogorov-Arnold Networks (KANs) for Probabilistic Time Series Forecasting 7. A Five-Dimensional Delay Field Model with Mellin Integrals: Unifying Cosmology and Quantum Resonances 8. A Tutorial on Discovering and Quantifying the Effect of Latent Causal Sources of Multimodal EHR Data 9. A Theoretical Study on Bridging Internal Probability and Self-Consistency for LLM Reasoning 10. A Survey of Scientific Large Language Models: From Data Foundations to Agent Frontiers
Transformative - C	1. A Theory of Everything 2. A Diagram-Hilbert-Space Framework: Mathematical Foundations, Projection Mechanism and Emergence of the Standard Model 3. A Semantic Generalization of Shannon's Information Theory and Applications 4. A Five-Dimensional Delay Field Model with Mellin Integrals: Unifying Cosmology and Quantum Resonances 5. A Brain Cell Type Resource Created by Large Language Models and a Multi-Agent AI System for Collaborative Community Annotation 6. A Framework for Hyper-Velocity Interplanetary Propulsion: The Cascade Combustion Ion Technology (CComIT) Drive 7. Absolute Abstraction: A Renormalisation Group Approach 8. A Pure Hypothesis Test for Inhomogeneous Random Graph Models Based on a Kernelised Stein Discrepancy 9. A Novel GPT-Based Framework for Anomaly Detection in System Logs 10. A Weakly Supervised Transformer for Rare Disease Diagnosis and Subphenotyping from EHRs with Pulmonary Case Studies
Toolmaker - A	1. A Yeast-Based High-Throughput Screening Platform for the Discovery of Novel pre-mRNA Splicing Modulators

Continued on next page

Table 3 – continued from previous page

Lens/Model	Top Selections
	2. A Pure Hypothesis Test for Inhomogeneous Random Graph Models Based on a Kernelised Stein Discrepancy 3. A Storm-Centric 250 m NEXRAD Level-II Dataset for High-Resolution ML Nowcasting 4. A Standardized Benchmark for Machine-Learned Molecular Dynamics using Weighted Ensemble Sampling 5. A Bacillus Sp. Strain Bgsc1 That Efficiently Degrades p-Nitrophenol 6. A Novel GPT-Based Framework for Anomaly Detection in System Logs 7. A Weakly Supervised Transformer for Rare Disease Diagnosis and Subphenotyping from EHRs with Pulmonary Case Studies 8. A Comprehensive, Open-Source Battery of Movement Imagery Ability Tests: Development and Psychometric Properties 9. A Simple Method for PMF Estimation on Large Supports 10. A Split-Client Approach to Second-Order Optimization
Toolmaker - B	1. A Yeast-Based High-Throughput Screening Platform for the Discovery of Novel pre-mRNA Splicing Modulators 2. A Pure Hypothesis Test for Inhomogeneous Random Graph Models Based on a Kernelised Stein Discrepancy 3. A Storm-Centric 250 m NEXRAD Level-II Dataset for High-Resolution ML Nowcasting 4. A Weakly Supervised Transformer for Rare Disease Diagnosis and Subphenotyping from EHRs with Pulmonary Case Studies 5. A Standardized Benchmark for Machine-Learned Molecular Dynamics using Weighted Ensemble Sampling 6. A Bacillus Sp. Strain Bgsc1 That Efficiently Degrades p-Nitrophenol 7. A Comprehensive, Open-Source Battery of Movement Imagery Ability Tests: Development and Psychometric Properties 8. A Sulfatide-Centered Ultra-High Resolution Magnetic Resonance MALDI Imaging Benchmark Dataset for MS1-Based Lipid Annotation Tools 9. A Novel GPT-Based Framework for Anomaly Detection in System Logs 10. A Simple Mean Field Model of Feature Learning
Toolmaker - C	1. A Yeast-Based High-Throughput Screening Platform for the Discovery of Novel pre-mRNA Splicing Modulators 2. A Pure Hypothesis Test for Inhomogeneous Random Graph Models Based on a Kernelised Stein Discrepancy 3. A Storm-Centric 250 m NEXRAD Level-II Dataset for High-Resolution ML Nowcasting 4. A Standardized Benchmark for Machine-Learned Molecular Dynamics using Weighted Ensemble Sampling 5. A Bacillus Sp. Strain Bgsc1 That Efficiently Degrades p-Nitrophenol 6. A Novel GPT-Based Framework for Anomaly Detection in System Logs

Continued on next page

Table 3 – continued from previous page

Lens/Model	Top Selections
	7. A Weakly Supervised Transformer for Rare Disease Diagnosis and Subphenotyping from EHRs with Pulmonary Case Studies 8. A Comprehensive, Open-Source Battery of Movement Imagery Ability Tests: Development and Psychometric Properties 9. A Simple Method for PMF Estimation on Large Supports 10. A Split-Client Approach to Second-Order Optimization